

5.1 Apprentissage dans le cas d'un modèle complètement observé

Notation La variable Y prend k valeurs et les variables X_1, \dots, X_s prennent r valeurs.

Exemple Dans le cas d'analyse de documents on peut utiliser l'hypothèse "sac de mots" c'est à dire définir X_i par $X_i = 1$ si le mot i appartient au document et 0 sinon. On peut alors, étant donné des observations X_1, \dots, X_s trouver à quel type Y de document (administratif, roman, ...) nous avons affaire.

Le modèle "Naïve Bayes" Dans ce modèle, les X_i sont indépendants entre eux, sachant Y . Ce modèle est faux dans le cas de l'analyse de documents mais cette hypothèse simplifie fortement les calculs. En pratique, cela marche.

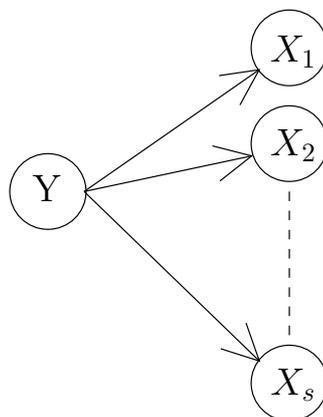


Figure 5.1. Exemple de modèle "Naïve Bayes".

Si on note $X = (X_1, \dots, X_s)$, on a alors

$$p(Y|X) = \frac{p(Y)p(X|Y)}{p(X)} \propto p(Y) \prod p(X_j|Y).$$

Les valeurs $p(X_j = \alpha, Y = \beta) = \eta_{j\alpha\beta}$ sont donc les paramètres du modèle que nous voulons estimer. Pour ce faire nous pouvons utiliser :

- une méthode générative : estimer chaque $\eta_{j\alpha\beta}$ indépendamment,
- une méthode discriminative : estimer directement $\max p(Y|X)$.

Ces méthodes sont expliquées dans les chapitres 6 et 7 du polycopié.

Cas général Dans le cas $G = (\{1, \dots, k\}, E)$ DAG, nous savons que $p(x) = \prod_{j=1}^k p(x_j|x_{\Pi_j})$. Soit θ l'ensemble des paramètres du modèle. Nous faisons maintenant deux hypothèses :

- paramétrisation découplée : $p(x|\theta) = \prod_{j=1}^k p(x_j|x_{\Pi_j}, \theta_j)$,
- données complètes (i.e. on observe toute les variables) et i.i.d.

Proposition 5.1 *Le maximum de vraisemblance se découple.*

Démonstration

$$\begin{aligned} p(\text{Donnees}|\theta) &= \prod_{i=1}^n p(x_1^i, \dots, x_k^i|\theta) = \prod_{i=1}^n \prod_{j=1}^k p(x_j^i|x_{\Pi_j}^i, \theta_j) \\ &= \prod_{j=1}^k \prod_{i=1}^n p(x_j^i|x_{\Pi_j}^i, \theta_j) = \prod_{j=1}^k f_j(\theta_j) \end{aligned}$$

■

5.2 Expectation Maximization

Dans le cas où les données ne sont pas complètes il n'est plus possible d'utiliser la méthode précédent, nous avons besoin d'un nouvel algorithm.

5.2.1 Cadre théorique

Notation X représente les variables aléatoires observées, Z les variables aléatoires cachées et θ les paramètres du modèle.

Situations pratiques d'utilisation :

1. Il y a des données manquantes (fréquent dans l'industrie).
2. Le modèle est plus simple si on introduit une variable cachée (cf figure 5.2).

Ici, on a $Z \in \{1, 2\}$ et $X|Z = i \sim \mathcal{N}(\mu_i, \Theta_i)$.

$$p(x) = \sum_z p(x, z) = \sum_z p(z)p(x|z) = \sum_i p(z = i)\mathcal{N}(x|\mu_i, \Theta_i).$$

La densité $p(x)$ est une combinaison convexe de densités normales.

On parle de modèle de mélanges ("mixtures").

Sans introduire de variable cachées, la représentation de $p(x)$ aurait posé problème.

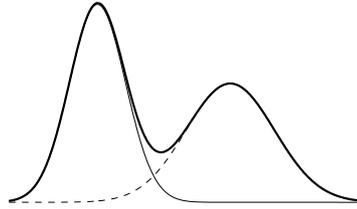


Figure 5.2. Exemple de distribution pour laquelle il est naturel d'introduire une variable cachée.

3. Estimer un paramètre caché. En anglais, on parle de clustering (cas non supervisé) ou de classification (cas supervisé), en français les termes utilisés sont classification et classification/discrimination.

Cadre classique On a des données i.i.d., x_1, \dots, x_i .

$$p(\text{Donnees}|\theta) = \prod_i p(x_i|\theta) = \prod_i \sum_{z_i} p(x_i, z_i|\theta)$$

$$\log p(\text{Donnees}|\theta) = \sum_i \log \sum_{z_i} p(x_i, z_i|\theta)$$

Deux solutions s'offrent alors à nous :

1. Maximiser directement s'il est possible d'utiliser la structure intrinsèque au problème étudié.
2. Utiliser l'algorithme Expectation Maximization (EM).

L'algorithme EM Nous introduisons la fonction $q(z|x)$ ayant les propriétés d'une probabilité, c'est à dire

$$q(z|x) \geq 0 \text{ et } \forall x \sum_z q(z|x) = 1.$$

Nous avons alors

$$\begin{aligned}
 \log p(x|\theta) &= \log \sum_z p(x, z|\theta) \\
 &= \log \sum_z \left(\frac{p(x, z|\theta)}{q(z|x)} \right) q(z|x) \\
 &\geq \sum_z q(z|x) \log \frac{p(x, z|\theta)}{q(z|x)}, \text{ par l'inégalité de Jensen} \\
 &= \sum_z q(z|x) \log p(x, z|\theta) - \sum_z q(z|x) \log q(z|x) \\
 &= \mathcal{L}(q, \theta)
 \end{aligned}$$

On rappelle l'inégalité de Jensen pour une fonction f concave $\mathbb{E}f(x) \leq f(\mathbb{E}x)$.

On a $\log p(x|\theta) = \mathcal{L}(q, \theta)$ si et seulement si pour tout z on a $\frac{p(x, z|\theta)}{q(z|x)} = \text{const}$ ce qui équivaut à $q(z|x) = p(z|x, \theta)$ (car $q(z|x)$ et $p(z|x, \theta)$ sont toutes deux des probabilités).

Proposition 5.2 *Pour tout q on a $\log p(x|\theta) \geq \mathcal{L}(q, \theta)$, avec égalité si et seulement si $q(z|x) = p(z|x, \theta)$.*

L'algorithme Expectation Maximization consiste à maximiser $\mathcal{L}(q, \theta)$ de manière alternée. On commence par initialiser θ_0 puis, pour tout $t \geq 0$ on calcule

- E-step : $q_{t+1} \in \operatorname{argmax}_q \mathcal{L}(q, \theta_t)$
On fait $q_{t+1}(z|x) = p(z|x, \theta_t)$. On trouve la meilleure borne inf.
- M-step : $\theta_{t+1} \in \operatorname{argmax}_\theta \mathcal{L}(q_{t+1}, \theta)$
C'est l'étape de maximisation, il faut trouver le maximum de la borne inf.

Proposition 5.3 1. $\forall t, p(x|\theta_{t+1}) \geq p(x|\theta_t)$

2. *Sous hypothèses de régularité on a convergence vers un point stationnaire de $p(x|\theta)$.*

3. *Nous sommes dans un cas non convexe*

- *l'optimum est rarement global*
- *la limite dépend de l'initialisation*
- *l'optimum global à souvent une vraisemblance ∞*

Recette

1. Écrire la vraisemblance complète $l_c = \log p(x, z|\theta)$.
2. E-step : espérance de l_c sous $p(z|x, \theta)$. On obtient une fonction de θ . On veut $q(z|x) = p(z|x, \theta)$
3. M-step : maximiser en fonction de θ .

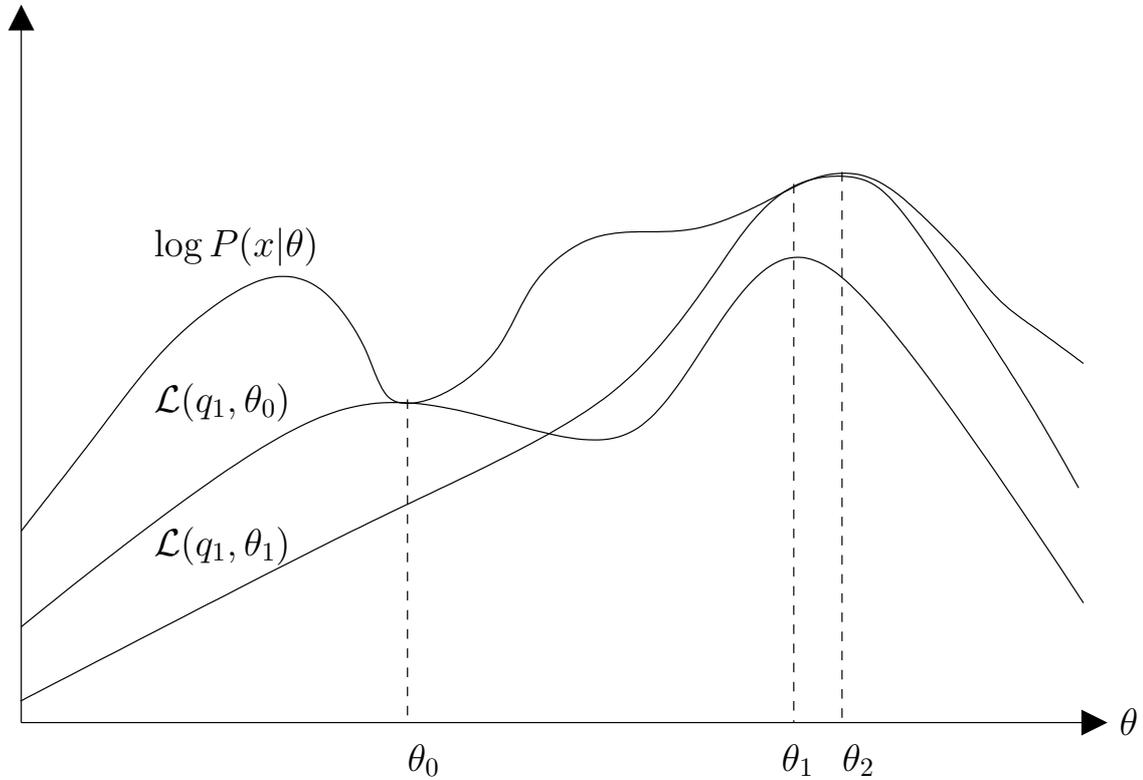


Figure 5.3. Exemple de modèle “Naïve Bayes”.

5.2.2 Sur un exemple : mixtures Gaussiennes

Notation Z prends q valeurs et suit une loi multinomiale Π , $X \in \mathbb{R}^d$ est tel que

$$X|Z = j \sim \mathcal{N}(\mu_j, \Sigma_j).$$

DESSIN

Nous avons des données $x_i \in \mathbb{R}^d$ i.i.d. et nous voulons estimer $p(z|x)$ et les paramètres μ, Σ, Π .

Calcul de $p(z|x)$

$$\begin{aligned} p(z = j|x) &= \frac{p(x|z = j)p(z = j)}{\sum_k p(x|z = k)p(z = k)} \\ &= \frac{\Pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}{\sum_k \Pi_k \mathcal{N}(x|\mu_k, \Sigma_k)} \\ &\propto \exp \left(\log \Pi_j - \frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) - \log \det(\Sigma_j)^{1/2} \right) \end{aligned}$$

Remarque : $p(z = j|x)$ est le softmax des valeurs

$$\log \Pi_j - \frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1}(x - \mu_j) - \log \det(\Sigma_j)^{1/2}.$$

Estimation de $\theta = (\Pi, \mu_j, \Sigma_j)$

Vraisemblance complète

$$\begin{aligned} l_c = \log p(x, z|\theta) &= \sum_{i=1}^n \log p(x_i, z_i|\theta) \\ &= \sum_{i=1}^n \sum_{k=1}^q \delta(z_i = k) \log p(x_i, k|\theta) \\ &= \sum_{i=1}^n \sum_{k=1}^q z_i^k (\log p(z = k|\theta) + \log p(x_i|z_i = k, \theta)), \text{ avec } z_i^k = \delta(z_i = k) \\ &= \sum_{i=1}^n \sum_{k=1}^q z_i^k (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)) \\ \langle l_c \rangle_{z|x, \theta} &= \sum_{i=1}^n \sum_{k=1}^q \langle z_i^k \rangle (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma_k)) \end{aligned}$$

E-step

$$\begin{aligned} \tau_i^k(t) &= \langle z_i^k \rangle \\ &= \mathbb{E}(\delta(z_i = k)) \\ &= p(z_i = k|x_i, \theta(t)) \\ &= \frac{\pi_k(t) \mathcal{N}(x_i|\mu_k(t), \Sigma_k(t))}{\sum_s \pi_s(t) \mathcal{N}(x_i|\mu_s(t), \Sigma_s(t))} \end{aligned}$$

M-step maximisation par rapport à π : $\sum_k (\sum_i \tau_i^k(t)) \log \pi_k$ Donc $\pi_k(t+1) = 1/n \sum_i \tau_i^k(t)$
par rapport à μ_k, Σ_k

$$\sum_i \tau_i^k(t) \log \mathcal{N}(x_i|\mu_k(t), \Sigma_k(t)) = \sum_i \tau_i^k(t) - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k) - \log \det(\Sigma_k)^{1/2}$$

$$\mu_k(t+1) = \frac{\sum_i \tau_i^k(t) x_i}{\sum_i \tau_i^k(t)}$$

$$\Sigma_k(t+1) = \frac{\sum_i \tau_i^k(t) (x_i - \mu_k(t+1))(x_i - \mu_k(t+1))^T}{\sum_i \tau_i^k(t)}$$

Initialisation Comme l'algorithme EM trouve un minimum local, son initialisation est très importante et on essaie donc de trouver une configuration initiale proche du maximum de vraisemblance global.

Pour cela, on utilise l'algorithme K -means qui, étant donné un échantillon de points x_i , que l'on suppose séparables en K clusters de centres μ_k , va chercher à déterminer ces μ_k .

IMPORTANT : la faiblesse majeure de cet algorithme est que l'on a besoin de savoir K à l'avance.

À chaque x_i , on assigne une les variables z_i^k telles que $z_i^k = 1$ ssi x_i appartient au cluster k . L'algorithme K -means va alors chercher à minimiser la fonctionnelle suivante, appelée "distortion" :

$$J(z, \mu) = \sum_{i=1}^n \sum_{k=1}^K z_i^k \|x_i - \mu_k\|^2$$

Algorithme L'algorithme effectue une minimisation alternée :

- minimisation par rapport à z : $z_i^k = 1$ pour $k \in \arg \min \|x_i - \mu_k\|^2$
- minimisation par rapport à μ : $\mu_k = \frac{\sum_i z_i^k x_i}{\sum_i z_i^k}$

Propriété K -means converge vers un minimum local, donc on l'utilise de la manière suivante :

- On lance K -means un grand nombre de fois, avec des initialisations aléatoires.
- On retient les μ pour lesquels on obtient la plus faible distortion.
- On les utilise pour initialiser l'algorithme EM.

5.3 Propriétés de la loi normale

Théorème 5.4 Soient deux variables aléatoire X_1 et X_2 . Si

$$X = (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma)$$

avec

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

et

$$X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}),$$

alors

$$P(x_1|x_2) \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$$

avec

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \text{ et } \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Démonstration

$$p(x_1, x_2) = \frac{\exp \left[-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right]}{(2\pi)^{(n_1+n_2)/2} \begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}^{1/2}}$$

On utilise l'astuce suivante:

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} \begin{pmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{pmatrix} \begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix}$$

avec $M/H = E - FH^{-1}G$

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} I & -FE^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} (E)^{-1} & 0 \\ 0 & (M/E)^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -E^{-1}G & I \end{pmatrix}$$

avec $M/E = H - GE^{-1}F$

On en déduit:

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix} = \det(M/H) \det(H) = \det(M/E) \det(E)$$

Et on obtient:

$$\begin{aligned} p(x_1, x_2) &= \frac{\exp \left[-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right]}{2\pi^{(n_1+n_2)/2} |\Sigma_{22}|^{1/2} |\Sigma/\Sigma_{22}|^{1/2}} \\ &= \frac{\exp \left[-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ x_2 - \mu_2 \end{pmatrix} \right]}{2\pi^{(n_1+n_2)/2} |\Sigma_{22}|^{1/2} |\Sigma/\Sigma_{22}|^{1/2}} \\ &= \frac{\exp \left[-\frac{1}{2} (x_2 - \mu_2)^T \Sigma_{22}^{-1} (x_2 - \mu_2) \right]}{2\pi^{n_2/2} |\Sigma_{22}|^{1/2}} \\ &= \frac{\exp \left[-\frac{1}{2} (x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2))^T (\Sigma/\Sigma_{22})^{-1} (x_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)) \right]}{2\pi^{n_1/2} |\Sigma/\Sigma_{22}|^{1/2}} \end{aligned}$$

Puis, on identifie avec la formule:

$$p(x_1, x_2) = p(x_2)p(x_1|x_2)$$

