Spring School - April 2016 - Spartan/Macsenet
Francis Bach

Slides generously provided by Guillaume Obozinski

Probabilistic models

Guillaume Obozinski

Ecole des Ponts - ParisTech

SOCN course 2014

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

References for further reading

Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

Kevin Murphy. Machine Learning: a Probabilistic Perspective. MIT Press, 2012.

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

Statistical concepts

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \mathrm{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \theta^x (1 - \theta)^{(1-x)}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \text{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Binomial model: $X \sim \text{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \text{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Binomial model: $X \sim \text{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

Multinomial model: $X \sim \mathcal{M}(n, \pi_1, \pi_2, \ldots, \pi_K) \qquad \Theta = [0, 1]^K$

$$p(x|\theta) = \binom{n}{x_1, \ldots, x_k} {\pi_1}^{x_1} \ \ldots \ {\pi_k}^{x_k}$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.
So $\mathbf{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\mathbf{y} = (0, 0, 0, 1, 0)^\top$.
So $\mathbf{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

$$\mathbb{P}(C = k) = \mathbb{P}(Y_k = 1) \quad \text{and} \quad \mathbb{P}(Y = y) = \prod_{k=1}^K \pi_k^{y_k}.$$

Bernoulli, Binomial, Multinomial

$Y \sim \text{Ber}(\pi)$	$(Y_1, \ldots, Y_K) \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$
$p(y) = \pi^y (1-\pi)^{1-y}$	$p(\mathbf{y}) = \pi_1^{y_1} \ldots \pi_K^{y_K}$
$N_1 \sim \text{Bin}(n, \pi)$	$(N_1, \ldots, N_K) \sim \mathcal{M}(n, \pi_1, \ldots, \pi_K)$
$p(n_1) = \binom{n}{n_1} \pi^{n_1} (1-\pi)^{n-n_1}$	$p(\mathbf{n}) = \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} \pi_1^{n_1} \ldots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \quad \text{and} \quad \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} = \frac{n!}{n_1! \ldots n_K!}$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}(\mu, \sigma^2)$

X real valued r.v., and $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$

X real valued r.v., and $\theta = \left(\mu, \sigma^2\right) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

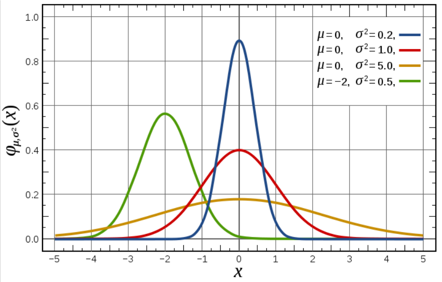
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Multivariate Gaussian model: $X \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

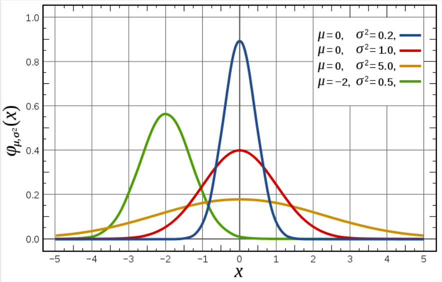
X r.v. taking values in $\mathbb{R}^d$. If $\mathcal{K}_n$ is the set of positive definite matrices of size $n \times n$, and $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Theta = \mathbb{R}^d \times \mathcal{K}_n$.

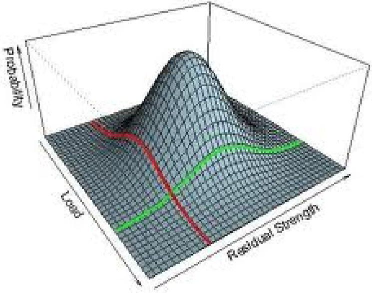
$$p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Gaussian densities



Gaussian densities





Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \ldots, X^{(n)}$$

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P.

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P.

This collection of observations is called

- the *sample* or the *observations* in statistics
- the *sample***s** in engineering
- the *training set* in machine learning

Outline

1 Statistical concepts

2 A short review of convex analysis and optimization

3 The maximum likelihood principle

4 Linear regression

5 Logistic regression

6 Fisher discriminant analysis

7 Clustering

8 The EM algorithm for the Gaussian mixture model

9 Hidden Markov models

A short review of convex analysis and optimization

Review: convex analysis

Convex function

$$\forall \lambda \in [0,1], \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Review: convex analysis

Convex function

$$\forall \lambda \in [0,1], \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Strictly convex function

$$\forall \lambda \in]0,1[, \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Review: convex analysis

Convex function

$$\forall \lambda \in [0,1], \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Strictly convex function

$$\forall \lambda \in]0,1[, \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Strongly convex function

$$\exists \mu > 0, \text{ s.t. } \quad \mathbf{x} \mapsto f(\mathbf{x}) - \mu \|\mathbf{x}\|^2 \quad \text{is convex}$$

Equivalently:

Review: convex analysis

Convex function

$$\forall \lambda \in [0,1], \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Strictly convex function

$$\forall \lambda \in]0,1[, \qquad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Strongly convex function

$$\exists \mu > 0, \text{ s.t. } \quad \mathbf{x} \mapsto f(\mathbf{x}) - \mu \|\mathbf{x}\|^2 \quad \text{is convex}$$

Equivalently:

$$\forall \lambda \in [0,1], \quad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda\, f(\mathbf{x}) + (1-\lambda)\, f(\mathbf{y}) - \mu\, \lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$$

The largest possible μ is called the strong convexity constant.

Minima of convex functions

Proposition (Supporting hyperplane)

If f is convex and differentiable at $\mathbf{x}$ *then*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

Minima of convex functions

Proposition (Supporting hyperplane)

If f is convex and differentiable at $\mathbf{x}$ then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

Convex function

All local minima are global minima.

Minima of convex functions

Proposition (Supporting hyperplane)

If f is convex and differentiable at $\mathbf{x}$ then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

Convex function

All local minima are global minima.

Strictly convex function

If there is a local minimum, then it is unique and global.

Strongly convex function

Minima of convex functions

Proposition (Supporting hyperplane)

If f is convex and differentiable at $\mathbf{x}$ then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

Convex function
All local minima are global minima.

Strictly convex function
If there is a local minimum, then it is unique and global.

Strongly convex function
There exists a unique local minimum which is also global.

Minima and stationary points of differentiable functions

Definition (Stationary point)

For f differentiable, we say that $\mathbf{x}$ is a stationary point if $\nabla f(\mathbf{x}) = 0$.

Minima and stationary points of differentiable functions

Definition (Stationary point)

For f differentiable, we say that $\mathbf{x}$ is a stationary point if $\nabla f(\mathbf{x}) = 0$.

Theorem (Fermat)

If f is differentiable at $\mathbf{x}$ and $\mathbf{x}$ is a local minimum, then $\mathbf{x}$ is stationary.

Minima and stationary points of differentiable functions

Definition (Stationary point)

For f differentiable, we say that $\mathbf{x}$ is a stationary point if $\nabla f(\mathbf{x}) = 0$.

Theorem (Fermat)

If f is differentiable at $\mathbf{x}$ and $\mathbf{x}$ is a local minimum, then $\mathbf{x}$ is stationary.

Theorem (Stationary point of a convex differentiable function)

If f is convex and differentiable at $\mathbf{x}$ and $\mathbf{x}$ is stationary then $\mathbf{x}$ is a minimum.

Minima and stationary points of differentiable functions

Definition (Stationary point)

For f differentiable, we say that $\mathbf{x}$ is a stationary point if $\nabla f(\mathbf{x}) = 0$.

Theorem (Fermat)

If f is differentiable at $\mathbf{x}$ and $\mathbf{x}$ is a local minimum, then $\mathbf{x}$ is stationary.

Theorem (Stationary point of a convex differentiable function)

If f is convex and differentiable at $\mathbf{x}$ and $\mathbf{x}$ is stationary then $\mathbf{x}$ is a minimum.

Theorem (Stationary points of a twice differentiable functions)

For f twice differentiable at $\mathbf{x}$

- *if $\mathbf{x}$ is a local minimum then $\nabla f(\mathbf{x}) = 0$ and $\nabla^2 f(\mathbf{x}) \succeq 0$.*
- *conversely if $\nabla f(\mathbf{x}) = 0$ and $\nabla^2 f(\mathbf{x}) \succ 0$ then $\mathbf{x}$ is a strict local minimum.*

Minima of differentiable functions under linear constraints

Theorem

If the function f is differentiable at $\mathbf{x}$, and $\mathbf{x}$ is a local minimum of

$$\min f(x) \quad s.t. \quad A\mathbf{x} = b$$

with $A \in \mathbb{R}^{n \times p}$ then $\mathbf{x}$ must satisfy

$$\nabla f(\mathbf{x}) + A^\top \lambda = 0,$$

for some $\lambda \in \mathbb{R}^n$.

Minima of differentiable functions under linear constraints

Theorem

If the function f is differentiable at $\mathbf{x}$, and $\mathbf{x}$ is a local minimum of

$$\min f(x) \quad s.t. \quad A\mathbf{x} = b$$

with $A \in \mathbb{R}^{n \times p}$ then $\mathbf{x}$ must satisfy

$$\nabla f(\mathbf{x}) + A^\top \lambda = 0,$$

for some $\lambda \in \mathbb{R}^n$.

More optimization later...

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

The maximum likelihood principle

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(x|\theta)$$

Sir Ronald Fisher
(1890-1962)

Maximum likelihood principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta \in \Theta} p(x|\theta)$$

Sir Ronald Fisher (1890-1962)

Case of i.i.d data

If $(x_i)_{1 \le i \le n}$ is an i.i.d. sample of size n:

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta \in \Theta} \prod_{i=1}^{n} p(x_i|\theta) = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \log\ p(x_i|\theta)$$

Examples of computation of the MLE

- Bernoulli model
- Multinomial model
- Gaussian model

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

Linear regression

Generative models vs conditional models

- X is the input variable
- Y is the output variable

A **generative model** is a model of the joint distribution $p(x, y)$.

Generative models vs conditional models

- X is the input variable
- Y is the output variable

A **generative model** is a model of the joint distribution $p(x, y)$.

A **conditional model** is a model of the conditional distribution $p(y|x)$.

Generative models vs conditional models

- X is the input variable
- Y is the output variable

A **generative model** is a model of the joint distribution $p(x, y)$.

A **conditional model** is a model of the conditional distribution $p(y|x)$.

Conditional models vs Generative models

- CM make less assumptions about the data distribution
- CM Require fewer parameters
- CM are typically harder to learn
- CM can typically not handle missing data or latent variables

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\mathbf{w}^\top X + b, \sigma^2)$$

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\mathbf{w}^\top X + b, \sigma^2)$$

or equivalently $Y = \mathbf{w}^\top X + b + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$.

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\mathbf{w}^\top X + b, \sigma^2)$$

or equivalently $Y = \mathbf{w}^\top X + b + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$.

The offset can be ignored up to a reparameterization.

$$Y = \tilde{\mathbf{w}}^\top \begin{pmatrix} x \\ 1 \end{pmatrix} + \epsilon.$$

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\mathbf{w}^\top X + b, \sigma^2)$$

or equivalently $Y = \mathbf{w}^\top X + b + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$.

The offset can be ignored up to a reparameterization.

$$Y = \tilde{\mathbf{w}}^\top \begin{pmatrix} x \\ 1 \end{pmatrix} + \epsilon.$$

Likelihood for one pair

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2} \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right)$$

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\mathbf{w}^\top X + b, \sigma^2)$$

or equivalently $Y = \mathbf{w}^\top X + b + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$

The offset can be ignored up to a reparameterization.

$$Y = \tilde{\mathbf{w}}^\top \begin{pmatrix} x \\ 1 \end{pmatrix} + \epsilon.$$

Likelihood for one pair

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2}\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right)$$

Negative log-likelihood

$$-\ell(\mathbf{w}, \sigma^2) = -\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}.$$

Probabilistic version of linear regression

$$\min_{\sigma^2, \mathbf{w}} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}$$

Probabilistic version of linear regression

$$\min_{\sigma^2, \mathbf{w}} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in $\mathbf{w}$

$$\min_{\mathbf{w}} \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

that we recognize as the usual linear regression, with

- $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and
- $\mathbf{X}$ the design matrix with rows equal to $\mathbf{x}_i^\top$.

Probabilistic version of linear regression

$$\min_{\sigma^2, \mathbf{w}} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in **w**

$$\min_{\mathbf{w}} \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

that we recognize as the usual linear regression, with

- $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and
- $\mathbf{X}$ the design matrix with rows equal to $\mathbf{x}_i^\top$.

Optimizing over σ^2, we find:

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{w}}_{MLE}^\top \mathbf{x}_i)^2$$

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

Logistic regression

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

Implies that

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$$

for

$$\sigma : z \mapsto \frac{1}{1 + e^{-z}},$$

the logistic function.

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

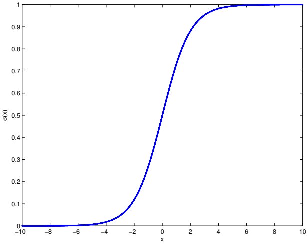
Implies that

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$$

for

$$\sigma : z \mapsto \frac{1}{1 + e^{-z}},$$

the logistic function.



Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

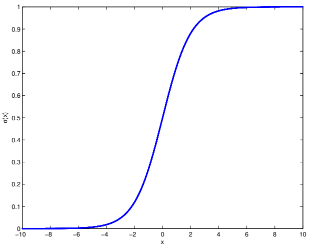
Implies that

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$$

for

$$\sigma : z \mapsto \frac{1}{1 + e^{-z}},$$

the logistic function.



- The logistic function is part of the family of *sigmoid functions*.
- Often called "the" sigmoid function.

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \mathbf{w}^\top \mathbf{x}$$

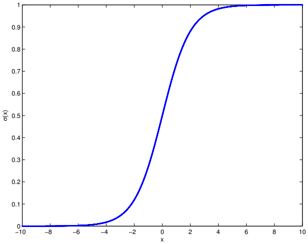
Implies that

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$$

for

$$\sigma : z \mapsto \frac{1}{1 + e^{-z}},$$

the logistic function.



- The logistic function is part of the family of *sigmoid functions*.
- Often called "the" sigmoid function.

Properties:

$$\begin{aligned}
\forall z \in \mathbb{R}, \quad \sigma(-z) &= 1 - \sigma(z), \\
\forall z \in \mathbb{R}, \quad \sigma'(z) &= \sigma(z)(1 - \sigma(z)) \\
&= \sigma(z)\sigma(-z).
\end{aligned}$$

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y | X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y|X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y(1-\eta)^{1-y} = \sigma(\mathbf{w}^\top \mathbf{x})^y \sigma(-\mathbf{w}^\top \mathbf{x})^{1-y}.$$

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y|X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y (1-\eta)^{1-y} = \sigma(\mathbf{w}^\top \mathbf{x})^y \sigma(-\mathbf{w}^\top \mathbf{x})^{1-y}.$$

Log-likelihood

$$\ell(\mathbf{w}) = y \log \sigma(\mathbf{w}^\top \mathbf{x}) + (1-y) \log \sigma(-\mathbf{w}^\top \mathbf{x})$$

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y|X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y (1-\eta)^{1-y} = \sigma(\mathbf{w}^\top \mathbf{x})^y \sigma(-\mathbf{w}^\top \mathbf{x})^{1-y}.$$

Log-likelihood

$$\begin{aligned} \ell(\mathbf{w}) &= y \log \sigma(\mathbf{w}^\top \mathbf{x}) + (1-y) \log \sigma(-\mathbf{w}^\top \mathbf{x}) \\ &= y \log \eta + (1-y) \log(1-\eta) \end{aligned}$$

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y|X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y (1-\eta)^{1-y} = \sigma(\mathbf{w}^\top \mathbf{x})^y \sigma(-\mathbf{w}^\top \mathbf{x})^{1-y}.$$

Log-likelihood

$$\begin{aligned} \ell(\mathbf{w}) &= y \log \sigma(\mathbf{w}^\top \mathbf{x}) + (1-y) \log \sigma(-\mathbf{w}^\top \mathbf{x}) \\ &= y \log \eta + (1-y) \log(1-\eta) \\ &= y \log \frac{\eta}{1-\eta} + \log(1-\eta) \end{aligned}$$

Likelihood for logistic regression

Let $\eta := \sigma(\mathbf{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y|X = \mathbf{x} \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y (1-\eta)^{1-y} = \sigma(\mathbf{w}^\top \mathbf{x})^y \sigma(-\mathbf{w}^\top \mathbf{x})^{1-y}.$$

Log-likelihood

$$\begin{aligned}
\ell(\mathbf{w}) &= y \log \sigma(\mathbf{w}^\top \mathbf{x}) + (1-y) \log \sigma(-\mathbf{w}^\top \mathbf{x}) \\
&= y \log \eta + (1-y) \log(1-\eta) \\
&= y \log \frac{\eta}{1-\eta} + \log(1-\eta) \\
&= y \mathbf{w}^\top \mathbf{x} + \log \sigma(-\mathbf{w}^\top \mathbf{x})
\end{aligned}$$

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} y_i \mathbf{w}^\top \mathbf{x}_i + \log \sigma(-\mathbf{w}^\top \mathbf{x}_i).$$

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} y_i \mathbf{w}^\top \mathbf{x}_i + \log \sigma(-\mathbf{w}^\top \mathbf{x}_i).$$

The log-likelihood is differentiable and concave.

$\Rightarrow$ Its global maxima are its stationary points.

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} y_i \mathbf{w}^\top \mathbf{x}_i + \log \sigma(-\mathbf{w}^\top \mathbf{x}_i).$$

The log-likelihood is differentiable and concave.
$\Rightarrow$ Its global maxima are its stationary points.

Gradient of ℓ

$$\begin{aligned} \nabla \ell(\mathbf{w}) &= \sum_{i=1}^{n} y_i \mathbf{x}_i - \mathbf{x}_i \frac{\sigma(-\mathbf{w}^\top \mathbf{x}_i)\sigma(\mathbf{w}^\top \mathbf{x}_i)}{\sigma(-\mathbf{w}^\top \mathbf{x}_i)} \\ &= \sum_{i=1}^{n} (y_i - \eta_i)\mathbf{x}_i \qquad \text{with} \qquad \eta_i = \sigma(\mathbf{w}^\top \mathbf{x}_i). \end{aligned}$$

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} y_i \mathbf{w}^\top \mathbf{x}_i + \log \sigma(-\mathbf{w}^\top \mathbf{x}_i).$$

The log-likelihood is differentiable and concave.
$\Rightarrow$ Its global maxima are its stationary points.

Gradient of ℓ

$$\begin{aligned}\nabla \ell(\mathbf{w}) &= \sum_{i=1}^{n} y_i \mathbf{x}_i - \mathbf{x}_i \frac{\sigma(-\mathbf{w}^\top \mathbf{x}_i)\sigma(\mathbf{w}^\top \mathbf{x}_i)}{\sigma(-\mathbf{w}^\top \mathbf{x}_i)} \\ &= \sum_{i=1}^{n} (y_i - \eta_i)\mathbf{x}_i \qquad \text{with} \qquad \eta_i = \sigma(\mathbf{w}^\top \mathbf{x}_i).\end{aligned}$$

Thus, $\nabla \ell(\mathbf{w}) = 0 \Leftrightarrow \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} y_i \mathbf{w}^\top \mathbf{x}_i + \log \sigma(-\mathbf{w}^\top \mathbf{x}_i).$$

The log-likelihood is differentiable and concave.
$\Rightarrow$ Its global maxima are its stationary points.

Gradient of ℓ

$$\begin{aligned} \nabla \ell(\mathbf{w}) &= \sum_{i=1}^{n} y_i \mathbf{x}_i - \mathbf{x}_i \frac{\sigma(-\mathbf{w}^\top \mathbf{x}_i)\sigma(\mathbf{w}^\top \mathbf{x}_i)}{\sigma(-\mathbf{w}^\top \mathbf{x}_i)} \\ &= \sum_{i=1}^{n} (y_i - \eta_i)\mathbf{x}_i \qquad \text{with} \qquad \eta_i = \sigma(\mathbf{w}^\top \mathbf{x}_i). \end{aligned}$$

Thus, $\quad \nabla \ell(\mathbf{w}) = 0 \Leftrightarrow \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$

No closed form solution !

Second order Taylor expansion

Need an iterative method to solve $$\sum_{i=1}^{n} \mathbf{x}_i(y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$$

Second order Taylor expansion

Need an iterative method to solve $$\sum_{i=1}^{n} \mathbf{x}_i(y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$$

- $\rightarrow$ Gradient descent (aka steepest descent)
- $\rightarrow$ Newton's method

Second order Taylor expansion

Need an iterative method to solve $\sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$

- $\rightarrow$ Gradient descent (aka steepest descent)
- $\rightarrow$ Newton's method

Hessian of ℓ

$$
\begin{aligned}
H\ell(\mathbf{w}) &= \sum_{i=1}^{n} \mathbf{x}_i (0 - \sigma'(\mathbf{w}^\top \mathbf{x}_i)\sigma'(-\mathbf{w}^\top \mathbf{x}_i)\mathbf{x}_i^\top) \\
&= \sum_{i=1}^{n} -\eta_i(1-\eta_i)\mathbf{x}_i\mathbf{x}_i^\top = -\mathbf{X}^\top \mathrm{Diag}(\eta_i(1-\eta_i))\mathbf{X}
\end{aligned}
$$

where $\mathbf{X}$ is the design matrix.

Second order Taylor expansion

Need an iterative method to solve $\sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$

→ Gradient descent (aka steepest descent)

→ Newton's method

Hessian of ℓ

$$\begin{aligned} H\ell(\mathbf{w}) &= \sum_{i=1}^{n} \mathbf{x}_i (0 - \sigma'(\mathbf{w}^\top \mathbf{x}_i) \sigma'(-\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i^\top) \\ &= \sum_{i=1}^{n} -\eta_i (1 - \eta_i) \mathbf{x}_i \mathbf{x}_i^\top = -\mathbf{X}^\top \mathrm{Diag}(\eta_i (1 - \eta_i)) \mathbf{X} \end{aligned}$$

where $\mathbf{X}$ is the design matrix.

→ Note that $-H\ell$ is p.s.d. $\Rightarrow \ell$ is concave.

Newton's method

Use the Taylor expansion

$$\ell(\mathbf{w}^t) + (\mathbf{w} - \mathbf{w}^t)^\top \nabla \ell(\mathbf{w}^t) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^t)^\top H\ell(\mathbf{w}^t)(\mathbf{w} - \mathbf{w}^t).$$

and minimize w.r.t. **w**.

Newton's method

Use the Taylor expansion

$$\ell(\mathbf{w}^t) + (\mathbf{w} - \mathbf{w}^t)^\top \nabla \ell(\mathbf{w}^t) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^t)^\top H\ell(\mathbf{w}^t)(\mathbf{w} - \mathbf{w}^t).$$

and minimize w.r.t. $\mathbf{w}$. Setting $\mathbf{h} = \mathbf{w} - \mathbf{w}^t$, we get

$$\max_{\mathbf{h}} \mathbf{h}^\top \nabla_{\mathbf{w}} \ell(\mathbf{w}) + \frac{1}{2}\mathbf{h}^\top H\ell(\mathbf{w})\mathbf{h}.$$

Newton's method

Use the Taylor expansion

$$\ell(\mathbf{w}^t) + (\mathbf{w} - \mathbf{w}^t)^\top \nabla \ell(\mathbf{w}^t) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^t)^\top H\ell(\mathbf{w}^t)(\mathbf{w} - \mathbf{w}^t).$$

and minimize w.r.t. $\mathbf{w}$. Setting $\mathbf{h} = \mathbf{w} - \mathbf{w}^t$, we get

$$\max_{\mathbf{h}} \mathbf{h}^\top \nabla_{\mathbf{w}} \ell(\mathbf{w}) + \frac{1}{2}\mathbf{h}^\top H\ell(\mathbf{w})\mathbf{h}.$$

I.e., for logistic regression, writing $\mathbf{D}_{\boldsymbol{\eta}} = \mathrm{Diag}\big((\eta_i(1-\eta_i))_i\big)$

$$\min_{\mathbf{h}} \quad \mathbf{h}^\top \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\eta}) - \frac{1}{2}\mathbf{h}^\top \mathbf{X}^\top \mathbf{D}_{\eta} \mathbf{X}\mathbf{h}$$

Newton's method

Use the Taylor expansion

$$\ell(\mathbf{w}^t) + (\mathbf{w} - \mathbf{w}^t)^\top \nabla \ell(\mathbf{w}^t) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^t)^\top H\ell(\mathbf{w}^t)(\mathbf{w} - \mathbf{w}^t).$$

and minimize w.r.t. $\mathbf{w}$. Setting $\mathbf{h} = \mathbf{w} - \mathbf{w}^t$, we get

$$\max_{\mathbf{h}} \mathbf{h}^\top \nabla_{\mathbf{w}} \ell(\mathbf{w}) + \frac{1}{2}\mathbf{h}^\top H\ell(\mathbf{w})\mathbf{h}.$$

I.e., for logistic regression, writing $\mathbf{D}_{\boldsymbol{\eta}} = \mathrm{Diag}\big((\eta_i(1-\eta_i))_i\big)$

$$\min_{\mathbf{h}} \quad \mathbf{h}^\top \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\eta}) - \frac{1}{2}\mathbf{h}^\top \mathbf{X}^\top \mathbf{D}_{\eta} \mathbf{X}\mathbf{h}$$

Modified normal equations

$$\mathbf{X}^\top \mathbf{D}_{\eta} \mathbf{X}\,\mathbf{h} - \mathbf{X}^\top \tilde{\mathbf{y}} \qquad \text{with} \qquad \tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\eta}.$$

Iterative Reweighted Least Squares (IRLS)

Assuming $\mathbf{X}^\top \mathbf{D}_{\boldsymbol{\eta}} \mathbf{X}$ is invertible, the algorithm takes the form

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + (\mathbf{X}^\top \mathbf{D}_{\boldsymbol{\eta}^{(t)}} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\eta}^{(t)}).$$

Iterative Reweighted Least Squares (IRLS)

Assuming $\mathbf{X}^\top \mathbf{D}_{\boldsymbol{\eta}} \mathbf{X}$ is invertible, the algorithm takes the form

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + (\mathbf{X}^\top \mathbf{D}_{\boldsymbol{\eta}^{(t)}} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\eta}^{(t)}).$$

This is called iterative reweighted least squares because each step is equivalent to solving the reweighted least squares problem:

$$\frac{1}{2} \sum_{i=1}^{n} \frac{1}{\tau_i^2} (\mathbf{x}_i^\top \mathbf{h} - \check{y}_i)^2$$

with

$$\tau_i^2 = \frac{1}{\eta_i^{(t)}(1 - \eta_i^{(t)})} \qquad \text{and} \qquad \check{y}_i = \tau_i^2 (y_i - \eta_i^{(t)}).$$

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

$$\mathbb{P}(Y = y | X = \mathbf{x}) = \sigma(y\, \mathbf{w}^\top \mathbf{x})$$

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

$$\mathbb{P}(Y = y | X = \mathbf{x}) = \sigma(y\,\mathbf{w}^\top \mathbf{x})$$

Log-likelihood

$$\ell(\mathbf{w}) = \log \sigma(y\mathbf{w}^\top \mathbf{x}) = -\log\left(1 + \exp(-y\mathbf{w}^\top x)\right)$$

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

$$\mathbb{P}(Y = y | X = \mathbf{x}) = \sigma(y\, \mathbf{w}^\top \mathbf{x})$$

Log-likelihood

$$\ell(\mathbf{w}) = \log \sigma(y\mathbf{w}^\top \mathbf{x}) = -\log\left(1 + \exp(-y\mathbf{w}^\top x)\right)$$

Log-likelihood for a training set

$$\ell(\mathbf{w}) = -\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^\top x_i)\right)$$

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

$$\mathbb{P}(Y = y|X = \mathbf{x}) = \sigma(y\,\mathbf{w}^\top \mathbf{x})$$

Log-likelihood

$$\ell(\mathbf{w}) = \log \sigma(y\mathbf{w}^\top \mathbf{x}) = -\log\left(1 + \exp(-y\mathbf{w}^\top x)\right)$$

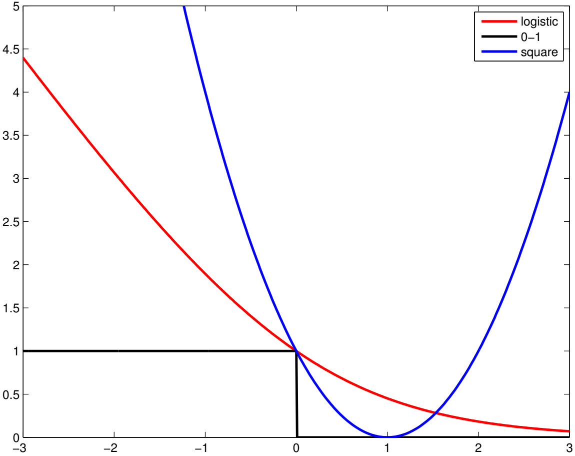
Log-likelihood for a training set

$$\ell(\mathbf{w}) = -\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^\top x_i)\right)$$

The negative log-likelihood takes the form of an empirical risk with loss

$$(a, y) = h(ya) \qquad \text{with} \qquad h : z \mapsto \log\left(1 + e^{-ya}\right)$$

Comparing losses



$\ell(a, 1)$ for several classification losses

(the logistic loss is scaled by $\log(2)^{-1}$)

Maximum likelihood for conditional models as ERM

Given a probabilistic model $p_\theta(y)$, define the loss function ℓ by

$$\ell : (\theta, y) \mapsto -\log p_\theta(y)$$

Maximum likelihood for conditional models as ERM

Given a probabilistic model $p_\theta(y)$, define the loss function ℓ by

$$\ell : (\theta, y) \mapsto -\log p_\theta(y)$$

Then the risk of a decision function f takes the form

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)] = -\mathbb{E}[\log p_{f(X)}(Y)],$$

where $p_{f(x)}(y)$ is a parameterization of $p(y|x)$.

Maximum likelihood for conditional models as ERM

Given a probabilistic model $p_\theta(y)$, define the loss function ℓ by

$$\ell : (\theta, y) \mapsto -\log p_\theta(y)$$

Then the risk of a decision function f takes the form

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)] = -\mathbb{E}[\log p_{f(X)}(Y)],$$

where $p_{f(x)}(y)$ is a parameterization of $p(y|x)$.

The ERM principle proposes to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i),$$

which is equivalent to the maximum likelihood principle.

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

Fisher discriminant analysis

Generative classification

$X \in \mathbb{R}^p$ and $Y \in \{0, 1\}$.

Generative classification

$X \in \mathbb{R}^p$ and $Y \in \{0, 1\}$. Instead of modeling directly $p(y \mid \mathbf{x})$ model $p(y)$ and $p(\mathbf{x} \mid y)$ and deduce $p(y \mid \mathbf{x})$ using Bayes rule.

Generative classification

$X \in \mathbb{R}^p$ and $Y \in \{0, 1\}$. Instead of modeling directly $p(y \mid \mathbf{x})$ model $p(y)$ and $p(\mathbf{x} \mid y)$ and deduce $p(y \mid \mathbf{x})$ using Bayes rule.
In classification $\mathbb{P}(Y = 1 \mid X = \mathbf{x}) =$

$$\frac{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\,\mathbb{P}(Y = 1)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\,\mathbb{P}(Y = 1) + \mathbb{P}(X = \mathbf{x} \mid Y = 0)\,\mathbb{P}(Y = 0)}$$

Generative classification

$X \in \mathbb{R}^p$ and $Y \in \{0,1\}$. Instead of modeling directly $p(y \mid \mathbf{x})$ model $p(y)$ and $p(\mathbf{x} \mid y)$ and deduce $p(y \mid \mathbf{x})$ using Bayes rule.
In classification $\mathbb{P}(Y = 1 \mid X = \mathbf{x}) =$

$$\frac{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\,\mathbb{P}(Y = 1)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\,\mathbb{P}(Y = 1) + \mathbb{P}(X = \mathbf{x} \mid Y = 0)\,\mathbb{P}(Y = 0)}$$

For example one can assume

- $\mathbb{P}(Y = 1) = \pi$
- $\mathbb{P}(X = \mathbf{x} \mid Y = 1) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- $\mathbb{P}(X = \mathbf{x} \mid Y = 0) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 = \mathbf{\Sigma}$.

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\widehat{\pi}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_0).$$

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\widehat{\pi}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_0).$$

Then we have

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \left(1 + \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\mathbb{P}(Y = 1)}\right)^{-1}$$

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\widehat{\pi}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_0).$$

Then we have

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = \mathbf{x}) &= \left(1 + \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\mathbb{P}(Y = 1)}\right)^{-1} \\
&= \left(1 + \frac{1 - \pi}{\pi} \frac{\exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)\right)}{\exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right)}\right)^{-1}
\end{aligned}$$

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\widehat{\pi}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_0).$$

Then we have

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = \mathbf{x}) &= \left(1 + \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\mathbb{P}(Y = 1)}\right)^{-1} \\
&= \left(1 + \frac{1 - \pi}{\pi} \frac{\exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)\right)}{\exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right)}\right)^{-1} \\
&= \left(1 + \exp\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + b\right)\right)^{-1}
\end{aligned}$$

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\widehat{\pi}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_0).$$

Then we have

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = \mathbf{x}) &= \left(1 + \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\mathbb{P}(Y = 1)}\right)^{-1} \\
&= \left(1 + \frac{1-\pi}{\pi} \frac{\exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0)\right)}{\exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right)}\right)^{-1} \\
&= \left(1 + \exp\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + b\right)\right)^{-1} \\
&= \sigma(\mathbf{w}^\top \mathbf{x} + b)
\end{aligned}$$

with $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ and $b = \log \frac{1-\pi}{\pi} + \frac{1}{2}\boldsymbol{\mu}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$.

LDA *vs* logistic regression

- Both lead to $\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$

LDA *vs* logistic regression

- Both lead to $\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$

Weaknesses of LDA

- Assumes a Gaussian model, which is likely to be quite wrong
- Requires to estimate $p(p+1)/2 + 2p + 1$ parameters vs $p + 1$

LDA *vs* logistic regression

- Both lead to $\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$

Weaknesses of LDA

- Assumes a Gaussian model, which is likely to be quite wrong
- Requires to estimate $p(p+1)/2 + 2p + 1$ parameters vs $p + 1$

Strengths of LDA

- Closed form
- Relevant if the model is a good match to the data.

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

Clustering

Supervised, unsupervised and semi-supervised classification

Supervised learning

Training set composed of pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
$\rightarrow$ Learn to classify new points in the classes

Supervised, unsupervised and semi-supervised classification

Supervised learning

Training set composed of pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
$\rightarrow$ Learn to classify new points in the classes

Unsupervised learning

Training set composed of pairs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
$\rightarrow$ Partition the data in a number of classes.
$\rightarrow$ Possibly produce a decision rule for new points.

Supervised, unsupervised and semi-supervised classification

Supervised learning

Training set composed of pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
$\rightarrow$ Learn to classify new points in the classes

Unsupervised learning

Training set composed of pairs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
$\rightarrow$ Partition the data in a number of classes.
$\rightarrow$ Possibly produce a decision rule for new points.

Transductive learning

Data available at train time composed of
train data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ + test data $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_n\}$
$\rightarrow$ Classify all the test data

Supervised, unsupervised and semi-supervised classification

Supervised learning

Training set composed of pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
$\rightarrow$ Learn to classify new points in the classes

Unsupervised learning

Training set composed of pairs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
$\rightarrow$ Partition the data in a number of classes.
$\rightarrow$ Possibly produce a decision rule for new points.

Transductive learning

Data available at train time composed of
train data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ + test data $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_n\}$
$\rightarrow$ Classify all the test data

Semi-supervised learning

Data available at train time composed of
labelled data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ + unlabelled data $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_n\}$
$\rightarrow$ Produce a classification rule for future points

Clustering

- Clustering is word usually used for unsupervised classification
- Clustering techniques can be useful to solve semi-supervised classification problem.

Clustering

- Clustering is word usually used for unsupervised classification
- Clustering techniques can be useful to solve semi-supervised classification problem.

Clustering is not a well-specified problem

- Classes might be impossible to infer from the distribution of X alone

Clustering

- Clustering is word usually used for unsupervised classification
- Clustering techniques can be useful to solve semi-supervised classification problem.

Clustering is not a well-specified problem

- Classes might be impossible to infer from the distribution of X alone
- Several goals possible:
  - Find the modes of the distribution
  - Find a set of denser **connected** regions supporting most of the density
  - Find a set of denser **convex** regions supporting most of the density
  - Find a set of denser **ellipsoidal** regions supporting most of the density
  - Find a set of denser **round** regions supporting most of the density

K-means

Key assumption: Data composed of K "roundish" clusters of similar sizes with centroids $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

K-means

Key assumption: Data composed of K "roundish" clusters of similar sizes with centroids $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

Problem can be formulated as: $$\min_{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K} \frac{1}{n} \sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2.$$

K-means

Key assumption: Data composed of K "roundish" clusters of similar sizes with centroids $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

Problem can be formulated as: $$\min_{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K} \frac{1}{n} \sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2.$$

Difficult (NP-hard) nonconvex problem.

K-means

Key assumption: Data composed of K "roundish" clusters of similar sizes with centroids $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

Problem can be formulated as: $$\min_{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_K} \frac{1}{n}\sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2.$$

Difficult (NP-hard) nonconvex problem.

K-means algorithm

1. Draw centroids at random
2. Assign each point to the closest centroid

$$C_k \leftarrow \left\{ i \mid \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 = \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \right\}$$

3. Recompute centroid as center of mass of the cluster

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{\mid C_k \mid} \sum_{i \in C_k} \mathbf{x}_i$$

4. Go to 2

K-means properties

Three remarks:

- K-means is greedy algorithm

K-means properties

Three remarks:

- K-means is greedy algorithm
- It can be shown that K-means converges in a finite number of steps.
- The algorithm however typically get stuck in local minima and it practice it is necessary to try several restarts of the algorithm with a random initialization to have chances to obtain a better solution.

K-means properties

Three remarks:

- K-means is greedy algorithm
- It can be shown that K-means converges in a finite number of steps.
- The algorithm however typically get stuck in local minima and it practice it is necessary to try several restarts of the algorithm with a random initialization to have chances to obtain a better solution.
- Will fail if the clusters are not round

Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

The EM algorithm for the Gaussian mixture model

Gaussian mixture model

- K components
- $\mathbf{z}$ component indicator
- $\mathbf{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$

Gaussian mixture model

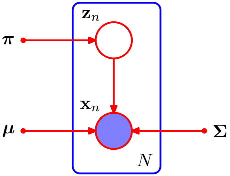
- K components
- $\mathbf{z}$ component indicator
- $\mathbf{z} = (z_1, \ldots, z_K)^\top \in \{0, 1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\mathbf{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

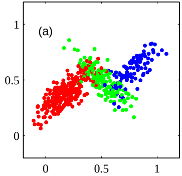
Gaussian mixture model

- K components
- $\mathbf{z}$ component indicator
- $\mathbf{z} = (z_1, \ldots, z_K)^\top \in \{0, 1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\mathbf{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Gaussian mixture model

- K components
- $\mathbf{z}$ component indicator
- $\mathbf{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\mathbf{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Estimation: $\underset{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}{\text{argmax}} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$





Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^{K} z_k = 1\}$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^{K} z_k = 1\}$

$p(\mathbf{x}) =$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^{K} z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z})$$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{z_k} =$$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{z_k} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{z_k} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \left(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1\le k\le K}\right)$ is now complicated

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{k=1}^K \Big[\pi_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big]^{z_k} = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1\le k\le K}\big)$ is now complicated
- No hope to find a simple solution to the maximum likelihood problem

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \prod_{k=1}^K \Big[\pi_k \, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big]^{z_k} = \sum_{k=1}^K \pi_k \, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \leq k \leq K}\big)$ is now complicated
- No hope to find a simple solution to the maximum likelihood problem
- By contrast the complete log-likelihood has a rather simple form:

$$\tilde{\ell}(\theta) =$$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{k=1}^{K} \Big[\pi_k \, \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big]^{z_k} = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1\le k\le K}\big)$ is now complicated
- No hope to find a simple solution to the maximum likelihood problem
- By contrast the complete log-likelihood has a rather simple form:

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$$

Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{k=1}^K \Big[\pi_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big]^{z_k} = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)_{1\le k\le K}\big)$ is now complicated
- No hope to find a simple solution to the maximum likelihood problem
- By contrast the complete log-likelihood has a rather simple form:

$$\tilde{\ell}(\theta) = \sum_{i=1}^M \log p(\mathbf{x}^{(i)},\mathbf{z}^{(i)}) = \sum_{i,\,k} z_k^{(i)} \log \mathcal{N}(x^{(i)};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

Applying maximum likelihood to the multinomial mixture

$\tilde{\ell}(\theta) =$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \le k \le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \le k \le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k^{(i)} = 1 \mid \mathbf{x}; \theta) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \le k \le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k^{(i)} = 1 \mid \mathbf{x}; \theta) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$\rightarrow$ Seems a chicken and egg problem...

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1\le k\le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k^{(i)} = 1 \mid \mathbf{x}; \theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

→ Seems a chicken and egg problem...

- In addition, we want to solve

$$\max_{\theta} \sum_i \log \Bigg(\sum_{\mathbf{z}^{(i)}} p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \Bigg)$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1\le k\le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k^{(i)} = 1 \mid \mathbf{x}; \theta) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$\rightarrow$ Seems a chicken and egg problem...

- In addition, we want to solve

$$\max_{\theta} \sum_i \log \Bigg(\sum_{\mathbf{z}^{(i)}} p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \Bigg) \quad \text{and not} \quad \max_{\substack{\theta,\\ \mathbf{z}^{(1)},\dots,\mathbf{z}^{(M)}}} \sum_i \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \le k \le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k^{(i)} = 1 \mid \mathbf{x}; \theta) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$\rightarrow$ Seems a chicken and egg problem...

- In addition, we want to solve

$$\max_{\theta} \sum_i \log \Big(\sum_{\mathbf{z}^{(i)}} p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \Big) \quad \text{and not} \quad \max_{\substack{\theta, \\ \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}}} \sum_i \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$$

- Can we still use the intuitions above to construct an algorithm maximizing the marginal likelihood?

Principle of the Expectation-Maximization Algorithm

$$\log p(\mathbf{x}; \boldsymbol{\theta}) \quad =$$

Principle of the Expectation-Maximization Algorithm

$$\log p(\mathbf{x}; \boldsymbol{\theta}) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$$

Principle of the Expectation-Maximization Algorithm

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})}$$

Principle of the Expectation-Maximization Algorithm

$$\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})}
\end{aligned}$$

Principle of the Expectation-Maximization Algorithm

$$
\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\quad = \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q)
\end{aligned}
$$

Principle of the Expectation-Maximization Algorithm

$$
\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&= \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})
\end{aligned}
$$

Principle of the Expectation-Maximization Algorithm

$$
\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&= \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})
\end{aligned}
$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$

Principle of the Expectation-Maximization Algorithm

$$\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&= \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})
\end{aligned}$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$
- Moreover: $\boldsymbol{\theta} \mapsto \mathcal{L}(q,\boldsymbol{\theta})$ is a **concave** function.

Principle of the Expectation-Maximization Algorithm

$$\begin{aligned}\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\ &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\ &= \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})\end{aligned}$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$
- Moreover: $\boldsymbol{\theta} \mapsto \mathcal{L}(q,\boldsymbol{\theta})$ is a **concave** function.
- Finally it is possible to show that

$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q||p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

Principle of the Expectation-Maximization Algorithm

$$
\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&= \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})
\end{aligned}
$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$
- Moreover: $\boldsymbol{\theta} \mapsto \mathcal{L}(q,\boldsymbol{\theta})$ is a **concave** function.
- Finally it is possible to show that

$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q||p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

So that if we set $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\boldsymbol{\theta}^{(t)})$ then

$$L(q,\boldsymbol{\theta}^{(t)}) = p(\mathbf{x};\theta^{(t)}).$$

Principle of the Expectation-Maximization Algorithm

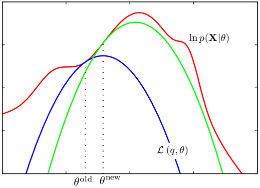
$$
\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\
&= \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})
\end{aligned}
$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$
- Moreover: $\boldsymbol{\theta} \mapsto \mathcal{L}(q,\boldsymbol{\theta})$ is a **concave** function.
- Finally it is possible to show that

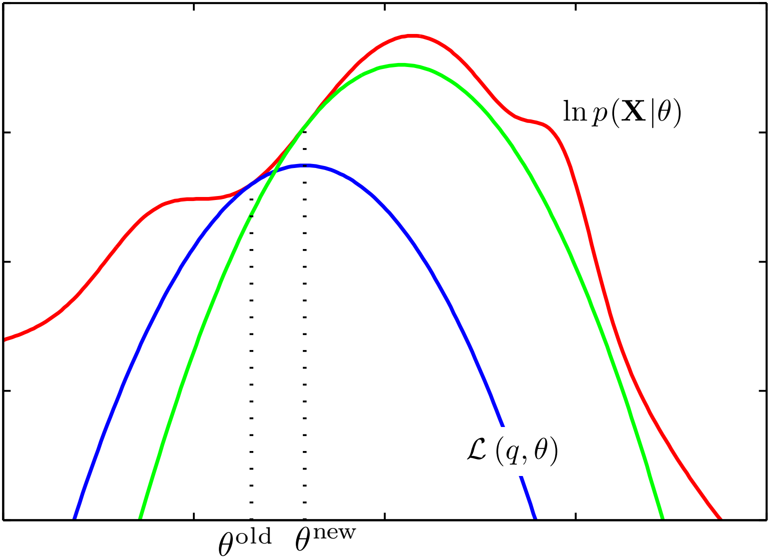
$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q||p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

So that if we set $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\boldsymbol{\theta}^{(t)})$ then

$$L(q,\boldsymbol{\theta}^{(t)}) = p(\mathbf{x};\theta^{(t)}).$$



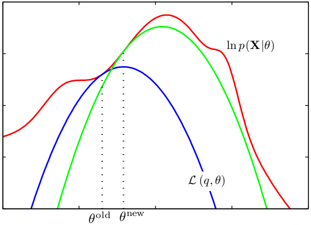
A graphical idea of the EM algorithm



Expectation Maximization algorithm

Expectation step

Maximization step



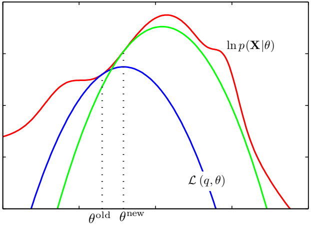
$$\begin{aligned} \boldsymbol{\theta}^{\text{old}} &= \boldsymbol{\theta}^{(t-1)} \\ \boldsymbol{\theta}^{\text{new}} &= \boldsymbol{\theta}^{(t)} \end{aligned}$$

Expectation Maximization algorithm

Expectation step

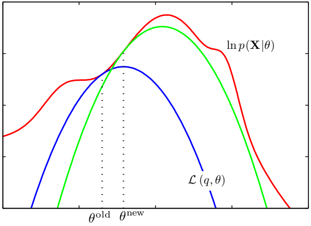
1. $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$

Maximization step



$$\begin{aligned} \boldsymbol{\theta}^{\text{old}} &= \boldsymbol{\theta}^{(t-1)} \\ \boldsymbol{\theta}^{\text{new}} &= \boldsymbol{\theta}^{(t)} \end{aligned}$$

Expectation Maximization algorithm



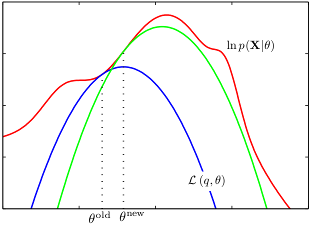
Expectation step

1. $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$
2. $\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \right] + H(q)$

Maximization step

$$\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{(t-1)}$$
$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{(t)}$$

Expectation Maximization algorithm



Expectation step

1. $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$
2. $\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q\left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\right] + H(q)$

Maximization step

1. $\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \mathbb{E}_q\left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\right]$

$$\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{(t-1)}$$
$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{(t)}$$

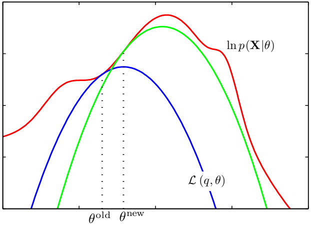
Expectation Maximization algorithm

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

1. $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$
2. $\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \right] + H(q)$



Maximization step

1. $\boldsymbol{\theta}^{(t)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \right]$

$$\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{(t-1)}$$
$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{(t)}$$

ENDWHILE

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] \quad =$$

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] = \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{X}, \mathbf{Z}; \theta)\right]$$

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{X}, \mathbf{Z}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}; \theta)\right]
\end{aligned}
$$

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{X}, \mathbf{Z}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k)\right]
\end{aligned}
$$

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{X}, \mathbf{Z}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k)\right] \\
&= \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log(\pi_k)
\end{aligned}
$$

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{X}, \mathbf{Z}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k)\right] \\
&= \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log(\pi_k) \\
&= \sum_{i,k} q_{ik}^{(t)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} q_{ik}^{(t)} \log(\pi_k)
\end{aligned}
$$

Expectation step for the Gaussian mixture

We computed previously $q_i^{(t)}(\mathbf{z}^{(i)})$, which is a multinomial distribution defined by

$$q_i^{(t)}(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}|\mathbf{x}^{(i)};\theta^{(t-1)})$$

Expectation step for the Gaussian mixture

We computed previously $q_i^{(t)}(\mathbf{z}^{(i)})$, which is a multinomial distribution defined by

$$q_i^{(t)}(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}; \theta^{(t-1)})$$

Abusing notation we will denote $(q_{i1}^{(t)}, \dots, q_{iK}^{(t)})$ the corresponding vector of probabilities defined by

$$q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$$

Expectation step for the Gaussian mixture

We computed previously $q_i^{(t)}(\mathbf{z}^{(i)})$, which is a multinomial distribution defined by

$$q_i^{(t)}(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}|\mathbf{x}^{(i)};\theta^{(t-1)})$$

Abusing notation we will denote $(q_{i1}^{(t)},\dots,q_{iK}^{(t)})$ the corresponding vector of probabilities defined by

$$q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)}=1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$$

$$q_{ik}^{(t)} = p(z_k^{(i)}=1 \mid \mathbf{x}^{(i)};\theta^{(t-1)}) = \frac{\pi_k^{(t-1)}\log\mathcal{N}(\mathbf{x}^{(i)},\boldsymbol{\mu}_k^{(t-1)},\boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^{K}\pi_j^{(t-1)}\log\mathcal{N}(\mathbf{x}^{(i)},\boldsymbol{\mu}_j^{(t-1)},\boldsymbol{\Sigma}_j^{(t-1)})}$$

Maximization step for the Gaussian mixture

$$\left(\boldsymbol{\pi}^t, (\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})_{1\leq k\leq K}\right) = \operatorname*{argmax}_\theta \mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right]$$

Maximization step for the Gaussian mixture

$$\left(\boldsymbol{\pi}^t, (\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})_{1\leq k\leq K}\right) = \underset{\theta}{\text{argmax}}\, \mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right]$$

This yields the updates:

$$\boxed{\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)}\, q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}}, \quad \boxed{\boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}}$$

and $$\boxed{\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}}$$

Final EM algorithm for the Multinomial mixture model

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

$$q_{ik}^{(t)} \leftarrow \frac{\pi_k^{(t-1)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}$$

Maximization step

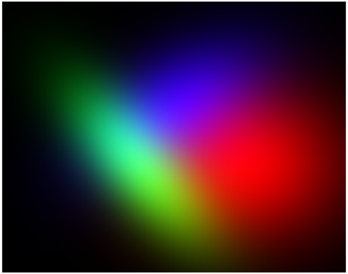
$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)} q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}, \quad \boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}$$

and $$\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}$$

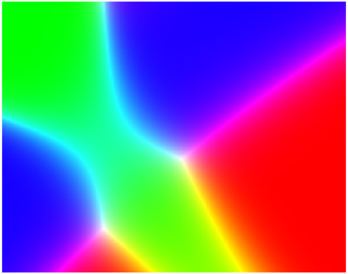
ENDWHILE

EM Algorithm for the Gaussian mixture model III

$p(\mathbf{x}|\mathbf{z})$



$p(\mathbf{z}|\mathbf{x})$



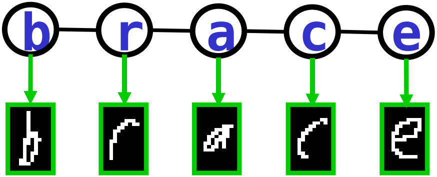
Outline

1. Statistical concepts
2. A short review of convex analysis and optimization
3. The maximum likelihood principle
4. Linear regression
5. Logistic regression
6. Fisher discriminant analysis
7. Clustering
8. The EM algorithm for the Gaussian mixture model
9. Hidden Markov models

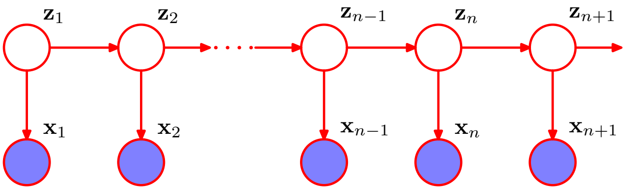
Hidden Markov models

Hidden Markov models (HMM)

- speech recognition
- natural language processing
- OCR
- biological sequences (proteins, DNA)



Hidden Markov Model(HMM)



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \prod_{n=1} p(\mathbf{x}_n | \mathbf{z}_n)$$

Homogeneous Markov chain

- $\mathbf{z}_n \in \{0, 1\}^K$ indicator variable for the state $(1, \ldots, K)$
- Homogeneous Markov chain: $\forall n,\ p(\mathbf{z}_n | \mathbf{z}_{n-1}) = p(\mathbf{z}_2 | \mathbf{z}_1)$
- $\mathbf{x}_n$ emitted symbol $(\{0, 1\}^K)$ / observation $(\mathbb{R}^d)$

Hidden Markov Model (HMM)

Parametrization

distribution of initial state $\quad p(\mathbf{z}_1; \pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$

Hidden Markov Model (HMM)

Parametrization

distribution of initial state	$p(\mathbf{z}_1; \pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$
transition matrix	$p(\mathbf{z}_n \vert \mathbf{z}_{n-1}; A) = \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j}\, z_{nk}}$

Hidden Markov Model (HMM)

Parametrization

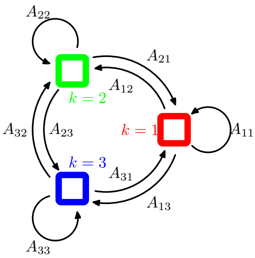
distribution of initial state	$p(\mathbf{z}_1; \pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$
transition matrix	$p(\mathbf{z}_n \mid \mathbf{z}_{n-1}; A) = \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j}\, z_{nk}}$
emission probabilities	$p(\mathbf{x}_n \mid \mathbf{z}_n; \phi)$ e.g. Gaussian Mixture

Hidden Markov Model (HMM)

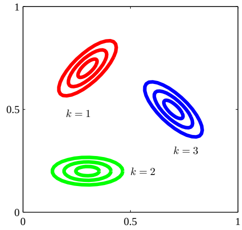
Parametrization

distribution of initial state	$p(\mathbf{z}_1; \pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$
transition matrix	$p(\mathbf{z}_n \mid \mathbf{z}_{n-1}; A) = \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j}\, z_{nk}}$
emission probabilities	$p(\mathbf{x}_n \mid \mathbf{z}_n; \phi)$ e.g. Gaussian Mixture

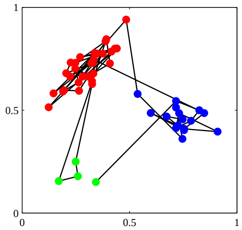
Interpretation



Transistions of $\mathbf{z}_n$



$p(\mathbf{x}_n \mid \mathbf{z}_n)$



Trajectory of $\mathbf{x}_n$

Maximum likelihood for HMMs

Applying the EM algorithm

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t) \qquad \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t)$$

Maximum likelihood for HMMs

Applying the EM algorithm

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t) \qquad \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t)$$

Espectation of the log-likelihood:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n|\phi_k)$$

Maximum likelihood for HMMs

Applying the EM algorithm

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t) \qquad \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t)$$

Espectation of the log-likelihood:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n|\phi_k)$$

When maximizing w.r.t. $\{\pi, A\}$ one obtains

$$\boxed{\pi_k^{t+1} = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}} \qquad \boxed{A_{jk}^{t+1} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}}$$

Maximum likelihood for HMMs

Applying the EM algorithm

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t) \qquad \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t)$$

Espectation of the log-likelihood:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n|\phi_k)$$

When maximizing w.r.t. $\{\pi, A\}$ one obtains

$$\boxed{\pi_k^{t+1} = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}} \qquad \boxed{A_{jk}^{t+1} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}}$$

If the emissions are Gaussians we have as well:

$$\boldsymbol{\mu}_k^{t+1} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad \boldsymbol{\Sigma}_k^{t+1} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

Maximum likelihood for HMMs

Application of the sum-product algorithm

In the context of HMM, the algorithm is known as *forward-backward*.

The following messages are propagated

- forward $\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1})$

Maximum likelihood for HMMs

Application of the sum-product algorithm

In the context of HMM, the algorithm is known as *forward-backward*.

The following messages are propagated

- forward $\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1})$
- backward $\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{z}_{n+1}|\mathbf{z}_n)$

Maximum likelihood for HMMs

Application of the sum-product algorithm

In the context of HMM, the algorithm is known as *forward-backward*.

The following messages are propagated

- forward $\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n)\sum_{\mathbf{z}_{n-1}}\alpha(\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})$
- backward $\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}}\beta(\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n)$

they satisfy the properties:

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n) \qquad \beta(\mathbf{z}_n) = p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)$$

Maximum likelihood for HMMs

Application of the sum-product algorithm

In the context of HMM, the algorithm is known as *forward-backward*.

The following messages are propagated

- forward $\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1})$
- backward $\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{z}_{n+1}|\mathbf{z}_n)$

they satisfy the properties:

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \qquad \beta(\mathbf{z}_n) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)$$

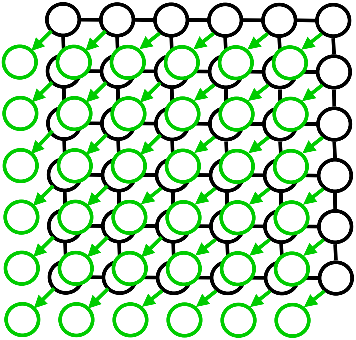
Finally we obtain the marginal probabilities:

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \boldsymbol{\theta}^t) = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X}|\boldsymbol{\theta}^t)}$$

et

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \frac{\alpha(\mathbf{x}_{n-1}) p(\mathbf{x}_n|\mathbf{z}_n) p(\mathbf{z}_n|\mathbf{z}_{n-1}) \beta(\mathbf{x}_n)}{p(\mathbf{X}|\boldsymbol{\theta}^t)}$$

Hidden Markov Field



Original image

Segmentation

Conclusions

Probabilistic models for interpretation

Probabilistic models for combining simple blocks

Probabilistic models for missing data

Probabilistic models for learning parameters and hyperparameters