Introduction aux modèles graphiques **2009/2010**

Cours 1 — 30 septembre

Enseignant: Francis Bach *Scribe: Marc Weber, Ruocong Zhang*

Pour information

– Page web du cours `http://www.di.ens.fr/~fbach/courses/fall2009/`

1.1 Introduction

1.1.1 Problèmes posés

Lorsque l'on veut réaliser des modélisations statistiques de données complexes, on se trouve confronté à des questions issues de deux problématiques principales :

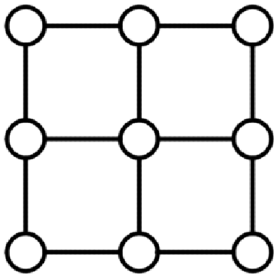
– Comment gérer la complexité des données à traiter ?

– Comment inférer les propriétés globales à partir de modèles locaux ?

Les problèmes rencontrés sont de trois types : la représentation des données (Comment obtenir un modèle global à partir d'un modèle local), l'inférence des lois (Comment utiliser le modèle), et l'apprentissage des modèles (Quels sont les paramètres du modèle ?).

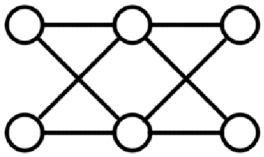
1.1.2 Exemples

– 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 :

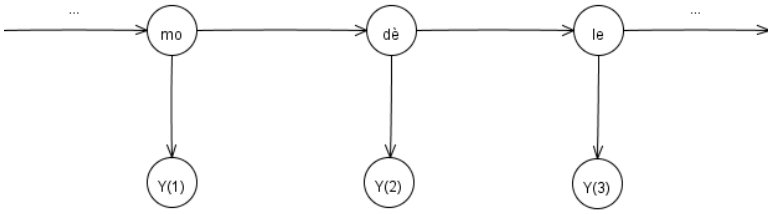


– Bioinformatique : soit une longue séquence de taille 10000 de base ADN. On considère une variable aléatoire discrète par base de cette séquence (en général à valeurs dans $\{A, C, G, T\}$). Le modèle utilisé pourra être une chaîne de Markov :

– Finance : On considère des actions évoluant dans un domaine temporel discret où l'on dispose des valeurs aux instants n. Il est légitime de supposer que l'évolution d'une action à l'instant n peut dépendre de l'évolution d'une autre action au temps $n-1$. Pour un modèle simplifié à deux actions, on aura donc le graphe de dépendance suivant :



– Traitement de la parole : On considère les syllabes d'un mot et la manière dont elles sont interprétées par l'oreille humaine ou par un ordinateur. A chaque syllabe d'un mot correspond un son aléatoire (la même syllabe sera prononcée différemment chaque fois). On cherche alors à remonter au mot prononcé en fonction des sons entendus. Dans ce cas il est possible d'utiliser un modèle de Markov caché.



– 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 il est souvent suffisant en pratique. L'algorithme utilisé pour la classification (par exemple spam vs non spam) est le "naive Bayes".

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.

1.2 Rappels de probabilités

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 dites "continues", i.e., à valeurs réelles ou vectorielles, $p(x_1, \ldots, x_n)$ représentera la densité par rapport à la mesure de Lebesgue.

1.2.1 Définitions

Définition 1.1 (Indépendance) *Deux variables aléatoires X et Y sont dites indépendantes, notées $X \perp Y$, si quelles que soient les valeurs x et y prises par X et Y, on a :*

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

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 :*

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

On notera cette relation d'indépendance $X \perp Y|Z$, qui se lit "X et Y sont indépendantes sachant Z".

1.2.2 Notations

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

Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire discrète (prend un nombre fini de valeurs) et $A = \{a_1, \ldots, a_k\}$ un sous-ensemble de $\{1, \ldots, n\}$. Nous utiliserons dans la suite du cours les abréviations suivantes pour la marginalisation de variables :

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

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

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

De même on notera la probabilité conditionnelle de la façon suivante :

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

Soient A et B deux opérateurs et soient $\mathcal{D}_A$ et $\mathcal{D}_B$ leurs domaines de définition respectifs. 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 utilisée pour simplifier l'écriture lors des différents calculs, notamment lorsqu'apparaissent des constantes ne dépendant pas des variables aléatoires considérées.

1.2.3 Autres rappels

Formule de Bayes

Soient A et B deux événements, alors

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

Marginalisation

On calcule en pratique les probabilités de la manière suivante :

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

On a ainsi pour tout sous-ensemble A de $\{1, \ldots, n\}$:

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

Exercices

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

1.3 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 modèle pour X, i.e., déterminer la loi de X, c'est-à-dire $p_\theta(x)$, en fonction d'un paramètre θ.
2. Estimer (ou "apprendre") θ à partir des observations $X_1, \ldots, X_n$.

1.3.1 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 (cf. cours de méthodes MCMC et applications) On étudie la loi $p_\theta(x)$ en supposant que θ une variable aléatoire. On définit alors la probabilité a priori : $p(\theta)$ et la vraisemblance : $p(x|\theta) = p_\theta(x)$, ce qui permet d'en déduire la loi a posteriori (par la règle de Bayes)

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

En particulier, le statisticien Bayésien essaiera 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, qui jouit de propriétés numériques (convexité) et statistiques (en théorie asymptotique) intéressantes [1].

Définition 1.3 (Estimateur de maximum de vraisemblance) *Soit une loi $p_\theta(x)$, avec $x \in \chi$, et des données $x_1, \ldots, x_n \in \chi$ i.i.d. La vraisemblance $L(\theta) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$ est alors égale à $\prod_{i=1}^n P_\theta(X_i = x_i) = \prod_{i=1}^n p_\theta(x_i)$ L'estimateur de maximum de vraisemblance (EMV) $\hat{\theta}$ est défini de la façon suivante :*

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\max_\theta \prod_{i=1}^n p_\theta(x_i)$$

1.3.2 Estimation de lois par maximum de vraisemblance

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

Loi de Bernoulli

Soit $p \in [0, 1]$ et X une variable à valeurs 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 (attention à la surcharge de notation, ici pratique).

On a :

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

Plutôt que de considérer la vraisemblance, il est plus pratique d'effectuer les calculs sur la log-vraisemblance $\ell(\theta)$, en appliquant le logarithme à l'équation. On en déduit :

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

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

$$\ell(\theta) = n_1 \log p + (n - n_1) \log(1-p)$$

Cette dernière 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 :

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

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

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

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 :

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

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) :

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

Remarque : Nous nous sommes ici contentés de la seconde contrainte ($\sum_i \pi_i = 1$), la contrainte de positivité $\pi \geq 0$ étant naturellement imposée par le logarithme qui tend vers $-\infty$ en 0. Sinon, il aurait fallu introduire un terme supplémentaire dans le lagrangien, de type $\sum_{j=1}^{k} \mu_j \pi_j$ avec $\mu_j \geq 0$.

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 paramétrée 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. 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.

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

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 paramétrée 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 2 du formulaire, ce qui donne :

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

Si on considère 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 $\Lambda = \Sigma^{-1}$.

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

Le terme sous la somme étant réel, il est égal à sa trace. On peut alors utiliser les propriétés de la trace de la façon suivante :

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

Définition 1.4 (Matrice de covariance empirique)

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

L'expression de la log-vraisemblance devient alors :

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

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 (on utilise les propositions 3 et 4 du formulaire) :

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

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.
(3) L'indépendance est sur les données, en lignes de 1 à n, tandis que les variables sont dépendantes entres elles, en colonnes de 1 à k.

1.4 Modèle à deux noeuds

1.4.1 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 stockées 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'apprentissage et des statistiques : les données sont stockées par lignes. La raison pour cela est que les données sont souvent stockées dans des fichiers texte pour lesquels le saut de ligne est naturel, contrairement au saut de colonne.

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-vraisemblance 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 5 du formulaire et on obtient :

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

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.
- X est une matrice de taille $n * k$ où n est le nombre d'observations et k le nombre de paramètres dont dépend la loi de Y. Une condition mathématique nécessaire pour que $X^\top X$ (de taille $k * k$) soit inversible est que n soit supérieur à k, puisque $rg(X^\top X) \leq min(n, k)$. Ceci peut être interprété de la manière suivante : si $n < k$, nous possédons moins d'observations qu'il n'y a de paramètres à apprendre. Nous sommes donc dans un cas de surapprentissage.
- Dans le cadre de la prédiction, les données permettent d'estimer $\hat{\theta}$, et pour tout nouveau $x \in \mathbb{R}^k$, de prédire y. La variable y sachant x sera prédite suivant une loi normale de moyenne $\hat{\theta}^\top x$, de variance $\hat{\sigma}^2$. On choisit la loi normale car d'une part, le bruit est souvent gaussien, d'autre part parce que la loi possède des propriétés intéressantes (théorème de la limite centrale et dérivation aisée par exemple).

1.4.2 Classification Linéaire (introduction)

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.

Considérons le cas q=2.
On définit la fonction logistique par $\sigma(z) = \frac{1}{1+e^{-z}}$.
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 :

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

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).