

## 1 Introduction

## 2 Deux exemples

- Partitionnement de données
- Classification

## 3 Maximum de vraisemblance

- Cas simples
- Cas compliqués
  - Algorithme Espérance Maximisation
  - Application au mélange de modèles
  - Preuve

## 4 Conclusion

# Mélange de modèles probabilistes

## Introduction

**Objectif** : améliorer les performances en combinant des modèles simples pour obtenir des modèles plus complexes.

**Approche** : probabiliste (différent du boosting)

- Hypothèse :
  - modèles simples probabilistes
  - mélange linéaire convexe
- Dérivation : maximiser la vraisemblance
  - des paramètres de chaque modèle simple
  - des coefficients du mélange

**Dans ce cours** :

- Deux usages : partitionnement et classification (régression)
- Deux modèles : Gaussien et régression logistique (régression linéaire)
- Maximisation de la vraisemblance dans des cas compliqués :  
Espérance-Maximisation (EM)

## 1 Introduction

## 2 Deux exemples

- Partitionnement de données
- Classification

## 3 Maximum de vraisemblance

- Cas simples
- Cas compliqués
  - Algorithme Espérance Maximisation
  - Application au mélange de modèles
  - Preuve

## 4 Conclusion

- 1 Introduction
  
- 2 Deux exemples
  - Partitionnement de données
  - Classification
  
- 3 Maximum de vraisemblance
  - Cas simples
  - Cas compliqués
    - Algorithme Espérance Maximisation
    - Application au mélange de modèles
    - Preuve
  
- 4 Conclusion

# Partitionnement de données

## Mélange de Gaussiennes

### Loi normale (densité)

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

$$\nu_{\mu, \Sigma}(\mathbf{x}) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{d/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

- $d$  : dimension
- $\mu$ ,  $\boldsymbol{\mu}$  : moyenne
- $\sigma^2$ ,  $\Sigma$  : variance

# Partitionnement de données

## Mélange de Gaussiennes

Partitionnement de données :

- Soit  $N$  individus  $X = \{\mathbf{x}_i | i = 1..N\}$ ,
- chacun décrit par  $d$  attributs,
- supposé issu de  $K$  groupes.

Hypothèse : modèle gaussien

- Les individus de chaque groupe sont des échantillons d'une gaussienne de paramètre  $\theta_k = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ . On note  $\Theta = \{\theta_k\}_{k=1..K}$
- Les proportions des groupes sont données par  $\boldsymbol{\pi} = \{\pi_k\}_{k=1..K}$
- Donc, les individus  $\mathbf{x}$  de  $X$  sont tirés selon le mélange de lois

## Mélange de Gaussiennes (densité)

$$f_{\Phi}(\mathbf{x}) = \sum_{k=1}^K \pi_k \nu_{\theta_k}(\mathbf{x}),$$

où  $\Phi = \{\Theta, \boldsymbol{\pi}\}$  et  $\nu_{\theta_k}$  est la densité d'une gaussienne.

- 1 Introduction
  
- 2 Deux exemples
  - Partitionnement de données
  - Classification
  
- 3 Maximum de vraisemblance
  - Cas simples
  - Cas compliqués
    - Algorithme Espérance Maximisation
    - Application au mélange de modèles
    - Preuve
  
- 4 Conclusion

# Classification

## Mélange de régressions logistiques

Régression logistique pour la classification (deux classes) :

Vraisemblance  $\mathbb{P}(Y|\mathbf{x}, \theta)$  du modèle régression logistique

- $\mathbb{P}(Y = 1|\mathbf{x}, \theta) = \text{sig}(\theta^T \mathbf{x})$
- $\mathbb{P}(Y = 0|\mathbf{x}, \theta) = 1 - \mathbb{P}(Y = 1|\mathbf{x}) = 1 - \text{sig}(\theta^T \mathbf{x})$
- $\mathbb{P}(Y = y|\mathbf{x}, \theta) = \text{sig}(\theta^T \mathbf{x})^y (1 - \text{sig}(\theta^T \mathbf{x}))^{1-y}$

- $\text{sig}(x) = \frac{1}{1+e^{-x}}$  : fonction sigmoïde
- $\theta$  : vecteur de paramètres du modèle log-linéaire
- Règle pour classifier (fonction de coût 0-1) :
  - Si  $\frac{\mathbb{P}(Y=1|\mathbf{x},\theta)}{\mathbb{P}(Y=0|\mathbf{x},\theta)} = \frac{\text{sig}(\theta^T \mathbf{x})}{1-\text{sig}(\theta^T \mathbf{x})} \geq 1$  alors  $\hat{y} = 1$
  - Si  $\frac{\mathbb{P}(Y=0|\mathbf{x},\theta)}{\mathbb{P}(Y=1|\mathbf{x},\theta)} = \frac{\text{sig}(\theta^T \mathbf{x})}{1-\text{sig}(\theta^T \mathbf{x})} < 1$  alors  $\hat{y} = 0$



# Classification

## Mélange de régressions logistiques

- La régression logistique ne fonctionne que pour des classes linéairement séparables.
- Idée :
  - Des sous groupes d'individus sont linéairement séparables.
  - Pour chaque groupe  $k$ , on apprend les paramètres  $\theta_k$  de la régression logistique.
  - On combine linéairement les classifieurs appris.

Vraisemblance  $\mathbb{P}(Y|\mathbf{x}, \Phi)$  du mélange de régressions logistiques

$$\mathbb{P}(Y = y|\mathbf{x}, \Phi) = \sum_{k=1}^K \pi_k \text{sig}(\theta_k^T \mathbf{x})^y (1 - \text{sig}(\theta_k^T \mathbf{x}))^{1-y},$$

avec  $\Phi = \{\Theta, \Pi\}$ ,  $\Theta = \{\theta_k\}_{k=1..K}$ ,  $\Pi = \{\pi_k\}_{k=1..K}$

- 1 Introduction
  
- 2 Deux exemples
  - Partitionnement de données
  - Classification
  
- 3 Maximum de vraisemblance
  - Cas simples
  - Cas compliqués
    - Algorithme Espérance Maximisation
    - Application au mélange de modèles
    - Preuve
  
- 4 Conclusion

# Vraisemblance

## Définition

### Log-vraisemblance $\mathcal{L}(X; \theta)$

Soit  $\theta$  les paramètres du modèle et  $X = \{\mathbf{x}_i\}_{i=1..N}$  des échantillons i.i.d.,

- Cas discret :

$$\mathcal{L}(X; \theta) = \log(\mathbb{P}(X|\theta)) \stackrel{\text{i.i.d.}}{=} \sum_{i=1}^N \log(\mathbb{P}(\mathbf{x}_i|\theta))$$

- Cas continue :

$$\mathcal{L}(X; \theta) = \log(f_\theta(X)) \stackrel{\text{i.i.d.}}{=} \sum_{i=1}^N \log(f_\theta(\mathbf{x}_i))$$

## 1 Introduction

## 2 Deux exemples

- Partitionnement de données
- Classification

## 3 Maximum de vraisemblance

- Cas simples
- Cas compliqués
  - Algorithme Espérance Maximisation
  - Application au mélange de modèles
  - Preuve

## 4 Conclusion

# Loi normale : maximum de vraisemblance I

## Annulation du gradient

- Pour la loi normale, dériver et annuler la log-vraisemblance « suffit » à retrouver les paramètres.

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu}_{ML})(\mathbf{x}_i - \boldsymbol{\mu}_{ML})^\top$$

- Démonstration ( $d = 1$ ) : on écrit la log-vraisemblance,

$$\mathcal{L}(X; \mu, \sigma) = \sum_{i=1}^N \log(\nu_{\mu, \sigma}(x_i))$$

$$= -N \log(\sigma) - \frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^N \left( \frac{x_i - \mu}{\sigma} \right)^2$$

# Loi normale : maximum de vraisemblance II

## Annulation du gradient

- Les dérivés partielles s'écrivent,

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu)$$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^N (x_i - \mu)^2$$

- On égalise à zéro :

$$\sum_{i=1}^N (x_i - \mu) = 0 \Rightarrow \mu = \frac{1}{N} \sum_{i=1}^N x_i$$
$$-N + \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 = 0 \Rightarrow \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

# Régression logistique : maximum de vraisemblance

Descente de gradient (rappel)

- On écrit la log-vraisemblance,

$$\mathcal{L}(Y; X, \theta) = \sum_{i=1}^N y_i \log(\text{sig}(\theta^T \mathbf{x}_i)) + (1 - y_i) \log(1 - \text{sig}(\theta^T \mathbf{x}_i))$$

- Calcul du gradient,

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^N \mathbf{x}_i (y_i - \text{sig}(\theta^T \mathbf{x}_i))$$

- Monté de gradient pour trouver le maximum,

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \frac{\partial \mathcal{L}}{\partial \theta} (\theta^{(t)})$$

## 1 Introduction

## 2 Deux exemples

- Partitionnement de données
- Classification

## 3 Maximum de vraisemblance

- Cas simples
- Cas compliqués
  - Algorithme Espérance Maximisation
  - Application au mélange de modèles
  - Preuve

## 4 Conclusion



# Espérance Maximisation I

Introduction de variables latentes et décomposition de la vraisemblance

**Problème** : maximiser la vraisemblance  $\mathcal{L}(X; \theta)$  de modèles complexes peut s'avérer très compliquée.

**Idée** : introduire de nouvelles variables  $Z$ , dites **cachées** ou **latentes**, qui vont venir **compléter** les observations  $X$  tel que leur connaissance facilite la maximisation de la vraisemblance.

- On note  $\Phi$  les paramètres de la distribution de  $X$
- En utilisant la règle de Bayes,

$$\mathbb{P}(X|\Phi) = \mathbb{P}(X|Z, \Phi) = \frac{\mathbb{P}(X, Z|\Phi)}{\mathbb{P}(Z|X, \Phi)}$$

- On appelle  $\mathcal{L}(X, Z; \Phi)$  la vraisemblance complétée, on a

$$\mathcal{L}(X; \Phi) = \mathcal{L}(X, Z; \Phi) - \mathcal{L}(Z; \Phi, X)$$

- L'algorithme EM fonctionne de façon itérative. On note  $\Phi^{(t)}$  les paramètres à l'étape  $t$ .

# Espérance Maximisation II

Introduction de variables latentes et décomposition de la vraisemblance

- Prenons l'espérance sur  $Z$  conditionnellement à  $\Phi^{(t)}$  et  $X$ ,

$$\mathcal{L}(X; \Phi) = \overbrace{\mathbb{E}_Z \left[ \mathcal{L}(X, Z; \Phi) \middle| \Phi^{(t)}, X \right]}^{Q(\Phi, \Phi^{(t)})} - \overbrace{\mathbb{E}_Z \left[ \mathcal{L}(Z; \Phi, X) \middle| \Phi^{(t)}, X \right]}^{H(\Phi, \Phi^{(t)})}$$

- Connaissant les paramètres à l'étape  $t$ , on cherche  $\Phi^{(t+1)}$  qui maximise  $\mathcal{L}(X, \Phi^{(t+1)})$ .
- On peut montrer qu'il suffit de maximiser  $Q(\Phi^{(t+1)}, \Phi^{(t)})$ .
- Cette maximisation se fait en deux étapes.

# Espérance Maximisation

## Étape E, étape M

- L'étape E, consiste à calculer de l'expression

$$Q(\Phi, \Phi^{(t)}) = \mathbb{E}_Z \left[ \mathcal{L}(X, Z; \Phi) \middle| \Phi^{(t)}, X \right]$$

C'est à dire à calculer l'espérance sur  $Z$  conditionnellement à  $\Phi^{(t)}, X$ .

- Dans l'étape M, on cherche le prochain ensemble de paramètres  $\Phi^{(t+1)}$  tel que

$$\Phi^{(t+1)} = \arg \max_{\Phi} Q(\Phi, \Phi^{(t)})$$

# Mélange de Gaussiennes I

## Étape E

- Pour le mélange de Gaussienne, on a

$$\mathbb{P}(\mathbf{x}_i | \Phi) = \sum_{k=1}^K \pi_k \nu_{\theta_k}(\mathbf{x}_i),$$

- On introduit les variables aléatoires  $z_{ik}$  qui valent 1 si l'individu  $\mathbf{x}_i$  est généré par la  $k$ -ième Gaussienne, 0 sinon. On pose,

$$\mathbb{P}(\mathbf{x}_i, z_{ik} | \Phi) = \sum_{k=1}^K (\pi_k \nu_{\theta_k}(\mathbf{x}_i))^{z_{ik}}$$

qui vérifie  $\sum_{z_{ik} \in \{0,1\}} \mathbb{P}(\mathbf{x}_i, z_{ik} | \Phi) = \mathbb{P}(\mathbf{x}_i | \Phi)$

- On écrit la log-vraisemblance des données complétées,

$$\mathcal{L}(X, Z; \Phi) = \sum_{i=1}^N \sum_{k=1}^K z_{ik} \log(\pi_k \nu_{\theta_k}(\mathbf{x}_i)),$$

# Mélange de Gaussiennes II

## Étape E

- Ainsi

$$\begin{aligned} Q(\Phi, \Phi^{(t)}) &= \mathbb{E}_Z \left[ \sum_{i=1}^N \sum_{k=1}^K z_{ik} \log(\pi_k \nu_{\theta_k}(\mathbf{x}_i)) \middle| \Phi^{(t)}, \mathbf{X} \right] \\ &= \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}_Z \left[ z_{ik} \middle| \Phi^{(t)}, \mathbf{x}_i \right] \log(\pi_k \nu_{\theta_k}(\mathbf{x}_i)) \end{aligned}$$

- Notons  $\gamma_{ik} = \mathbb{E}_Z [z_{ik} | \Phi^{(t)}, \mathbf{x}_i]$ , on a

$$\begin{aligned} \gamma_{ik} &= \mathbb{P}(z_{ik} = 1 | \Phi^{(t)}, \mathbf{x}_i) = \frac{\mathbb{P}(z_{ik} = 1, \mathbf{x}_i | \Phi^{(t)})}{\mathbb{P}(\mathbf{x}_i | \Phi^{(t)})} \\ &= \frac{\pi_k \nu_{\theta_k}(\mathbf{x}_i)}{\sum_{k'=1}^K \pi_{k'} \nu_{\theta_{k'}}(\mathbf{x}_i)} \end{aligned}$$

- Intuition : probabilité d'être généré par le  $k$ -ième modèle divisé normalisé par la probabilité d'être généré par l'ensemble (le mélange) des modèles.

# Mélange de Gaussiennes I

## Étape M

- Dans l'étape M, on cherche

$$\Phi^{(t+1)} = \arg \max_{\Phi} \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} \log (\pi_k \nu_{\theta_k}(\mathbf{x}_i))$$

avec  $\Phi = \{\Pi, \Theta\}$  sous la contrainte  $\sum_k \pi_k = 1$

- On utilise le lagrangien,

$$L(\Pi, \Theta, \lambda) = \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} \log (\pi_k \nu_{\theta_k}(\mathbf{x}_i)) - \lambda \left( \sum_k \pi_k - 1 \right)$$

- On dérive par rapport à chaque composante et on annule.
- Commençons par  $\pi_k$ ,

$$\frac{\partial L}{\partial \pi_k} = \frac{\sum_{i=1}^N \gamma_{ik}}{\pi_k} - \lambda = 0$$

# Mélange de Gaussiennes II

## Étape M

- Ainsi, pour tout  $k$ ,  $\sum_{i=1}^N \gamma_{ik} = \lambda \pi_k$
- On somme sur  $k$  et on réinjecte la contrainte pour trouver,

$$N = \sum_{k=1}^K \sum_{i=1}^N \gamma_{ik} = \sum_{k=1}^K \lambda \pi_k = \lambda$$

- Finalement,

$$\pi_k = \frac{1}{N} \sum_{i=1}^N \gamma_{ik}$$

- Passons à  $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ , on a,

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \frac{\partial \mathcal{L}_k(X; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \quad \text{et} \quad \frac{\partial L}{\partial \boldsymbol{\Sigma}_k} = \frac{\partial \mathcal{L}_k(X; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k}$$

avec,

$$\mathcal{L}_k(X; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{1}{2} \log(|\boldsymbol{\Sigma}|) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^N \gamma_{ik} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

# Mélange de Gaussiennes III

## Étape M

- Similairement au cas avec une seule Gaussienne, on trouve

$$\mu_k = \frac{\sum_{i=1}^N \gamma_{ik} \mathbf{x}_i}{\sum_{i=1}^N \gamma_{ik}}$$
$$\Sigma_k = \frac{\sum_{i=1}^N \gamma_{ik} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{\sum_{i=1}^N \gamma_{ik}}$$

- Intuition : chaque modèle est appris avec l'ensemble des individus mais pondérés par leur coefficient d'appartenance à ce modèle.



# Mélange de régression logistiques I

## Étape E

- Pour le mélange de régressions logistiques, on a

$$\mathbb{P}(y_i | \mathbf{x}_i, \Phi) = \sum_{k=1}^K \pi_k \text{sig}(\theta_k^T \mathbf{x}_i)^{y_i} (1 - \text{sig}(\theta_k^T \mathbf{x}_i))^{1-y_i}$$

- On introduit les variables aléatoires  $z_{ik}$  qui valent 1 si l'individu  $\mathbf{x}_i$  est appartient au groupe  $k$ , 0 sinon. On pose,

$$\mathbb{P}(y_i | \mathbf{x}_i, \Phi) = \sum_{k=1}^K \left( \pi_k \text{sig}(\theta_k^T \mathbf{x}_i)^{y_i} (1 - \text{sig}(\theta_k^T \mathbf{x}_i))^{1-y_i} \right)^{z_{ik}}$$

qui vérifie  $\sum_{z_{ik} \in \{0,1\}} \mathbb{P}(\mathbf{x}_i, z_{ik} | \Phi) = \mathbb{P}(\mathbf{x}_i | \Phi)$

# Mélange de régression logistiques II

## Étape E

- On écrit la log-vraisemblance,

$$\mathcal{L}(Y, Z; X, \Phi) = \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left( \log(\pi_k) + y_i \log \left( \text{sig}(\theta_k^T \mathbf{x}_i) \right) \right. \\ \left. + (1 - y_i) \log \left( 1 - \text{sig}(\theta_k^T \mathbf{x}_i) \right) \right)$$

- Ainsi

$$Q(\Phi, \Phi^{(t)}) = \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}_Z \left[ z_{ik} \mid \Phi^{(t)}, \mathbf{x}_i, y_i \right] \left( \log(\pi_k) + y_i \log \left( \text{sig}(\theta_k^T \mathbf{x}_i) \right) \right. \\ \left. + (1 - y_i) \log \left( 1 - \text{sig}(\theta_k^T \mathbf{x}_i) \right) \right)$$

- On note  $\gamma_{ik} = \mathbb{E}_Z [z_{ik} | \Phi^{(t)}, \mathbf{x}_i, y_i]$ , on a

$$\begin{aligned}\gamma_{ik} &= \mathbb{P} \left( z_{ik} = 1 \mid \Phi^{(t)}, \mathbf{x}_i, y_i \right) = \frac{\mathbb{P} \left( z_{ik} = 1, y_i \mid \Phi^{(t)}, \mathbf{x}_i \right)}{\mathbb{P} \left( y_i \mid \Phi^{(t)}, \mathbf{x}_i \right)} \\ &= \frac{\pi_k \operatorname{sig}(\theta_k^T \mathbf{x}_i)^{y_i} (1 - \operatorname{sig}(\theta_k^T \mathbf{x}_i))^{1-y_i}}{\sum_{k'=1}^K \pi_{k'} \operatorname{sig}(\theta_{k'}^T \mathbf{x}_i)^{y_i} (1 - \operatorname{sig}(\theta_{k'}^T \mathbf{x}_i))^{1-y_i}}\end{aligned}$$

- Intuition : probabilité d'être généré par le  $k$ -ième modèle divisé normalisé par la probabilité d'être généré par l'ensemble (le mélange) des modèles.

# Mélange de régressions logistiques I

## Étape M

- Dans l'étape M, on cherche

$$\Phi^{(t+1)} = \arg \max_{\Phi} \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} \left( \log(\pi_k) + y_i \log \left( \text{sig}(\theta_k^T \mathbf{x}_i) \right) + (1 - y_i) \log \left( 1 - \text{sig}(\theta_k^T \mathbf{x}_i) \right) \right)$$

avec  $\Phi = \{\Pi, \Theta\}$  sous la contrainte  $\sum_k \pi_k = 1$

- On utilise le lagrangien  $L(\Pi, \Theta, \lambda)$ , que l'on dérive par rapport à chaque composante et égalise à zéros.
- Comme pour le mélange de Gaussiennes, on a

$$\pi_k = \frac{1}{N} \sum_{i=1}^N \gamma_{ik}$$

# Mélange de régressions logistiques II

## Étape M

- Passons à  $\theta_k = \{\mu_k, \Sigma_k\}$ . La résolution se fait par descente de gradient,

$$\frac{\partial L}{\partial \theta_k} = \sum_{i=1}^N \gamma_{ik} \mathbf{x}_i \left( y_i - \text{sig}(\theta_k^\top \mathbf{x}_i) \right)$$

$$\theta_k^{(t'+1)} = \theta_k^{(t')} + \alpha_t \frac{\partial L}{\partial \theta_k} \left( \theta_k^{(t')} \right)$$

- Intuition : le calcul du gradient pour un modèle se fait à l'aide de tous les individus mais pondérés par leur coefficient d'appartenance à ce modèle.

# Espérance Maximisation I

## Preuve de convergence

- Pour prouver la convergence, on va montrer que chaque itération de EM augmente la log-vraisemblance.
- On calcule l'incrément à chaque pas,

$$\begin{aligned}\mathcal{L}(X; \Phi^{(t)}) &= Q(\Phi^{(t)}, \Phi^{(t)}) - H(\Phi^{(t)}, \Phi^{(t)}) \\ \mathcal{L}(X; \Phi^{(t+1)}) &= Q(\Phi^{(t+1)}, \Phi^{(t)}) - H(\Phi^{(t+1)}, \Phi^{(t)}) \\ \mathcal{L}(X; \Phi^{(t+1)}) - \mathcal{L}(X; \Phi^{(t)}) &= \left( Q(\Phi^{(t+1)}, \Phi^{(t)}) - Q(\Phi^{(t)}, \Phi^{(t)}) \right) \\ &\quad - \left( H(\Phi^{(t+1)}, \Phi^{(t)}) - H(\Phi^{(t)}, \Phi^{(t)}) \right)\end{aligned}$$

- On montre avec l'inégalité de Jensen que

$$H(\Phi^{(t+1)}, \Phi^{(t)}) - H(\Phi^{(t)}, \Phi^{(t)}) \leq 0$$

- De plus, l'algorithme EM assure que

$$Q(\Phi^{(t+1)}, \Phi^{(t)}) \geq Q(\Phi^{(t)}, \Phi^{(t)})$$

# Espérance Maximisation II

## Preuve de convergence

- Donc,

$$\mathcal{L}(X; \Phi^{(t+1)}) - \mathcal{L}(X; \Phi^{(t)}) \geq 0$$

- Comme la log-vraisemblance est majorée et qu'à chaque pas  $\mathcal{L}(X; \Phi^{(t)})$  augmente, l'algorithme converge.

- 1 Introduction
  
- 2 Deux exemples
  - Partitionnement de données
  - Classification
  
- 3 Maximum de vraisemblance
  - Cas simples
  - Cas compliqués
    - Algorithme Espérance Maximisation
    - Application au mélange de modèles
    - Preuve
  
- 4 Conclusion



## Quelques remarques

- L'algorithme EM est très utilisé, a de nombreuses variantes et s'applique a une très large variété de modèle.
- Le mélange de Gaussiennes est équivalent au  $K$ -moyennes quand les variances tendent vers 0.
- L'algorithme EM peut diverger en cas de singularités dans la fonction de vraisemblance. Ce problème peut être réglé par traitement bayésien.
- Une approche bayésienne permet aussi de sélectionner automatiquement le nombre de composantes dans le mélange.