L'algorithme "Expectation-Maximization"

J.-B. Gasnier [1]

Machine Learning Group - 15 mars 2017

1. Mines ParisTech, PSL, C.M.M., Fontainebleau

Rappel et contexte

Cadre du problème

- On dispose de n observations $(X_1, ..., X_n)$ (i.i.d.) d'une v.a. X
- On cherche une loi P permettant de modéliser X

Quelques définitions

- P_θ : la loi P de paramètre θ
- $L_\theta(x_i) = L(x_i; \theta) = L(x_i|\theta)$: densité de X_i
- $L_n(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} L(x_i|\theta)$: Vraisemblance de l'échantillon pour la loi P_θ

Maximum de vraisemblance

Notre but ultime

- $L_n(x_1, ..., x_n; \theta)$ permet de quantifier à quel point il est probable que notre observation découle de la loi P_θ
- On ne connait pas θ, c'est notre but.

Nouvelle définition (encore)

- On définit un estimateur de θ au sens du Maximum de Vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} L_n(x_1, ..., x_n; \theta)$$

- La vraisemblance est un produit. Better maximiser une somme :

$$log(L_n(x_1, ..., x_n; \theta)) = \sum_{i=1}^{n} log(L(x_i|\theta))$$

Exemple avec le cas d'une loi binomiale

- On lance 30 fois une pièce : $22\times$ pile ; $8\times$ faces
- X le nombre de pile obtenus sur 30 lancés
- P est une loi Binomiale $B(30, p)$
- $\theta = p$ est notre paramètre à déterminer

Solution (cf. papier)

- On peut démontrer que $p = \frac{22}{30}$ est le meilleur paramètre, au sens MV
- Maximisation de la log-Vraisemblance
- Critère sur la dérivée de la log-Vraisemblance

L'algorithme EM

Pourquoi un algorithme ?

- Il n'est pas toujours possible de maximiser la log-vraisemblance
- Algorithme permet de procéder par étapes successives pour améliorer les estimations de θ

Cadre de l'algorithme

- Maximisation de $P(X|\theta)$ impossible
- On rajoute des données cachées $Z = (z_1, ..., z_n)$ t.q. si on connaissait Z, on pourrait maximiser $P(X, Z|\theta)$. (oui, ça parait abstrait dit comme ça... Et c'est normal.)
- Avec ces randoms données Z et un premier vecteur de paramètre θ_m, on va pouvoir commencer notre identification de $\hat{\theta}$...

Etape E : Expectation

- En pratique, on ne connait pas Z
- On n'a donc pas accès à la vraisemblance des données...
- On estime cette vraisemblance naturellement via :

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}\left[\, log(P(\mathbf{X}, \mathbf{z}|\theta))\,\right]$$

Interprétation

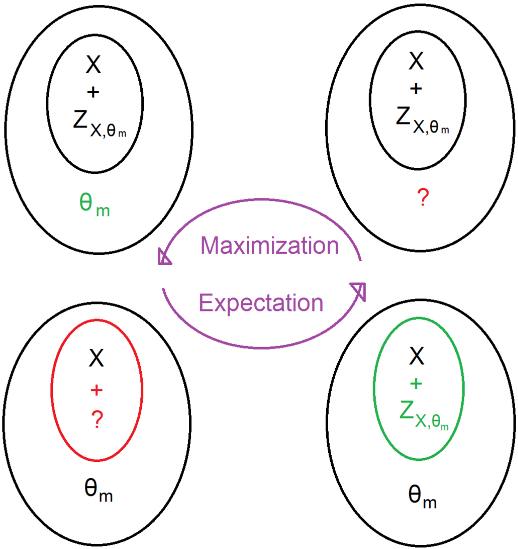
- Avec notre paramètre actuel θ_m, on trouve les données Z les plus probables.
- Grace à ça, on a enfin un jeu de donnée complet.
- On regarde la vraisemblance de telles données pour une loi P_θ en général.

Etape M : Maximisation

- On détermine un nouveau paramètre θ_{m+1} qui maximise la vraisemblance pour de telles données

Et on itère !

L'algorithme consiste à alterner ces deux phases
- Phase E : Raffiner les données manquantes
- Phase M : Raffiner les paramètres estimés du modèle qu'on cherche à fitter.



Un joli dessin pour visualiser ce qu'il se passe !

Demonstration de la croissance de vraisemblance, ou Pourquoi ça marche ? (Attention, ça va piquer)

Pourquoi ça marche ?

- Tout repose sur l'inégalité de Jensen :

Si
1. f convexe sur un intervalle I
2. $(x_i)_{i \in [\![1,n]\!]} \in I^n$
3. $(\lambda_i)_{i \in [\![1,n]\!]}$ t.q. $\sum_{i=1}^{n} \lambda_i = 1$

Alors $\quad f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$

Des définitions (Yeah !)

- Supposons qu'on soit à l'étape m de l'algorithme. On connait θ_m. On cherche un meilleur θ.
- $\Delta(\theta, \theta_m) = log(P(\mathbf{X}|\theta)) - log(P(\mathbf{X}|\theta_m))$
- $\Delta(\theta, \theta_m)$: différence de vraisemblance entre P_θ et P_{θ_m}
- On veut θ t.q. $\Delta(\theta, \theta_m) \geq 0$

- On ne connait toujours pas $P(\mathbf{X}|\theta)$...
Damned.
Oh wait !

MORE DEFINITIONS

- Cherchons $\delta(\theta, \theta_m)$ une fonction telle que :

$$\begin{array}{rcl} \Delta(\theta, \theta_m) & \geq & \delta(\theta|\theta_m) \quad \forall\theta \\ \delta(\theta_m|\theta_m) & = & 0 \end{array}$$

- Si une telle fonction a un maximum, il sera positif ou nul
- Maximiser δ revient donc à assurer que $\Delta \geq 0$
- On assure ainsi la croissance de la vraisemblance si on trouve un nouveau θ qui maximise δ

Démonstration : 1^e étape - minorer $\Delta(\theta_,\theta_m)$

$\Delta(\theta, \theta_m) = log(P(\mathbf{X}|\theta)) - log(P(\mathbf{X}|\theta_m))$

Démonstration : 1^e étape - minorer $\Delta(\theta,\theta_m)$

$\Delta(\theta, \theta_m) = log(P(\mathbf{X}|\theta)) - log(P(\mathbf{X}|\theta_m))$

$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - 1\times log(P(\mathbf{X}|\theta_m))$

Démonstration : 1^e étape - minorer $\Delta(\theta_,\theta_m)$

$$\Delta(\theta, \theta_m) = log(P(\mathbf{X}|\theta)) - log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - 1 \times log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - \underbrace{\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m)}_{=1} log(P(\mathbf{X}|\theta_m))$$

Démonstration : 1^e étape - minorer $\Delta(\theta_, \theta_m)$

$$\Delta(\theta, \theta_m) = log(P(\mathbf{X}|\theta)) - log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - 1 \times log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - \underbrace{\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m)}_{=1} log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z \frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)} \cdot P(\mathbf{z}|\mathbf{X}, \theta_m)) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$$

Démonstration : 1^e étape - minorer $\Delta(\theta_, \theta_m)$

$$\Delta(\theta, \theta_m) = log(P(\mathbf{X}|\theta)) - log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - 1 \times log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)) - \underbrace{\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m)}_{=1} log(P(\mathbf{X}|\theta_m))$$

$$= log(\sum_z \frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)} \cdot P(\mathbf{z}|\mathbf{X}, \theta_m)) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$$

$$\geqslant \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)}) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$$

Démonstration : 2^e étape - récrire le minorant

$\Delta(\theta, \theta_m) \geqslant$
$\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)}) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$

Démonstration : 2^e étape - récrire le minorant

$$\Delta(\theta, \theta_m) \geqslant$$
$$\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta_m)}) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$$

$$= \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta_m)P(\mathbf{X}|\theta_m)})$$

Démonstration : 2^e étape - récrire le minorant

$\Delta(\theta, \theta_m) \geqslant$
$\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)}) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$

$= \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)P(\mathbf{X}|\theta_m)})$

$= \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X},\mathbf{z}|\theta)}{P(\mathbf{X},\mathbf{z}|\theta_m)})$

Démonstration : 2^e étape - récrire le minorant

$\Delta(\theta, \theta_m) \geqslant$
$\sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)}) - \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(P(\mathbf{X}|\theta_m))$

$= \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X}|\mathbf{z},\theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X},\theta_m)P(\mathbf{X}|\theta_m)})$

$= \sum_z P(\mathbf{z}|\mathbf{X}, \theta_m) log(\frac{P(\mathbf{X},\mathbf{z}|\theta)}{P(\mathbf{X},\mathbf{z}|\theta_m)})$

$\hat{=}\, \delta(\theta|\theta_m)$
On a bien : $\Delta(\theta, \theta_m) \geqslant \delta(\theta|\theta_m)$ (par construction) et il est évident que $\delta(\theta_m|\theta_m) = 0$

On définit donc logiquement :

$$\theta_{m+1} = \arg\max_{\theta} \delta(\theta|\theta_m)$$

Et on est assuré que le nouveau paramètre θ_{m+1} est plus vraisemblable que θ_m.

Mise en pratique

Problème des deux pièces

- On dispose de deux pièces ayant respectivement des probabilités p_1 et p_2 de faire pile.
- Le joueur choisit une pièce au hasard.
- Il fait trois lancer.
- Il réitère l'expérience autant de fois qu'on le veut (par exemple, 5 fois).
- On n'enregistre que la liste des résultats, mais pas la pièce utilisée. Dans notre exemple, ça donnerait :

$$\{P, P, F\}, \{F, P, F\}, \{P, F, F\}, \{F, F, P\}, \{P, F, F\}$$

- Question : déterminer p_1 et p_2.
- Donnée manquante : quelle pièce a généré quel lancer ?

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.
- On jouera n tours de jeu.

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.
- On jouera n tours de jeu.
- $(Z_i)_{1 \leqslant i \leqslant n}$ une suite de V.A. i.i.d. suivant la loi de Bernouilli $B(\lambda)$. Si $Z_i = 1$, le joueur utilisera la pièce 1 au i-ème tour. Sinon, si $Z_i = 2$, il utilisera la pièce 2.

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.
- On jouera n tours de jeu.
- $(Z_i)_{1 \leqslant i \leqslant n}$ une suite de V.A. i.i.d. suivant la loi de Bernouilli $B(\lambda)$. Si $Z_i = 1$, le joueur utilisera la pièce 1 au i-ème tour. Sinon, si $Z_i = 2$, il utilisera la pièce 2.
- La pièce 1 a une probabilité p_1 d'obtenir pile.
- La pièce 2 a une probabilité p_2 d'obtenir pile.

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.
- On jouera n tours de jeu.
- $(Z_i)_{1 \leqslant i \leqslant n}$ une suite de V.A. i.i.d. suivant la loi de Bernouilli $B(\lambda)$. Si $Z_i = 1$, le joueur utilisera la pièce 1 au i-ème tour. Sinon, si $Z_i = 2$, il utilisera la pièce 2.
- La pièce 1 a une probabilité p_1 d'obtenir pile.
- La pièce 2 a une probabilité p_2 d'obtenir pile.
- Notre expérience est donc paramétrée par $\theta = (\lambda, p_1, p_2)$

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.
- On jouera n tours de jeu.
- $(Z_i)_{1 \leqslant i \leqslant n}$ une suite de V.A. i.i.d. suivant la loi de Bernouilli $B(\lambda)$. Si $Z_i = 1$, le joueur utilisera la pièce 1 au i-ème tour. Sinon, si $Z_i = 2$, il utilisera la pièce 2.
- La pièce 1 a une probabilité p_1 d'obtenir pile.
- La pièce 2 a une probabilité p_2 d'obtenir pile.
- Notre expérience est donc paramétrée par $\theta = (\lambda, p_1, p_2)$
- Le tour de jeu utilisant la pièce k est une série de 3 réalisations i.i.d. selon la loi $B(p_k)$.

Formalisons le problème

- Un tour de jeu est composé de 3 lancers.
- On jouera n tours de jeu.
- $(Z_i)_{1 \leqslant i \leqslant n}$ une suite de V.A. i.i.d. suivant la loi de Bernouilli $B(\lambda)$. Si $Z_i = 1$, le joueur utilisera la pièce 1 au i-ème tour. Sinon, si $Z_i = 2$, il utilisera la pièce 2.
- La pièce 1 a une probabilité p_1 d'obtenir pile.
- La pièce 2 a une probabilité p_2 d'obtenir pile.
- Notre expérience est donc paramétrée par $\theta = (\lambda, p_1, p_2)$
- Le tour de jeu utilisant la pièce k est une série de 3 réalisations i.i.d. selon la loi $B(p_k)$.
- Notons H_i le nombre de pile obtenus au i-ème lancer. Conditionnellement à la variable $Z_i = k$, la vraisemblance de X_i s'écrit donc : $L(X_i|p_k) = p_k^{H_i}(1 - p_k)^{(3-H_i)}$

Ecriture de l'agorithme

- On veut, sachant θ_m, trouver θ_{m+1} t.q. :

$$\theta_{m+1} = \arg\max_{\theta}\{\mathbb{E}_{Z|X,\theta_m}[log(L_n(X,Z|\theta)]\}$$

$$\theta_{m+1} = \arg\max_{\theta}\{\sum_{i=1}^{n}\mathbb{E}_{Z|X,\theta_m}[log(L(X_i,Z_i|\theta)]\}$$

- On définit $\tilde{p}_i = P(Z = 1|X_i, \theta_m)$ la probabilité que ce soit la pièce 1 qui ait généré le i-ème lancé, d'après les paramètres actuels θ_m et l'observation X_i. On parle dans le langage bayésien de probabilité *a posteriori*.

- On peut alors récrire :

$$\theta_{m+1} = \arg\max_{\theta}\{\sum_{i=1}^{n}\tilde{p}_i log(P(X_i, Z_i = 1|\theta) + (1-\tilde{p}_i) log(P(X_i, Z_i = 2|\theta)\}$$

Et tout est connu, sauf $\tilde{p}_i$. (Si si, je vous jure !)

ECBL (Ellipse calculatoire de Bôté Légendaire)

Résultat des courses

- On a exprimé $E_{\tilde{p}_i}(\theta) = \mathbb{E}_{Z|X,\theta_m}[log(L_n(X,Z|\theta)]$ comme une fonction des paramètres actuels (cachés dans $\tilde{p}_i$) et on n'a plus qu'à dériver cette expression pour trouver son maximum et en déduire les paramètres améliorés qui maximisent cette expression. Gradient nul, bla bla bla :

$$\frac{\partial E_{\tilde{p}_i}}{\partial \lambda}(\lambda_{m+1}, (p_1)_{m+1}, (p_2)_{m+1}) = 0$$

$$\frac{\partial E_{\tilde{p}_i}}{\partial p_1}(\lambda_{m+1}, (p_1)_{m+1}, (p_2)_{m+1}) = 0$$

$$\frac{\partial E_{\tilde{p}_i}}{\partial p_2}(\lambda_{m+1}, (p_1)_{m+1}, (p_2)_{m+1}) = 0$$

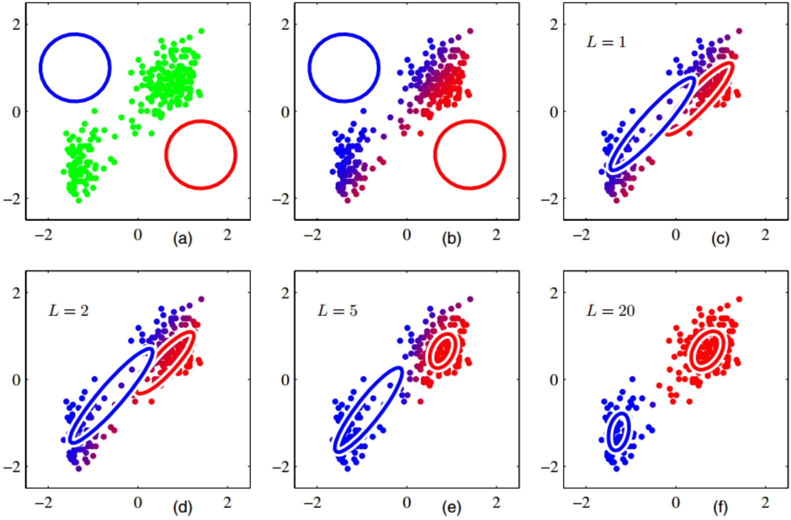
Loi des mélanges, ou mélange des lois #Chiasme #Hashtag

Loi de mélange

- Idée générale : notre variable est issue d'un mélange de plusieurs lois (par exemple, des gaussiennes).
- On ignore les proportions de chaque loi (c'est la donnée manquante).
- L'algorithme EM peut donc nous aider ici encore à déterminer quelles sont ces lois (paramètres à déterminer), et dans quelles proportions.

Application au cas des mélanges de gaussiennes

- On suppose que nos observations sont des réalisations d'une variable qui suit un mélange de lois gaussiennes.
- Paramètre à déterminer : moyenne μ et matrice de covariance Σ pour chaque loi

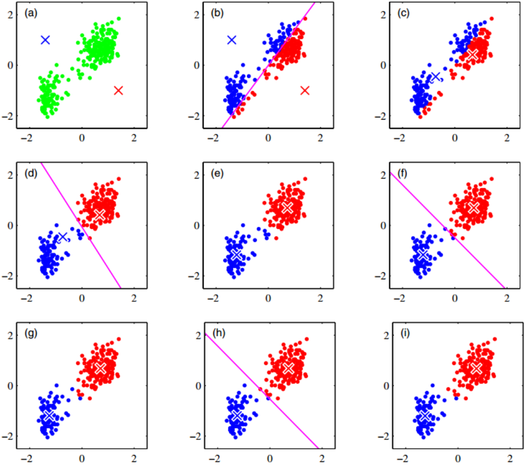


(Image from Bishop Book)

A vue d'oeil, on dirait 2 lois gaussiennes : $X \sim \sum_{i=1}^{2} \pi_i \times \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

K-means et EM

Lien avec l'algorithme des K-means (Image from Bishop Book)



Lien entre EM et K-means

- K-means : définit une et une seule classe par observations en fonction des centres actuels
- EM : définit pour chaque observations des probabilités d'appartenir à différentes lois (c'est ce que représentent les probabilités *a posteriori*)

Ré-écriture

- Avec une réécriture astucieuse, on peut percevoir les K-means comme un cas particulier D'EM pour des mélanges de gaussiennes isotrope (la covariance n'est plus un paramètre à fixer) :

$$\gamma(z_{nk}) = \frac{\pi_k exp^{\frac{-\|x_n-\mu_k\|^2}{2\sigma}}}{\sum_j \pi_k exp^{\frac{-\|x_n-\mu_j\|^2}{2\sigma}}}$$

(Probabilité a posteriori que l'observation x_n suive la loi $\mathcal{N}(\mu_k, \sigma)$)