Introduction aux modèles graphiques **2008/2009**

Cours 1 — 1 octobre

Enseignant: Francis Bach *Scribe: Grégoire Mesnil, Nicolas Schmidt*

Pour information

– Page web du cours `http://www.di.ens.fr/~fbach/courses/fall2008/`

1.1 Introduction

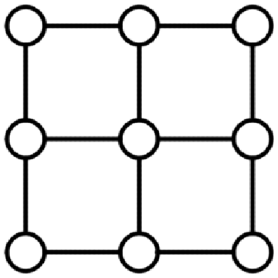
Dans ce cours, nous considèrerons le plus souvent un ensemble $\{X_1, X_2 \ldots, X_n\}$ de variables aléatoires **discrètes** et nous noterons x_i la réalisation de la variable X_i pour tout $i = 1, \ldots, n$. Nous garderons à l'esprit que n est en pratique assez grand.

Les X_i peuvent être définis simplement par la donnée de leur loi jointe $P(X_1 = x_1, \ldots, X_n = x_n)$ (nous verrons que ce n'est pas la meilleure manière de procéder en particulier lorsque n est grand).

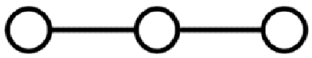
Dans le cadre des variables dite "continues", i.e., à valeurs rélles ou vectorielles, $p(x_1, \ldots, x_n)$ représentera la densité par rapport à la mesure de Lebesgue.

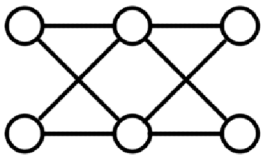
1.1.1 Exemples

– Image : soit une image monochromatique composée de $100 * 100$ pixels. On considère une variable aléatoire discrète par pixel, on a donc $n = 10000$. Le modèle utilisé pourra être une grille de cette forme :



– Bioinformatique : soit une longue séquence de taille 10000 de base ADN. On considère une variable aléatoire discrète par base de cette séquence (en général à valeurs dans $\{a, c, g, t\}$). Le modèle utilisé pourra être une chaîne de Markov :
– Finance : soit 500 actions dont on dispose des valeurs toutes les minutes sur 7 jours. On utilise le modéle suivant :





- Traitement de la parole : soit un signal sonore d'une seconde échantillonné à 10 khz, on obtient alors un modèle mélangeant discret (les mots reconnus) et continu (mesure de la pression de l'air). À cette fréquence, le modèle contiendrait 10000 variables aléatoires seulement pour mesurer la pression de l'air.

- Texte : soit un texte de 1000000 mots. On modélise le texte par un vecteur où chaque composante du vecteur est égale au nombre d'occurences de chaque mot clé. On utilise ici le modèle "bag of words", qui est assez faible car il ne prend pas en compte l'ordre des mots rencontrés dans le texte, mais souvent suffisante en pratique.

On peut déjà constater qu'il est trop faible de considérer un modèle où les variables aléatoires sont toutes indépendantes les unes des autres et qu'il est trop coûteux de supposer que chaque variable est liée à toutes les autres. Il faudra donc faire des hypothèses respectant un certain compromis entre un modèle explicite et un temps de calcul associé raisonnable.

1.1.2 Définitions

Définition 1.1 ***Indépendance***
Deux variables aléatoires X et Y sont indépendantes si quelles que soient les valeurs x et y prises par X et Y, on a :

$$P(X = x, Y = y) = P(X = x)p(Y = y)$$

On notera $X \perp Y$.

Définition 1.2 ***Indépendance conditionnelle***
Soient X, Y, et Z trois variables aléatoires. On dit que X est indépendante de Y sachant Z si X, Y et Z vérifient l'une des deux assertions équivalentes suivantes :

- $\forall x \forall y \forall z, P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$

- $\forall x \forall y \forall z, P(X = x|Y = y, z = z) = P(X = x|Z = z)$

On notera $X \perp Y|Z$ qui se lit X et Y sont indépendantes sachant Z.

1.1.3 Notations

On dira qu'un ensemblee de variables aléatoires est **i.i.d.** lorsque qu'elles sont indépendantes et identiquement distribuées.

Soit X une variable aléatoires discréte (prend un nombre fini de valeur) et $A = \{a_1, \ldots, a_k\}$ une partie de $\{1, \ldots, n\}$. Nous utiliserons dans la suite du cours les abréviations suivantes pour la marginalisation de variables :

$$P\left(X_A = x_A\right) = P\left(X_{a_1} = x_{a_1}, \ldots, X_{a_k} = x_{a_k}\right) = p\left(x_A\right)$$

$$\sum_{x_{a_1}} \sum_{x_{a_2}} \cdots \sum_{x_{a_k}} p\left(x_{a_1}, x_{a_2}, \ldots, x_{a_k}\right) = \sum_{x_A} p\left(x_A\right)$$

En particulier, si $A = \{1\}$, on notera $p(X_1 = x_1) = p(x_1)$

De mÃªme on notera la probabilité conditionnelle de la faÃ§on suivante :

$$P\left(X = x|Y = y\right) = p\left(x|y\right)$$

Soient A et B deux opérateurs et soient $\mathcal{D}_A$ et $\mathcal{D}_B$ leurs domaines de définition respectif. Soit $(a, b) \in \mathcal{D}_A \times \mathcal{D}_B$. On dira que $A(a) \propto B(b)$ lorsque $A - B$ est constant, ou A/B est constant (selon le contexte).
Cette notation sera utilisé pour simplifier l'écriture lors des différents calcules, notament lorsque apparait des constantes ne dépendent pas des variables aléatoires considérée.

1.1.4 Rappels

Formule de Bayes

Soient A et B deux événements, alors

$$P\left(A|B\right) = \frac{P\left(B|A\right) P\left(A\right)}{P\left(B\right)}$$

Marginalisation

On calcule en pratique les probabilités de la maniére suivante :

$$p\left(x_1\right) = \sum_{x_2} \sum_{x_2} \cdots \sum_{x_n} p\left(x_1, x_2, \ldots, x_n\right)$$

On a ainsi pour toute partie A de $\{1, \ldots, n\}$:

$$p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$$

Exercices

- J'ai 2 enfants dont 1 fille, quelle est la probabilité que l'autre soit un garÃ§on ?
- J'ai 3 enfants dont 2 filles, quelle est la probabilité que l'autre soit un garÃ§on ?
- J'ai 1 fille, quelle est la probabilité que celui qui va naître soit un garÃ§on ?

1.2 Modèle à un noeud

Soit X une variable aléatoire avec des observations $X_1, \ldots, X_n$ i.i.d.
Notre objectif est le suivant :

1. Décrire un modéle pour X, i.e., déterminer la loi de X, c'est à dire $p_\theta(x)$, en fontion d'un paramétre θ.
2. Estimer (ou "apprendre") θ à partir des observations $X_1, \ldots, X_n$.

1.2.1 Estimation de paramètre à partir de données i.i.d.

Soit X un variable aléatoire de loi $p_\theta(x)$. Il existe deux philosophies différentes pour estimer θ.

Philosophie Bayésienne On étudie la loi $p_\theta(x)$ en supposant que θ une variable aléatoire. On définit alors la probabilité à priori : $p(\theta)$ et la vraisemblance : $p(x|\theta) = p_\theta(x)$, ce qui permet d'en déduire la loi à postériori (par la régle de Bayes)

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

En particulier, le statisticien Bayésien esssaiera de ne jamais utiliser un estimateur ponctuel de θ, mais utilisera toujours l'ensemble de la loi a posteriori. Dans certains cas, le mode ou la moyenne de cette distribution sont utilisées. Dans le cas du mode, on parle de "maximum a posteriori" (MAP).

Philosophie fréquentiste Il faut trouver un bon estimateur $\hat{\theta}(x_1, \ldots, x_n)$ et l'évaluer. L'estimateur utilisé dans ce cours sera le maximum de vraisemblance : $\max_\theta p_\theta(x)$, qui jouit de propriétés numériques (convexité) et statistiques (en théorie asymptotique) intéressantes [1].

1.2.2 Estimation de lois par maximum de vraisemblance

Les définitions des différentes suivantes se trouve, par exemple, dans [2]

Loi de Bernoulli

Soit $p \in [0, 1]$ et X une variable à valeur dans $\{0, 1\}$, de loi définie comme suit :

$$\begin{cases} p(X=1) = p \\ p(X=0) = 1-p \end{cases}$$

Dans ce cas le paramètre θ à estimer est le réel p. Calcul du maximum de vraisemblance :

On a :

$$\begin{aligned} p(x_1, \ldots, x_n) &= \prod_{i=1}^{n} p(x_i) \text{ à cause de l'indépendance,} \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \text{ car les } X_i \text{ sont indentiquement distribuées.} \end{aligned}$$

On en déduit :

$$-\log(p(x_1, \ldots, x_n)) = -\log(p)(\sum_{i=1}^{n} x_i) - \log(1-p)(n - \sum_{i=1}^{n} x_i)$$

On pose $n_1 = \sum_{i=1}^{n} x_i$ ="nombre de 1". On obtient alors :

$$-\log(p(x_1, \ldots, x_n)) = -n_1 \log p - (n_1 - m)\log(1-p)$$

Cette derniére fonction est convexe par rapport à p. On peut donc déterminer son minimum en annulant son gradient ce qui revient à déterminer p tel que $-\frac{n_1}{p} + \frac{n-n_1}{1-p} = 0$. La solution est $p = \frac{n_1}{n}$, qui est la fréquence empirique de l'observation 1 (estimateur naturel).

On a finalement :

$$p = n_1/n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Loi multinomiale

Soit une variable aléatoire X prenant ses valeurs dans $\{1, \ldots, q\}$. La loi est paramétrée par un vecteur $\pi \in \mathbb{R}^q$ tel que $\pi \geq 0$ et $\sum_i \pi_i = 1$. Soit un échantillon $x_1, \ldots, x_n$ i.i.d.

(indépendant et identiquement distribué). La vraisemblance est donnée par

$$\begin{aligned} p_\pi(x_1,\dots,x_n) &= \prod_{j=1}^{n} p_\pi(x_j) = \prod_{j=1}^{n}\prod_{i=1}^{q} \pi_i^{\delta(x_j=i)} \\ &= \prod_{i=1}^{q} \pi_i^{\sum_{j=1}^{n}\delta(x_j=i)} \\ &= \prod_{i=1}^{q} \pi_i^{n_i} \end{aligned}$$

où $n_i = \sum_{j=1}^{n} \delta(x_j = i)$ est le nombre de valeurs i observées dans l'échantillon.

Le maximum de vraisemblance est donné par :

$$\max_{\pi\geq 0, \sum_i \pi_i = 1} \log\left(\prod_{i=1}^{q} \pi_i^{n_i}\right)$$

$$\Longleftrightarrow$$

$$\min_{\pi\geq 0, \sum_i \pi_i = 1} -\left(\sum_{i=1}^{q} n_i \log \pi_i\right)$$

On introduit les multiplicateurs de Lagrange afin de passer la contrainte $\sum_i \pi_i = 1$ dans la fonction objectif. On obtient le Lagrangien suivant :

$$\mathcal{L}(\pi,\lambda) = -\sum_{i=1}^{q} n_i \log \pi_i + \lambda(\sum_{i=1}^{q} \pi_i - 1)$$

On doit désormais minimiser le Lagrangien par rapport à $\pi \geq 0$ ce qui revient à annuler son gradient. On obtient le système suivant :

$$\left\{\begin{array}{rcl} \frac{\partial}{\partial \pi_1}\mathcal{L}(\pi,\lambda) &=& -\frac{n_1}{\pi_1} + \lambda \quad = \quad 0 \\ & \vdots & \\ \frac{\partial}{\partial \pi_q}\mathcal{L}(\pi,\lambda) &=& -\frac{n_q}{\pi_q} + \lambda \quad = \quad 0 \\ \frac{\partial}{\partial \lambda}\mathcal{L}(\pi,\lambda) &=& \sum_{i=1}^{q}\pi_i - 1 \quad = \quad 0 \end{array}\right.$$

En sommant les q premières équations de notre système, on obtient $\lambda = \sum_{j=1}^{q} n_j$. On remarque que la contrainte réapparaît dans la dernière équation du système.

On obtient donc la solution suivante (fréquence empirique) :

$$\pi_i = \frac{n_i}{\sum_{j=1}^{q} n_j}$$

NB : La page [3] donne quelque précisions sur les multiplicateurs de Lagrange. Pour une introduction à l'optimisation convexe avec contraintes, voir par exemple, le livre [4].

Loi gaussienne

Soit une variable $x \in \mathbb{R}$. Nous supposons qu'elle suit une loi normale parametrée par sa moyenne μ et sa variance σ^2 :

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Soit un échantillon $x_1, \ldots, x_n$ i.i.d. (indépendant et identiquement distribué). La log-vraisemblance est donnée par :

$$\begin{aligned}
\ell\left(\mu, \sigma^2\right) &= \log p\left(x_1, \ldots, x_n | \mu, \sigma^2\right) \\
&= \log \prod_{i=1}^{n} p\left(x_i | \mu, \sigma^2\right) \\
&= \sum_{i=1}^{n} \left(-\log\left(\sqrt{2\pi}\sigma\right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
&\propto -n \log\left(\sigma\right) + \sum_{i=1}^{n} -\frac{(x_i - \mu)^2}{2\sigma^2}
\end{aligned}$$

En dérivant par rapport à μ et σ^2, nous trouvons les estimateurs $\widehat{\mu}$ et $\widehat{\sigma}^2$ qui maximisent la vraisemblance :

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Notons que cette valeur est exactement la moyenne empirique.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

Cette valeur est quasiment égale à la variance empirique (il faudrait changer n par $n-1$ dans le dénominateur).

Loi gaussienne multivariée

Soit une variable $x \in \mathbb{R}^k$. Nous supposons qu'elle suit une loi normale multivariée parametrée par un vecteur de moyennes $\mu \in \mathbb{R}^k$ et une matrice de covariance $\Sigma \in \mathbb{R}^{k\times k}$:

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}}\frac{1}{\sqrt{\det \Sigma}} e^{\frac{-(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}}$$

Soit un échantillon $x_1, \ldots, x_n$ i.i.d. . La log-vraisemblance est donnée par :

$$\begin{aligned}
\ell(\mu, \Sigma) &= \log p(x_1, \ldots, x_n | \mu, \Sigma) \\
&= \log \prod_{i=1}^{n} p(x_i \mid \mu, \Sigma) \\
&= \sum_{i=1}^{n}\left(-\log(2\pi)^{\frac{k}{2}} - \frac{1}{2}\log(\det\Sigma) - \frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu)\right)
\end{aligned}$$

Dans ce cas, notre fonction est convexe et il est possible de la minimiser. On peut dériver par rapport à μ pour trouver l'estimateur qui maximise la log-vraisemblance. Afin de calculer la dérivée, nous utiliserons la proposition suivante :

Proposition 1.3 *Soit un vecteur $v \in \mathbb{R}^k$ et une matrice $Q \in \mathbb{R}^{k\times k}$:*

$$\frac{\partial}{\partial v}\left(v^\top Q v\right) = 2Qv$$

En appliquant cette proposition, il vient :

$$\begin{aligned}
\frac{\partial}{\partial \mu}\ell(\mu, \Sigma) &= \sum_{i=1}^{n}\left(\Sigma^{-1}(x_i - \mu)\right) \\
&= \Sigma^{-1}\left(n\mu - \sum_{i=1}^{n}(x_i)\right)
\end{aligned}$$

Si on considére que cette expression vaut zéro on trouve l'estimateur du vecteur de moyennes :

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Pour calculer l'estimateur de la matrice de covariance, on manipule tout d'abord l'expression de la log-vraisemblance pour faciliter les opérations.
On notera $\Lambda = \Sigma^{-1}$.

$$\ell(\mu, \Sigma) \quad \propto \quad \frac{1}{2}\mathrm{n}\log\det\Lambda - \frac{1}{2}\sum_{i=1}^{n}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)$$

Le dernier terme peut être traité de la façon suivante :

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n}\left((x-\mu)^{\top}\Lambda(x-\mu)\right) &= \frac{1}{2}\sum_{i=1}^{n}\mathrm{Trace}\left((x-\mu)^{\top}\Lambda(x-\mu)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\mathrm{Trace}\left(\Lambda(x-\mu)(x-\mu)\right)^{\top} \\
&= \frac{1}{2}\mathrm{Trace}\left(\Lambda\sum_{i=1}^{n}(x-\mu)(x-\mu)^{\top}\right)
\end{aligned}$$

Définition 1.4 (Matrice de covariance empirique)

$$\widehat{\Sigma} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x-\mu)(x-\mu)^{\top}$$

L'expression de la log-vraisemblance devient alors :

$$\ell(\mu, \Lambda) \quad \propto \quad \frac{1}{2}\mathrm{n}\log\det\Lambda - \frac{n}{2}\mathrm{Trace}\left(\Lambda\widehat{\Sigma}\right)$$

La fonction est la somme d'une fonction concave et d'une fonction linéaire, elle est donc concave. On pourrait essayer de dériver par rapport à chaque élement Λ_{ij}. Mais il est plus aisé de dériver par rapport à toute la matrice :

$$\nabla\ell(\mu, \Lambda) \quad = \quad \frac{n}{2}\Lambda^{-1} - \frac{n}{2}\widehat{\Sigma}$$

Si cette expression est égale à zéro on obtient alors :

$$\Lambda^{-1} = \widehat{\Sigma}$$

L'estimateur de la matrice de covariance est donc la matrice de covariance empirique.


(1) Ne jamais dériver par rapport à chaque élément de la matrice Σ ou Λ.
(2) Toujours vérifier dans les produits matriciels que les dimensions sont compatibles.

1.3 Modéle à deux noeuds

1.3.1 Régression linéaire

On modélise le rapport entre une variable $x \in \mathbb{R}^k$ et une variable $y \in \mathbb{R}$. On notera x^i chaque composante de x. On suppose que la probabilité de y conditionnée à x suit une loi normale :

$$p(y \mid x) = \mathcal{N}(\theta^\top x, \sigma^2)$$


Astuce classique pour ramener le cas affine au cas linéaire : Dans les cas où la moyenne de la distribution gaussienne est de la forme $\theta^\top x + \theta_0$, il suffira de redéfinir x par $\widetilde{x} = (x, 1) \in \mathbb{R}^{k+1}$.

Les données utilisées pour estimer les paramètres sont de la forme $(x_i^1 \dots x_i^q, y_i)$ avec $i = 1 \dots n$, $x_i^j \in \mathbb{R}$ et $y_i \in \mathbb{R}$, et stockés dans une matrice X où $X \in \mathbb{R}^{n \times q}$ est une matrice dont la ligne k est de la forme $(x_k^1 \dots x_k^q)$, et un vecteur y à n dimensions.


Attention à la convention classique de l'appretissage et des statistiques : les données sont stockées par lignes.

Il s'agit de données i.i.d. par paires :

$$\left(x_i^1 \dots x_i^q, y_i\right) \perp \left(x_j^1 \dots x_j^q, y_j\right) i \neq j$$

On utilise la fonction de log-vraisamblance afin de trouver un estimateur pour chacun des paramètres :

$$\begin{aligned}
\ell\left(\theta, \sigma^2\right) &= \sum_{i=1}^{n} \left(-\log\left(\sqrt{2\pi}\right) - \log\sigma - \frac{\left(y_i - \theta^\top x_i\right)^2}{2\sigma^2} \right) \\
&\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \theta^\top x_i\right)^2 \\
&\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\|y - X\theta\right\|^2
\end{aligned}$$

où $X \in \mathbb{R}^{n \times q}$ est une matrice dont la ligne k est de la forme $(x_k^1 \dots x_k^q)$.
Pour dériver par rapport à θ on utilisera la proposition suivante :

Proposition 1.5 *Soit un vecteur $v \in \mathbb{R}^k$ et une matrice $Q \in \mathbb{R}^{k\times k}$:*

$$\frac{\partial}{\partial v}(Qv)^\top (Qv) = 2QQ^\top v$$

On obtient alors :

$$\frac{\partial}{\partial \theta}\ell\left(\theta, \sigma^2\right) = -\frac{1}{2\sigma^2}X^\top (X\theta - y)$$

Si cette expression vaut zéro on obtient les **Equations Normales** :

$$X^\top X\theta = X^\top Y$$

Les estimateurs pour θ et σ^2 sont alors :

$$\widehat{\theta} = \left(X^\top X\right)^{-1} X^\top Y \,;\, \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\theta}^\top x_i\right)^2$$

Le chapitre 6 de [5] décrit l'estimation de θ à l'aide d'algorithmes itératifs.

1.3.2 Classification Linéaire

Ici nous considérons le cas où les sorties prennent leurs valeurs parmi un nombre fini de possibilités : $Y \in \{1, .., q\}$ et où les entrées sont vectorielles ie : $X = (X^1, \ldots, X^k) \in \mathbb{R}^k$. Cette situation s'apparente à un problème de classification.

On pourrait appliquer la méthode de régression linéaire décrite dans la section précedente, mais deux problèmes se posent :

- La densité parametrée qu'on obtient n'est pas restreinte sur les points y_i. Il faut particulariser la distribution continue en ces points.

- La minimisation de l'écart quadratique pénalise les classifications qui seraient parfaites. La distance $\left\|y_i - \theta^\top x\right\|$ peut être grande, tandis que x correspond à la classe i.

Il faut alors utiliser d'autres méthodes. Il y a deux méthodes principales pour approcher ce problème : la méthode discriminative et la méthode générative.

Méthode discriminative

Pour cette méthode on suppose qu'on connait explicitement une expression de $p(y|x, \theta)$, i.e., on connait f telle que $p(y|x, \theta) = f(x, \theta)$. La méthode du maximum de vraisemblance permet de trouver un estimateur de θ de maniére à ce que notre modéle colle au modèle prédictif : le but n'étant pas de modéliser à la fois x et y mais de modéliser x sachant y. Ceci est à contraster avec la méthode dite générative.

Méthode générative

Ici nous n'avons pas d'expression explicite de $p(y|x)$. On définit une loi jointe $p(x, y)$ à partir de laquelle on peut calculer $p(y|x)$ par la règle de Bayes. Afin d'estimer les paramètres, la méthode générative va maximiser la vraisemblance jointe $p(x, y)$ par rapport aux paramètres (au lieu de maximiser la vraisemblance conditionnelle dans le cas de la méthode discriminative).

Le modèle est le suivant :

- Y suit une loi multinomiale (de vecteur Π)
- $\forall j \in \{1, \ldots, q\} \quad X|Y = j$ suit une loi normale : $\mathcal{N}(\mu_j, \Sigma_j)$.

Le but est de trouver les μ_j, Σ_j et Π tels que l'on s'adapte le mieux possible aux couples observés (x_i, y_i).

D'aprés le régle de Bayes on a donc :

$$\begin{aligned} p(y = i|x) &\propto p(x|y = i) \times p(y = i) \\ &\propto \frac{1}{(2\pi)^{k/2}} \frac{1}{\det(\Sigma_i)^{1/2}} \exp\Big(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i) \Big) \times \Pi_i \\ &\propto \frac{1}{\det(\Sigma_i)^{1/2}} \exp\Big(-\frac{1}{2} x^\top \Sigma_i^{-1} x - \frac{1}{2}\mu_i^\top \Sigma_i^{-1}\mu_i + \mu_i^\top \Sigma_i^{-1} x \Big) \times \Pi_i \end{aligned}$$

Si on fait l'hypothése supplémentaire correspondant à LDA (**Linear Discriminant Analysis**) , i.e., $\forall i \in \{1, \ldots, q\} \quad \Sigma_i = \Sigma$, on a :

$$\begin{aligned} p(y = i|x) &\propto \Pi_i \exp\big(-\frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i \big) \exp\big(x^\top \Sigma^{-1}\mu_i \big) \\ &\propto \exp(b_i) \exp(x^\top \theta_i) \end{aligned}$$

Si on renormalise en considérant que $b_i = -\frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i + \log(\Pi_i)$ et $\theta_i = \Sigma^{-1}\mu_i$, on obtient :

$$p(y = i|x) = \frac{e^{x^\top \theta_i + b_i}}{\sum_{j=1}^{q} e^{x^\top \theta_j + b_j}}$$

Définition 1.6 (fonction softmax) *La fonction softmax allant de $\mathbb{R}^q$ dans $\mathbb{R}^q$ associe à chaque vecteur $(z_1, \ldots, z_q)$ le vecteur $S(z_1, \ldots, z_q)$ dont la i-ème composante s'écrit : $S(z_1, \ldots, z_q)_i = \frac{e^{z_i}}{\sum_{j=1}^{q} e^{z_j}}$. On a bien : $\sum_{j=1}^{q} S(z_1, \ldots, z_q) = 1$ et $S(z_1, \ldots, z_q) \geq 0$*

Avec ce formalisme on peut dire que :

$$p(y = i|x) = S(x^\top \theta_1 + b_1, \ldots, x^\top \theta_q + b_q)_i$$

Si on considére le cas où $\Sigma_i \neq \Sigma_j \quad i \neq j$, nommé QDA (**Quadratic Discriminant Analysis**), alors les termes quadratiques ne s'annulent pas. Voir DM)

La maximization de la vraisemblance jointe permet d'estimer les paramêtres μ_j, Π_j et Σ (voir DM) qui permettent alors de calculer θ_j et b_j. Dans le cadre discriminatif, les paramètres θ_j et b_j sont directement estimés.

Regression logistique le cas q=2

On peut de la mÃªme faÃ§on considérer que $Y \in \{0, 1\}$, $Y \in \{1, -1\}$ ou $Y \in \{1, 2\}$, le choix dépendant de ce que l'on veut faire. Pour ce cours, nous considérerons que $Y \in \{0, 1\}$.

On a donc d'après ce qui précéde :

$$\begin{aligned} p(y = 1|x) &= \frac{e^{x^\top \theta_1 + b_1}}{e^{x^\top \theta_0 + b_0} + e^{x^\top \theta_1 + b_1}} \\ &= \frac{1}{1 + e^{-\left(x^\top(\theta_1 - \theta_0) + b_1 - b_0\right)}} \\ &= \sigma\left(x^\top(\theta_1 - \theta_0) + b_1 - b_0\right) \end{aligned}$$

où σ est la fonction sigmoïde i.e., $\sigma(z) = \frac{1}{1+e^{-z}}$.

En utilisant la même astuce que pour la régression linéaire, on peut s'affranchir du terme constant et considérer le modèle : $p(Y = 1|x, \theta) = \sigma(\theta^\top x)$.

Calculons désormais la log-vraisemblance afin de la maximiser pour obtenir l'estimateur $\hat{\theta}$:

$$\begin{aligned} \ell(\theta) &= \sum_i \log p(y = y_i|x_i, \theta) \\ &= \sum_i \log \left(p(y = 1|x_i, \theta)^{\delta_{y_i=1}} p(y = 0|x_i, \theta)^{\delta_{y_i=0}} \right) \\ &= \sum_i y_i \log \left(p(y = 1|x_i, \theta)\right) + (1 - y_i) \log \left(1 - p(y = 1|x_i, \theta)\right) \end{aligned}$$

Il vient :

$$\ell(\theta) = \sum_i y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log \left(1 - \sigma(\theta^\top x_i)\right)$$

On note maintenant que $\log(\sigma(z)) = \log\left(\frac{1}{1+e^{-z}}\right) = -\log(1 + e^{-z})$ et $\log(1 - \sigma(z)) = \log \sigma(-z)$ sont concaves. Donc la log-vraisemblance est aussi concave et on peut lui trouver un maximum. Bien qu'il n'y ait pas de formule analytique pour exprimer ce maximum on peut utiliser des méthodes numériques d'approximation du maximum pour s'en rapprocher (ex :méthode itérative de Newton pour minimiser une fonction convexe $f \in C^\infty$).

Bibliographie

[1] Gilbert Saporta. *Probabilités, analyses des données et statistiques*. Technip, 1990.

[2] Aim Fuchs Dominique Foata. *Calcul des probabilités*. 2ème édition. Dunod, 2003.

[3] http ://fr.wikipedia.org/wiki/multiplicateur_de_lagrange.

[4] Frédéric Bonnans. *Optimisation continue, Cours et problèmes corrigés*. Dunod, 2003.

[5] Michael Jordan. *An introduction to graphical models*. (en préparation).