Introduction aux modèles graphiques **2010/2011**

Cours 9 — 1^{er} décembre

Enseignant: Guillaume OBOZINSKI *Scribe: Nicolas* CORDIER, *Vivien* FÉCAMP

9.1 Naive Bayes

9.1.1 Introduction

Remarque : Contrairement à ce que son nom semble indiquer, « Naive Bayes » n'est pas une méthode bayésienne.

Considérons le problème de classification suivant : $x \in \mathbb{X}^p \longmapsto y \in \{1, 2, \ldots, M\}$.

Ici, $x = (x_1, x_2, \ldots, x_p)$ est un vecteur de descripteurs (ou « traits » ou « features ») : $\forall i \in \{1, 2, \ldots, p\}, x_i \in \mathbb{X}$, avec $\mathbb{X} = \{1, 2, \ldots, K\}$ (ou $\mathbb{X} = \mathbb{R}$).

But : apprendre $p(y|x)$

Une méthode très naïve conduirait à une explosion combinatoire : $\theta \in \mathbb{R}^{K^p}$.

Par la formule de Bayes :

$$p(y|x) = \frac{p(x|y)\, p(y)}{p(x)}$$

Le modèle Naive Bayes consiste à faire l'hypothèse assez forte que les descripteurs x_i sont indépendants conditionnellement à la classe, d'où :

$$p(x|y) = \prod_{i=1}^{p} p(x_i|y)$$

La formule de Bayes donne donc :

$$p(y|x) = \frac{p(y) \prod_{i=1}^{p} p(x_i|y)}{p(x)} = \frac{p(y) \prod_{i=1}^{p} p(x_i|y)}{\sum_{y'} p(y') \prod_{i=1}^{p} p(x_i|y')}$$

On considère le cas où les descripteurs prennent des valeurs discrètes. Le modèle graphique considéré ne contient alors que des variables discrètes. On peut toujours écrire un modèle discret comme un famille exponentielle : en effet on peut écrire

On paramétrise la distribution en écrivant

$$\log p(x_i = k | y = k') = \delta(x_i = k,\, y = k')\, \theta_{ikk'}$$

et

$$\log p(y = k') = \delta(y = k')\, \theta_{k'}$$

Notons que les indicatrices $\delta(x_i = k, y = k')$ et $\delta(y = k')$ sont les *statistiques suffisantes* pour le modèle de la distribution jointe de y et des variables x_i et que $\theta_{ikk'}$ et $\theta_{k'}$ sont ses *paramètres canoniques*. On peut alors écrire :

$$\log p(y, x_1, \ldots, x_p) = \sum_{i,k,k'} \delta(x_i = k, y = k')\theta_{ikk'} + \sum_{k'} \delta(y = k')\theta_{k'} - A((\theta_{ikk'})_{i,k,k'}, (\theta_{k'})_{k'})$$

où $A((\theta_{ikk'})_{i,k,k'}, (\theta_{k'})_{k'})$ est la fonction de log-partition.

Nous avons réécrit le modèle de la distribution jointe de $(y, x_1, \ldots, x_p)$ comme une famille exponentielle, or l'estimateur du maximum de vraisemblance dans une famille exponentielle dont les paramètres canoniques ne sont pas couplés est aussi, comme nous l'avons vu dans le cours sur les familles exponentielles, l'estimateur du maximum d'entropie sous contrainte d'égalité de moments des statistiques suffisantes (les moments doivent être égaux aux moments empiriques).

Donc si on introduit

$$N_{ikk'} = \#\{(x_i, y) = (k, k')\}$$
$$N = \sum_{i,k,k'} N_{ikk'},$$

l'estimateur du maximum de vraisemblance doit satisfaire les contraintes de moment

$$\widehat{p}\,(y = k') = \frac{\sum_{i,k} N_{ikk'}}{N} \qquad \text{et} \qquad \widehat{p}\,(x_i = k | y = k') = \frac{N_{ikk'}}{\sum_{k"} N_{ik"k'}},$$

qui le caractérisent entièrement.

On peut alors écrire les estimateurs des paramètres canoniques du modèle génératif comme

$$\widehat{\theta}_{ikk'} = \log \widehat{p}\,(x_i = k | y = k') \qquad \text{et} \qquad \widehat{\theta}_{k'} = \log \widehat{p}\,(y = k')\,.$$

Néanmoins notre but est d'obtenir un modèle de classification, c'est-à-dire un modèle de la seule loi conditionnelle. En partant du modèle génératif estimé et en appliquant la règle de Bayes on obtient

$$\log \widehat{p}\,(y = k'|x) = \sum_{i=1}^{p} \log \widehat{p}\,(x_i | y = k') + \log \widehat{p}\,(y = k') - \log \sum_{k'} \left(\widehat{p}\,(y = k') \prod_{i=1}^{p} \widehat{p}\,(x_i | y = k') \right)$$

On peut réécrire ce modèle conditionnel sous forme de famille exponentielle

$$\log p\,(y|x) = \sum_{i,k,k'} \delta(x_i = k, y = k')\theta_{ikk'} + \sum_{k'} \delta(y = k')\theta_{k'} - \log p(x)$$

Ses statistiques suffisantes et ses paramètres canoniques sont les mêmes que ceux du modèle génératif, mais vus comme des fonctions de la variables aléatoire y, x étant fixé (on pourrait écrire $\phi_{x,i,k,k'}(y) = \delta(x_i = k, y = k')$). Quant à la fonction de log-partition, elle est maintenant égale à $\log p(x)$.

Attention : $\widehat{\theta}_{ikk'}$ est l'estimateur du maximum de vraisemblance dans le modèle génératif et en général il ne se confond pas avec le maximum de vraisemblance dans le modèle conditionnel.

9.1.2 Avantages et inconvénients

Avantages :
- Possible en ligne.
- Computationnellement acceptable.

Inconvénients :
- Génératif : les modèles génératifs fournissent de bons estimateurs lorsque le modèle est "juste", ou en terme statistique *bien spécifié*, ce qui signifie que le processus qui génère les données réelles induit une distribution qui est celle du modèle génératif. Lorsque le modèle est *mal spécifié* (ce qui est de loin le cas le plus courant) on aura intérêt à utiliser une méthode discriminative.

9.1.3 Méthode discriminative

Le problème que nous avons considéré dans la section précédente est un modèle génératif pour la classification en K classes. Comment apprendre de façon discriminative un classifieur à K classes ? Est-il possible d'utiliser la même famille exponentielle ?

Nous avions étudié la régression logistique à deux classes :

$$p\left(y = 1|x\right) = \frac{\exp\left(\omega^T x\right)}{1 + \exp\left(\omega^T x\right)}$$

Etudions la régression logistique multi-classe (K classes) :

$$
\begin{aligned}
p\left(y=k'|x\right) &= \frac{\exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i=k\right)\theta_{ikk'}\right)}{\sum_{k"=1}^{M}\exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i=k\right)\theta_{ikk"}\right)} \\
&= \exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i=k\right)\theta_{ikk'} - \log\left(\sum_{k"=1}^{M}\exp\left(\sum_{i=1}^{p}\sum_{k=1}^{K}\delta\left(x_i=k\right)\theta_{ikk"}\right)\right)\right) \\
&= \exp\left(\theta_{k'}^{T}\phi\left(x\right) - \log\left(\sum_{k"=1}^{M}\exp\left(\theta_{k"}^{T}\phi\left(x\right)\right)\right)\right) \\
&= \frac{\exp\left(\theta_{k'}^{T}\phi\left(x\right)\right)}{\sum_{k"=1}^{M}\exp\left(\theta_{k"}^{T}\phi\left(x\right)\right)}
\end{aligned}
$$

Bien que nous ayons construit le modèle sur la base de considérations différentes le modèle obtenu (c'est-à-dire l'ensemble des distributions possibles) est la même famille exponentielle que dans le modèle de Naive Bayes.

En revanche, le modèle appris sera différent dans une approche discriminative que dans un approche générative : l'apprentissage de la régression logistique multi-classe est obtenu en maximisant la vraisemblance des classes $y^{(j)}$ d'un ensemble d'apprentissage, les $x^{(j)}$ étant fixés, c'est-à-dire en calculant l'estimateur du maximum de vraisemblance dans le modèle conditionnel. Contrairement à ce qui se passe pour le modèle génératif, l'estimateur ne peut être obtenu en forme analytique et l'apprentissage requiert de résoudre numériquement le problème d'optimisation.

9.2 Méthodes Bayésiennes

9.2.1 Introduction

Vocabulaire :
- a priori : $p(\theta)$
- vraisemblance : $p(x|\theta)$
- vraisemblance marginale : $\int p(x|\theta)\, p(\theta)\, d\theta$
- a posteriori : $p(\theta|x)$

La formulation bayésienne permet d'introduire de l'information à priori dans le processus d'estimation. Par exemple, imaginons que nous jouions à pile ou face :
- avec un jeton « inconnu », nous n'avons aucune information a priori : nous utiliserons la loi uniforme pour $p(\theta)$.
- avec une pièce de monnaie « normale », nous utiliserons une distribution de masse importante autour de 0,5 pour $p(\theta)$.

Pour un bayésien, proposer un estimateur "ponctuel", comme l'estimateur du maximum de vraisemblance qui propose une seule valeur de θ n'est pas satisfaisant car l'estimateur lui-même ne rend pas compte de l'incertitude intrinsèque au processus d'estimation ou d'apprentissage. Son estimateur sera donc la densité a postériori obtenue en utilisant la règle de Bayes, qui s'écrit en notation continue

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

Ainsi, le bayésien spécifie l'incertitude avec des distributions qui forment sont estimateur, plutôt que de combiner un estimateur avec des intervalles de confiance.

Si le bayésien est contraint de produire un estimateur ponctuel, il utilise l'espérance de la quantité sous-jacent sous la distribution a postériori ; par exemple pour θ :

$$\mu_{post} = \mathbb{E}[\theta|D] = \mathbb{E}[\theta|x_1, x_2, \ldots, x_n] = \int \theta p(\theta|x_1, x_2, \ldots, x_n)\, d\theta$$

9.2.2 Maximum a posteriori (MAP)

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} p(\theta|x_1, x_2, \ldots, x_n) \\ &= \arg\max_{\theta} p(x_1, x_2, \ldots, x_n|\theta)\, p(\theta)\end{aligned}$$

car, par la formule de Bayes :

$$p(\theta|x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n|\theta)\, p(\theta)}{p(x)}$$

Le Maximum a posteriori n'est pas vraiment Bayésien, c'est plutôt une petite modification apportée à un estimateur fréquentiste.

9.2.3 Probabilité prédictive

Dans le paradigme bayésien, la probabilité d'une observation future x^* sera estimée par la *probabilité prédictive* :

$$\begin{aligned} p(x^*|D) &= p(x^*|x_1, x_2, \ldots, x_n) \\ &= \int p(x^*|\theta)\, p(\theta|x_1, x_2, \ldots, x_n)\, d\theta \end{aligned}$$

$$\begin{aligned} p(\theta|x_1, x_2, \ldots, x_n) &\propto p(x_n|\theta)\, p(x_1|\theta)\, p(x_2|\theta) \ldots p(x_{n-1}|\theta)\, p(\theta) \\ &\propto p(x_n|\theta)\, p(\theta|x_1, x_2, \ldots, x_{n-1})\, p(x_1, x_2, \ldots, x_{n-1}) \\ &\propto p(x_n|\theta)\, p(\theta|x_1, x_2, \ldots, x_{n-1})\, \frac{p(x_1, x_2, \ldots, x_{n-1})}{p(x_1, x_2, \ldots, x_n)} \end{aligned}$$

Un calcul séquentiel est possible puisque :

$$p(\theta|x_1, x_2, \ldots, x_n) = \frac{p(x_n|\theta)\, p(\theta|x_1, x_2, \ldots, x_{n-1})}{p(x_n|x_1, x_2, \ldots, x_{n-1})}$$

Vocabulaire :
- nouvel a priori : $p(\theta|x_1, x_2, \ldots, x_{n-1})$
- vraisemblance : $p(x_n|\theta)$
- nouvel a posteriori : $p(\theta|x_1, x_2, \ldots, x_n)$

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta)\, p(\theta)\, d\theta$$

9.2.4 Situations échangeables

9.2.4.1 Echangeabilité

Les variables aléatoires X_1, X_2, …, X_n sont échangeables si elles ont la même distribution que $X_{\widehat{\Sigma}(1)}$, $X_{\widehat{\Sigma}(2)}$, …, $X_{\widehat{\Sigma}(n)}$ pour toute permutation des indices $\widehat{\Sigma}$.

9.2.4.2 Théorème de Finetti

Si X_1, X_2, …, X_n sont échangeables, alors il existe un processus stochastique G tel que :

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|G)\, d\mu(G)$$

où $d\mu(G)$ est la généralisation de "$p(G)\, dG$" pour un processus stochastique.

9.2.4.3 Pourquoi s'intéresser aux situations échangeables ?

Des variables i.i.d sont un cas particulier de variables échangeables que l'on rencontre en pratique. Cependant lorsque des données i.i.d sont composées d'observations qui ne sont pas scalaires, les différentes composantes ne sont le plus souvent pas indépendantes. Dans certain cas ces composantes sont néanmoins échangeables. Par exemple dans un texte les mots présentés en séquence ne sont pas échangeables à cause de la syntaxe, mais si nous oublions l'ordre des mots comme dans le modèle « bag-of-words », alors il deviennent échangeables. C'est le principe de modélisation utilisé dans le modèle Latent Dirichlet Allocation.

9.2.5 Exemple de modèle

9.2.5.1 Variable de Bernoulli

Considérons des variables aléatoires $X_i \in \{0, 1\}$. Nous supposerons les X_i i.i.d. sachant θ.

$$p\left(x|\theta\right) = \theta^x \left(1-\theta\right)^{1-x}$$

9.2.5.2 A priori

Introduisons la *distribution* Beta dont la densité sur l'intervalle $[0, 1]$ est

$$p(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

où $B(\alpha, \beta)$ dénote la *fonction* Beta :

$$\forall \alpha > 0,\, \forall \beta > 0,\, B\left(\alpha, \beta\right) = \int_0^1 \theta^{\alpha-1} \left(1-\theta\right)^{\beta-1} d\theta$$

et la fonction Gamma :

$$\Gamma\left(x\right) = \int_0^{+\infty} t^{x-1} \exp\left(-t\right) dt$$

Nous pouvons montrer que :

$$B\left(\alpha, \beta\right) = \frac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)}$$

Nous choisissons comme distribution a priori sur θ la distribution Beta :

$$p\left(\theta\right) \propto \theta^{\alpha-1} \left(1-\theta\right)^{\beta-1}$$

$$p\left(\theta\right) = \frac{\theta^{\alpha-1} \left(1-\theta\right)^{\beta-1}}{B\left(\alpha, \beta\right)}$$

9.2.5.3 A posteriori

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

Or :

$$p(x,\theta) = \theta^x (1-\theta)^{1-x} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

Donc :

$$p(\theta|x) \propto \frac{\theta^{x+\alpha-1}(1-\theta)^{1-x+\beta-1}}{B(\alpha,\beta)}$$

$$p(\theta|x) = \frac{\theta^{x+\alpha-1}(1-\theta)^{1-x+\beta-1}}{B(x+\alpha,1-x+\beta)}$$

Donc si au lieu de considérer une seule variable, nous observons un échantillon i.i.d., la distribution jointe s'écrit

$$\theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

Introduisons :

$$k = \sum_{i=1}^{n} x_i$$

On a alors :

$$p(\theta|x_1, x_2, \ldots, x_n) = \frac{\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}$$

9.2.6 Distributions

$$\theta \sim Beta(\alpha,\beta)$$

Pour $\alpha = \beta = 1$, nous avons un a priori uniforme.
Pour $\alpha = \beta > 1$, nous avons une courbe en cloche.
Pour $\alpha = \beta < 1$, nous avons une courbe en U.
$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$
$\mathbb{V}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{(\alpha+\beta)} \times \frac{\beta}{(\alpha+\beta)} \times \frac{1}{(\alpha+\beta+1)}$
Pour $\alpha > 1$ et $\beta > 1$, nous avons le mode : $\frac{\alpha-1}{\alpha+\beta-2}$.
Dans le cas qui nous intéresse ici, notons D les données :

$$\theta_{post} = \mathbb{E}[\theta|D] = \frac{\alpha+k}{\alpha+\beta+n} = \frac{\alpha}{(\alpha+\beta)} \times \frac{(\alpha+\beta)}{(\alpha+\beta+n)} + \frac{n}{(\alpha+\beta+n)} \times \frac{k}{n}$$

Nous voyons que l'espérance a posteriori du paramètre est une combinaison convexe de l'estimateur du maximum de vraisemblance et de l'espérance a priori. Il se rapproche asymptotiquement de l'estimateur du maximum de vraisemblance.

Si on utilise un a priori uniforme, $\mathbb{E}|\mathbb{D}[\theta] = \frac{k+1}{n+2}$. Laplace avait proposé de corriger l'estimateur des fréquences, dont il ne trouvait pas naturel qu'il ne soit pas défini en l'absence d'observation. Il proposait ainsi de compter deux observation virtuelles, une valant 0, l'autre valant 1 de sorte que l'estimateur vaille $\frac{1}{2}$ en l'absence d'observations. Cette correction de l'estimateur de fréquence est connu sous le nom de *correction de Laplace*.

La variance de la distribution a posteriori décroit en $\frac{1}{n}$

$$\mathbb{V}[\theta|D] = \theta_M(1-\theta_M)\frac{1}{(\alpha+\beta+n)}$$

Nous avons ainsi une distribution de plus en plus piquée autour de θ_M, de la même façon que dans une approche fréquentiste les intervalles de confiance se resserrent autour de l'estimateur lorsque le nombre d'observations augmente.

9.2.7 Propriété ludique

$$p(x_1, x_2, \ldots, x_n) = \frac{B(k+\alpha, n-k+\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+k)\Gamma(\beta+n-k)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)\Gamma(\beta)} \tag{9.1}$$

Utilisons une propriété de la fonction Gamma :

$$\Gamma(n+1) = n!$$

$$\text{et} \qquad \forall x > -1,\ \Gamma(x+1) = x\Gamma(x)$$

de sorte que

$$\Gamma(\alpha+k) = (\alpha+k-1)(\alpha+k-2)\ldots\alpha\Gamma(\alpha)$$

Notons $\alpha^{[k]} = \alpha(\alpha+1)\ldots(\alpha+k-1)$ et simplifions l'expression 9.1 :

$$p(x_1, x_2, \ldots, x_n) = \frac{\alpha^{[k]}\beta^{[n-k]}}{(\alpha+\beta)^{[n]}}$$

Remarquons l'analogie avec les urnes de Pólya : considérons $(\alpha+\beta)$ boules de couleur : α sont noires, β sont blanches. Imaginons que nous tirions une première boule de couleur noire, la probabilité de cet événement est :

$$\mathbb{P}(X_1 = 1) = \frac{\alpha}{\alpha+\beta}$$

Après tirage, nous replaçons dans l'urne la boule tirée, et nous ajoutons une boule supplémentaire de même couleur que la boule tirée. Imaginons que nous tirions une seconde boule de couleur noire, la probabilité de cet événement est :

$$\mathbb{P}(X_1 = 1,\, X_2 = 1) = \mathbb{P}(X_1 = 1)\,\mathbb{P}(X_2 = 1|X_1 = 1) = \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+1}{\alpha+\beta+1}$$

En revanche :

$$\mathbb{P}(X_1 = 1,\, X_2 = 0) = \frac{\alpha}{\alpha+\beta} \times \frac{\beta}{\alpha+\beta+1}$$

De façons plus générale, on montre par récurrence que la probabilité marginale d'obtenir une certaine séquence de couleurs par tirage avec les urnes de Pólya est la même que la probabilité marginale d'obtenir la même séquence dans le modèle marginal obtenu en intégrant par rapport à un a priori θ. D'une part, ceci montre que (de façon presque paradoxale) les tirages d'une urne de Pólya sont échangeables ; d'autre part le mécanisme des urnes de Pólya, et son échangeabilité fournira un bon outil pour faire de l'échantillonnage de Gibbs dans les modèles bayésiens de ce type.

9.2.8 A priori conjugué

Soit $\mathbb{F}$ un ensemble. On suppose $p(x|\theta)$ connu, on en déduit $p(\theta) \in \mathbb{F}$ tel que $p(\theta|x) \in \mathbb{F}$. On dit que $p(\theta)$ est conjugué au modèle $p(x|\theta)$.

9.2.8.1 Modèle exponentiel

Considérons :

$$\begin{aligned} p(x|\theta) &= \exp\left(\langle\theta, \phi(x)\rangle - A(\theta)\right) \\ p(\theta) &= \exp\left(\langle\alpha, \theta\rangle - \tau A(\theta) - B(\alpha, \tau)\right) \end{aligned}$$

Pour $p(x|\theta)$, θ est le paramètre canonique. Pour $p(\theta)$, α est le paramètre canonique et θ est la statistique suffisante. Remarquons que B ne désigne plus la distribution Beta.

$$p(\theta|x) \propto p(x|\theta)\,p(\theta) \propto \exp\left(\langle\theta, \phi(x)\rangle - A(\theta) + \langle\alpha, \theta\rangle - \tau A(\theta) - B(\alpha, \tau)\right)$$

Définissons :

$$\bar{\phi} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$$

Alors :

$$p(\theta|x_i) \propto \exp\left(\langle\theta, \alpha + \phi(x_i)\rangle - (\tau+1)\,A(\theta) - B(\alpha + \phi(x_i), \tau+1)\right)$$

$$p(\theta|x_1, x_2, \ldots, x_n) \propto \exp\left(\left\langle\theta, \alpha + n\bar{\phi}\right\rangle - (\tau+n)\,A(\theta) - B\left(\alpha + n\bar{\phi}, \tau+n\right)\right)$$

$$p\left(x_1, x_2, \ldots, x_n\right) \propto \exp\left(B\left(\alpha, \tau\right) - B\left(\alpha + n\bar{\phi}, \tau + n\right)\right)$$

Comme la famille est exponentielle,

$$\nu_{post} = \mathbb{E}\left[\theta|D\right] = \nabla_\alpha B\left(\alpha + n\bar{\phi}, \tau + n\right)$$

θ_{MAP} est obtenu par :

$$\nabla_\theta p\left(\theta|x_1, x_2, \ldots, x_n\right) = 0$$
$$\alpha + n\bar{\phi} = \left(\tau + n\right)\nabla_\theta A\left(\theta\right) = \left(\tau + n\right)\mu\left(\theta\right)$$

Nous obtenons donc $\mu_{MAP} = \mu\left(\theta\right)$ dans l'équation précédente. Donc :

$$\mu_{MAP} = \frac{\alpha + n\bar{\phi}}{\tau + n} = \frac{\alpha}{\tau} \times \frac{\tau}{\tau + n} + \frac{n}{\tau + n}\bar{\phi}$$

9.2.8.2 Gaussienne univariée

Avec un a priori sur μ mais pas sur σ^2

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

$$p\left(\mu|\mu_0, \tau^2\right) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \mu_0\right)^2}{\tau^2}\right)$$

Donc :

$$\begin{aligned} p\left(D|\mu, \sigma^2\right) &= p\left(x_1, x_2, \ldots, x_n|\mu, \sigma^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(x_i - \mu\right)^2}{\sigma^2}\right) \end{aligned}$$

$$\begin{aligned} p\left(\mu|D\right) &= p\left(\mu|x_1, x_2, \ldots, x_n\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\left(\mu - \mu_0\right)^2}{\tau^2} + \sum_{i=1}^{n}\frac{\left(x_i - \mu\right)^2}{\sigma^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\tau^2} + \sum_{i=1}^{n}\frac{\mu^2 - 2\mu x_i + x_i^2}{\sigma^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\mu^2\Lambda - 2\mu\eta + \left(\frac{\mu_0^2}{\tau^2} + \sum_{i=1}^{n}\frac{x_i^2}{\sigma^2}\right)\right)\right) \end{aligned}$$

où :

$$\Lambda = \frac{1}{\tau^2} + \frac{n}{\sigma^2}$$
$$\eta = \frac{\mu_0}{\tau^2} + \frac{n\overline{x}}{\sigma^2}$$
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Donc :

$$\begin{aligned}
\mu_{post} &= \mathbb{E}\left[\mu|D\right] \\
&= \frac{\eta}{\Lambda} \\
&= \frac{\frac{\mu_0}{\tau^2} + \frac{n\overline{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \\
&= \frac{\sigma^2\mu_0 + n\tau^2\overline{x}}{\sigma^2 + n\tau^2} \\
&= \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu_0 + \frac{n\tau^2}{\sigma^2 + n\tau^2}\overline{x}
\end{aligned}$$

Et :

$$\begin{aligned}
\widehat{\Sigma}^2_{post} &= \mathbb{V}\left[\mu|D\right] \\
&= \frac{1}{\Lambda} \\
&= \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}
\end{aligned}$$

La variance décroît bien en $\frac{1}{n}$.

Avec un a priori sur σ^2 mais pas sur μ Nous obtenons alors $p\left(\sigma^2\right)$ sous la forme d'une Inverse Gamma.

Avec un a priori sur μ et sur σ^2 A priori gaussien pour x et pour μ, a priori Inverse Gamma pour σ^2. Se rapporter au chapitre 9 du polycopié de cours.

9.2.8.3 Généralisation de la distribution Beta

Dirichlet, qui est la conjuguée de la Multinômiale.

$$p\left(\theta_1, \theta_2, \ldots, \theta_k\right) = \frac{\Gamma\left(\alpha_1 + \alpha_2 + \ldots + \alpha_k\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\alpha_2\right)\ldots\Gamma\left(\alpha_k\right)}\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\ldots\theta_k^{\alpha_k - 1}d\mu\left(\theta\right)$$

où μ désigne la mesure uniforme sur $\left\{s \in \mathbb{R}^k \mid \sum_i s_i = 1\,;\, \forall i,\, s_i \geq 0\right\}$ (simplexe).

9.3 Sélection de modèles

9.3.1 Introduction

Considérons deux modèles $M_1 \subset M_2$ avec $\Theta_1 \subset \Theta_2$. Définissons

$$\widehat{\Theta}_{M_i} = \arg\max_{\theta \in \Theta_i} \log\left(p_\theta\left(x_1, x_2, \ldots, x_n\right)\right)$$

où $i \in \{1, 2\}$.

Nous ne pouvons pas utiliser le maximum de vraisemblance comme score, puisque nécessairement $\log\left(p_{\widehat{\Theta}_{M_2}}\right) \geq \log\left(p_{\widehat{\Theta}_{M_1}}\right)$.

Nous nous intéressons à la capacité de généralisation du modèle : nous voulons éviter le sur-apprentissage. Une façon de résoudre ce problème est de sélectionner la taille du modèle par validation croisée. C'est la méthode la plus utilisée en pratique. Nous n'en parlerons pas dans ce cours.

Nous présentons dans cette partie du cours les *facteurs de Bayes*, qui fournissent l'outil bayésien principal pour la sélection de modèle et nous montrons son lien avec la pénalité BIC (Bayésian Information Criterion) qui est utilisé par les fréquentistes pour "corriger" le maximum de vraisemblance et qui à de bonne propriétés. Le problème de la sélection de modèle, et le problème de sélection de variables sont des problématiques riches et complexes qui font encore l'objet de recherche actives. Il existe d'autres pénalités que la pénalité BIC et d'autres approches à la sélection de modèles.

Si p_0 est la distribution des données réelles, nous souhaitons choisir entre différents modèles $(M_i)_{i \in I}$ en maximisant $\mathbb{E}_{p_0}\left[\log\left(p_{M_i}\left(X^* | D\right)\right)\right]$, où X^* est un nouvel échantillon de test distribué selon p_0 (en réalité il s'agit encore du principe du maximum de vraisemblance mais en espérance sur de nouvelles données).

Dans le cadre Bayésien, on peut calculer la probabilité marginale des données pour un modèles donné

$$\int p\left(x_1, x_2, \ldots, x_n | \theta\right) p\left(\theta | M_i\right) d\theta = p\left(D | M_i\right)$$

et, en utilisant la règle de Bayes, calculer la probabilité a posteriori du modèle

$$p\left(M_i | D\right) = \frac{p\left(D | M_i\right) p\left(M_i\right)}{p\left(D\right)}$$

9.3.2 Facteur de Bayes

Introduisons le facteur de Bayes (« Bayes factor »), qui permet de comparer deux modèles :

$$\frac{p\left(M_1 | D\right)}{p\left(M_2 | D\right)} = \frac{p\left(D | M_1\right) p\left(M_1\right)}{p\left(D | M_2\right) p\left(M_2\right)}$$

La probabilité marginale des données

$$p\left(D|M_i\right) = p\left(x_1, x_2, \ldots, x_n|M_i\right)$$

peut se décomposer de façon séquentielle en utilisant :

$$p\left(x_n|x_1, x_2, \ldots, x_{n-1}, M\right) = \int p\left(x_n|\theta\right) p\left(\theta|x_1, x_2, \ldots, x_{n-1}, M\right) d\theta.$$

En effet on a :

$$p(D|M) = p(x_n|x-1, \ldots, x_{n-1}, M)\, p(x_{n-1}|x-1, \ldots, x_{n-2}, M) \ldots\ p(x_1|M)$$

de telle sorte que

$$\frac{1}{n}\log p\left(D|M_i\right) = \frac{1}{n}\sum_{i=1}^{n}\log p(x_i|x_1, \ldots, x_{i-1}, M) \simeq \mathbb{E}_{p_0}\left[\log p_M\left(X|D\right)\right]$$

9.3.3 Proposition

Le score Bayésien est approximé par le score BIC (« Bayesian information criterion »).

$$\log p\left(D|M\right) = \log p_{\widehat{\theta}_{MV}}\left(D\right) - \frac{K}{2}\log\left(n\right) + O\left(1\right)$$

avec $p_{\widehat{\theta}_{MV}}\left(D\right)$ la distribution des données lorsque le paramètre est l'estimateur du maximum de vraisemblance $\widehat{\theta}_{MV}$, K est le nombre de paramètres du modèle et n le nombre de données.

Dans la section suivante, nous ébauchons une preuve de ce résultat dans le cas d'une famille exponentielle donnée par $p\left(x|\theta\right) = \exp\left(\left\langle\theta, \phi\left(X\right)\right\rangle - A\left(\theta\right)\right)$.

9.3.4 Méthode de Laplace

$$\begin{aligned} p\left(D|M\right) &= \int \prod_{i=1}^{n} p\left(x_i|\theta\right) p\left(\theta\right) d\theta \\ &= \int \exp\left(\left\langle\theta, n\bar{\phi}\right\rangle - n\, A\left(\theta\right)\right) p\left(\theta\right) d\theta \end{aligned}$$

$$\begin{aligned} \langle\theta, n\bar{\phi}\rangle - n\, A(\theta) = {} & \langle\widehat{\theta}, n\bar{\phi}\rangle - n\, A(\widehat{\theta}) + \langle\theta - \widehat{\theta}, n\bar{\phi}\rangle \\ & - n(\theta - \widehat{\theta})^T \nabla_\theta A(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^T n \nabla^2_\theta A(\widehat{\theta})(\theta - \widehat{\theta}) \\ & + \mathrm{R}_n \end{aligned}$$

où R_n désigne un reste négligeable devant les autres termes.

Or le maximum de vraisemblance est le dual du maximum d'entropie : $\max H(p_\theta)$ tel que $\mu(\theta) = \bar{\phi}$.

$$\mu(\widehat{\theta}) = \bar{\phi}$$

$$p(D|M) \simeq \exp(\langle \widehat{\theta}, n\bar{\phi} \rangle - n\, A(\widehat{\theta})) \times \int \exp\left(-\frac{1}{2}(\theta - \widehat{\theta})^T n\widehat{\Sigma}(\theta - \widehat{\theta})\right) p(\theta) d\theta$$

Or :

1. l'information de Fisher est égale à $\widehat{\Sigma}^{-1}$
2. $\displaystyle\int \exp\left(-\frac{1}{2}\left(\theta - \widehat{\theta}\right)^T n\widehat{\Sigma}\left(\theta - \widehat{\theta}\right)\right) p\left(\theta\right) d\theta \simeq c\sqrt{\left(2\pi\right)^k \left|\frac{\widehat{\Sigma}^{-1}}{n}\right|}$

Donc :

$$\begin{aligned}
\log p\left(D|M\right) &= \log p_{\widehat{\theta}}\left(X\right) + \frac{1}{2}\log\left(\left(2\pi\right)^k \left|\frac{\widehat{\Sigma}^{-1}}{n}\right|\right) \\
&= \log p_{\widehat{\theta}}\left(X\right) + \frac{k}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\left(\frac{1}{n}\right)^k \left|\widehat{\Sigma}^{-1}\right|\right) \\
&= \log p_{\widehat{\theta}}\left(X\right) + \frac{k}{2}\log\left(2\pi\right) - \frac{k}{2}\log\left(n\right) + \frac{1}{2}\log\left(\left|\widehat{\Sigma}^{-1}\right|\right)
\end{aligned}$$

La motivation principale pour présenter le critère BIC est qu'un théorème stipule que le score BIC est consistant, c'est-à -dire que lorsque le nombre de données est suffisamment grand il choisit avec probabilité tendant vers un modèle qui vérifie :

$$M_k \in \mathrm{Argmax}_M\, \mathbb{E}_{p_0}\left[\log\left(p_{\widehat{\theta}_{MV}}\left(X\,;\,M\right)\right)\right]$$