Introduction to probabilistic graphical models **2015/2016**

Lecture 10 — December 3, 2015

Lecturer: Simon Lacoste-Julien *Scribes: Gauthier Gidel and Lilian Besson*

Note: These scribed notes have only been lightly proofread.

10.1 Bayesian Method

10.1.1 Introduction

Vocabulary:

- *a priori* or prior: $p\left(\theta\right)$
- *likelihood*: $p\left(x|\theta\right)$
- *marginal* likelihood: $\int p\left(x|\theta\right)p\left(\theta\right)d\theta$
- *a posteriori* or posterior: $p\left(\theta|x\right)$

Caricature Bayesian vs Frequentist:

1. the *Bayesian* is "optimistic": he thinks that he can come up with good models and obtain a method by "pulling the Bayesian crank" (basically a high dimensional integral),
2. the *frequentist* is more "pessimistic" and uses analysis tools.

The Bayesian formulation enables us to introduce the a priori information in the process of estimation. For instance , let's imagine that we play heads or tails. The Bayesian model is:

$$X_i \in \{0,1\}, \qquad X_i|\theta \sim Ber(\theta), \quad p(x_i|\theta) = \theta^{x_i}\left(1-\theta\right)^{1-x_i}$$

the graphical model associated is represented on Figure 10.1.

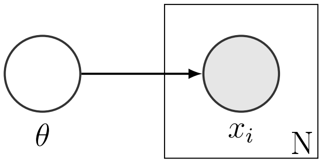

Figure 10.1. Graphical model of the biased coin game

Now we can compute the posterior:

$$p(\theta|x_{1:n}) \propto p(x_{1:n}|\theta)p(\theta)$$

then

$$p(\theta|x_{1:n}) = \theta^{n_1}\left(1-\theta\right)^{n-n_1}\mathbf{1}_{[0,1]}(\theta) = Beta(\alpha,\beta)$$

where $n_1 = \sum_{i=1}^{n} x_i$ is the number of 1, $\beta = n - n_1 + 1$ and $\alpha = n_1 + 1$.

Question: what is the probability of head on the next flip?

- Frequensist: $\hat{\theta}_{ML} = n_1/n$ by a maximum likelihood approach.
- Bayesian: $p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\theta)p(\theta|x_{1:n})d\theta$, where $p(\theta|x_{1:n})d\theta$ is the posterior distribution. Then,

  $$\hat{\theta}_B = \frac{\alpha}{\alpha+\beta} = \frac{n_1+1}{n+2}$$

  hence,

  $$\hat{\theta}_B = \frac{n_1}{n}\left[\frac{n}{n+2}\right] + \frac{1}{2}\left[\frac{2}{n+2}\right] = \rho_n\hat{\theta}_{ML} + \left(1-\rho_n\right)\hat{\theta}_{prior}$$

  is a convex combination of $\hat{\theta}_{ML}$ and $\hat{\theta}_{prior}$. Then we can notice that for $n = 0$, the quantity $\hat{\theta}_B = \frac{1}{2}$ whereas $\hat{\theta}_{ML}$ is not defined. It underlines the importance of the prior distibution:

  - with an “unknown” coin, we’ve got the information a priori : we’ll use the uniform law for $p\left(\theta\right)$.
  - with a “normal” coin , we’ll use a distribution with an important concentration of mass around 0,5 for $p\left(\theta\right)$.

  For a Bayesian, offering a “limited” estimator, as the maximum likelihood estimator, which gives a unique value for θ, is not enough because the estimator itself do not translate the inherent uncertainty of the learning process. Thus, its estimator will be the density a posteriori, obtained from the Bayes rule, which is written in continuous notations as:

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

  The Bayesian specifies the uncertainty with distributions that form its estimator, rather than combining an estimator with confidence intervals.

  If the Bayesian is forced to produce a limited estimator, he uses the expectation of the underlying quantity under the a posteriori distribution; for instance for θ:

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

For more details about Bayesians see subsection B.1 and B.1.1 in annex.

We then need to show that $\hat{\theta}_{ML} \to \theta^*$. Its variance is the variance of a Beta law

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \left(\frac{n_1}{n}\right)\left(1-\frac{n_1}{n}\right)\cdot O\left(\frac{1}{n}\right) = \hat{\theta}_{ML}\left(1-\hat{\theta}_{ML}\right)O\left(\frac{1}{n}\right)$$

then the posterior covariance vanishes and

$$\hat{\theta}_B \overset{a.s.}{\to} \hat{\theta}_{ML} \overset{a.s.}{\to} \theta^*$$

where θ^* is the "true" parameter of the model.

10.1.2 Bernstein von Mises Theorem

It says that if prior puts non-zero mass around the true model θ^*, then posterior asymptotically concentrate around θ^* as a Gaussian.

Revisiting example Consider repeating several times the experiment above: T coins picked randomly each flipped n times. (Figure 10.2)

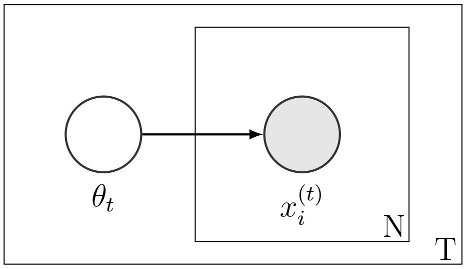

Figure 10.2. Graphical model of the biased coin game repeated T times

As a frequentist, empirical distribution on $x_{1:n}$ will converge (as $T \to \infty$) to

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

where $p(\theta)$ is the distribution of coins of parameter θ in the jar and $\prod_{i=1}^n p(x_i|\theta)$ is the mixture distribution. Note that $X_1, \ldots, X_n$ are **NOT** independent.

On the other hand, for all $\pi \in \mathcal{S}_n$

$$p(x_1, \ldots, x_n) = p\left(x_{\pi(1)}, \ldots, x_{\pi(n)}\right)$$

10.1.3 Exchangeable situations

Exchangeablility

The random variables X_1, X_2, ..., X_n are exchangeable if they have the same distribution as $X_{\pi(1)}$,$X_{\pi(2)}$, ..., $X_{\pi(n)}$ for any permutation of indices $\pi \in \mathcal{S}_n$.

Infinite Exchangeablility

The definition naturally generalizes to infinite families (indexed by $\mathbb{N}$). The random variables $X_1, X_2, \ldots$ are exchangeable if every finite subfamily $X_{i_1}, \ldots, X_{i_n}$ is exchangeable.

de Finetti's theorem

X_1, X_2, ... are infinitely exchangeable, if and only if $\exists!\ p(\theta)$ (on some space Θ) such that

$$\forall n \in \mathbb{N},\ p\left(x_1, x_2, \ldots, x_n\right) = \int \left(\prod_{i=1}^{n} p\left(x_i|\theta\right)\right) p(\theta) d\theta$$

Why do we care about exchangeable situations?

The i.i.d. variables are a particular case of the situation of exchangeable variables, that we see in practice. However when the i.i.d. data are combined with non scalar observations, the different components are no longer independent. In some cases, those components are nonetheless exchangeable. For instance in a text, words are shown as sequences that are not exchangeable because of the syntax. But if we forget the order of the words as in the "bag of word" model, then the components are exchangeable. It's the basic principle used in the LDA model.

Multinomial example

Let $X|\theta \sim Mult(\theta, 1)$ where $\theta \in \Delta_k$ i.e.

$$p(X = l|\theta) = \theta_l \quad \text{and} \quad \sum_{l=1}^{k} \theta_l = 1,\ 0 \leq \theta_l \leq 1.$$

for that distribution we have,

$$\hat{\theta}_l^{ML} = \frac{n_l}{n}$$

hence if $k \geq n$ there exists a l such that $\hat{\theta}_l^{ML} = 0$.

In that case this frequentist model overfits. In the Bayesian model one puts a prior on $\Delta_k = \Theta$, but which one? A convenient property of prior families is "conjugacy", introduced below:

Conjugacy Consider a family of distribution

$$F = \{p(\theta|\alpha) \ : \ \alpha \in \mathcal{A}\} .$$

One says that F is a "conjugate family" for the observation model $p(x|\theta)$ if the posterior

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

belongs to the same family F than the prior, i.e.

$$\exists\, \alpha' \in \mathcal{A} \quad s.t \quad p(\theta|x,\alpha) = p(\theta|\alpha')$$

For the multinomial distribution it gives us

$$p(x_{1:n}|\theta) = \prod_{l=1}^{n} p(x_l|\theta) = \prod_{l=1}^{n} \theta_l^{n_l}$$

so if $p(\theta) \propto \prod_{l=1}^{n} \theta_l^{\alpha_l}$, then $p(x_{1:n}|\theta) \propto \prod_{l=1}^{n} \theta_l^{\beta_l}$.

Dirichlet Distribution

The Dirichlet distribution is the conjugate of the Multinomial law (see on Wikipédia for more details).

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

Where μ stands for the uniform measure on $\Delta_K = \left\{ s \in \mathbb{R}^K \mid \sum_i s_i = 1\, ;\, \forall i,\, s_i \geq 0 \right\}$ (K-dim simplex).

- $\mathbb{E}\left[\theta_l|\alpha_1, \ldots, \alpha_K\right]$,
- $\mathbb{V}(\theta_l) \equiv O\left(\frac{1}{\sum_{j=1}^{K} \alpha_j}\right)$,
- If $\alpha_l = 1$ for all l then one gets an uniform distribution,
- if $k = 2$ one gets the Beta distribution,
- if there exists l such that $\alpha_l < 1$ one gets a $\cup$ shape distribution,
- if $\alpha_l \geq 1$ for all l, one gets a $\cap$ (unimodal bump).

For the multinomial model, if the we assume that the prior is

$$p(\theta) = Dir(\theta|\alpha)$$

then the posterior is

$$p(\theta|x_{1:n}) \propto \prod_{l=1}^{K} \theta_l^{n_l+\alpha_l-1}$$

and the posterior mean is

$$\mathbb{E}\left[\theta_l|x_{1:n}\right] = \frac{n_l + \alpha_l}{n + \sum\limits_{j=1}^{K} \alpha_j}$$

for instance with $\alpha_l = 1$ for all l it adds 1, "smoothing" the maximum likelihood estimator.

$$\mathbb{E}\left[\theta_l|x_{1:n}\right] = \frac{n_l + 1}{n + K}$$

NB One can consider that posterior can be used for prior of next observation. This is the *sequential approach.*

10.2 Bayesian linear regression

Let us assume that

$$y = \omega^T x + \epsilon \tag{10.1}$$

where $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. Then the observation issue

$$p(y|x) = \mathcal{N}\left(y \,|\, \omega^T x \,,\, \sigma^2\right)$$

Then if we also choose a Gaussian prior on ω.

$$p(\omega) = \mathcal{N}\left(\omega \,;\, 0, \frac{I_n}{\lambda}\right)$$

then the posterior is also a Gaussian with the following parameters

- covariance: $\hat{\Sigma}_n = \lambda I_n + \frac{X^T X}{\sigma^2}$
- mean: $\hat{\mu}_n = \hat{\Sigma}_n^{-1}\left(X^T \vec{y}/\sigma^2\right)$

where

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

the covariance and the mean are the same as the ones for the *ridge regression* with $\tilde{\lambda} = \lambda\sigma^2$.

As a Bayesian: compute predictive distribution

$$\begin{aligned} p(y_{new}|x_{new}, x_{1:n}, y_{1:n}) &= \int_\omega p(y_{new}|x_{new}, \omega) p(\omega|data) d\omega \\ &= \mathcal{N}\left(y_{new}|\hat{\mu}_n^T x_{new}, \sigma^2_{predictive}\right) \end{aligned}$$

where

$$\sigma^2_{predictive}(x_{new}) = \sigma^2 + x_{new}^T \hat{\Sigma}_n x_{new},$$

the real number σ comes from the noise model and the second quantity of the right hand side comes from the posterior covariance.

10.3 Model Selection

10.3.1 Introduction

Let's consider two models $M_1 \subset M_2$ with $\Theta_1 \subset \Theta_2$. We define:

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

where $i \in \{1, 2\}$.

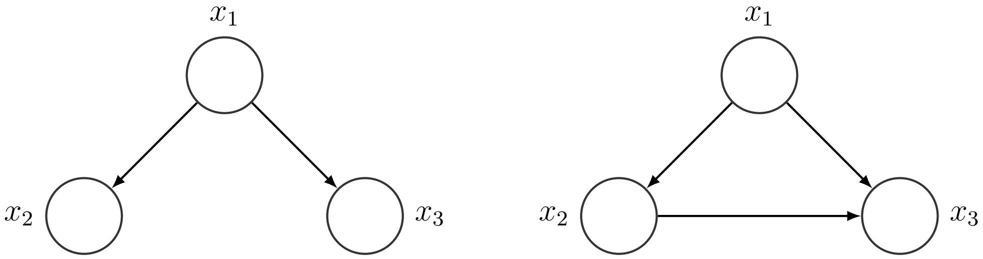

Figure 10.3. Example of Model Section for $n = 2$ (M_1 on the l.h.s and M_2 on the r.h.s)

We can't use the maximum likelihood as a score since we have by definition:

$$\log\left(p_{\widehat{\Theta}_{M_2}}\right) \geq \log\left(p_{\widehat{\Theta}_{M_1}}\right).$$

We are interested in the capacity of the generalisation of the model: we'd like to avoid over-fitting. Commonly, one way of dealing with that task is to select the size of the model by cross-validation. Here, we'll not develop it furthermore.

In this part we present the *Bayes factors*, which gives us the main Bayes principal for selecting models. Also we will show the link with the penalised version BIC, (Bayesian Information Criterion) which is used by the frequentists so as to "correct" the maximum likelihood and which has good proprieties. The issue with the selection model ask is the issue with the selection of the variables which are an active topic of research. There are others ways of penalising the maximum likelihood and of selecting models.

If p_0 is the distribution of the real data, we wish to choose between difference models $(M_i)_{i \in I}$ by maximising $\mathbb{E}_{p_0}\left[\log\left(p_{M_i}\left(X^*|D\right)\right)\right]$, where X^* is a new test sample distributed as p_0 (in fact, it's still the maximum likelihood principle but we take the expectation on new data).

In the Bayesian framework, we can compute the marginal probability of data for a given model

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

and, by applying the Bayes rule, compute the a posteriori probability of the model:

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

10.3.2 Bayes Factor

Let's introduce the Bayes factors, which enables us to compare two models:

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

The marginal probability of data

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

can decompose itself in a sequential way by using:

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

Indeed, we get:

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

Such as

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

10.3.3 Bayesian Information Criterion

The Bayesian score is approximated by the BIC:

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

With $p_{\widehat{\theta}_{MV}}\left(D\right)$ the data's distribution when the parameter is the maximum likelihood estimator $\widehat{\theta}_{MV}$, K is the number of parameters of the model and n the number of observations.

In the following section, we outline the proof of this result in the case of an exponential family given by $p\left(x|\theta\right) = \exp\left(\langle\theta, \phi\left(X\right)\rangle - A\left(\theta\right)\right)$.

10.3.4 Laplace's Method

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

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

where R_n is a negligible rest.

But the maximum likelihood is the dual of the maximum entropy: $\max H(p_\theta)$ such that $\mu(\theta) = \bar{\phi}$.

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

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

However:

1. the information of fisher is equal to $\widehat{\Sigma}^{-1}$
2. $\displaystyle\int \exp\left(-\frac{1}{2}\left(\theta - \widehat{\theta}\right)^T n\widehat{\Sigma}\left(\theta - \widehat{\theta}\right)\right) p\left(\theta\right) d\theta \simeq c\sqrt{\left(2\pi\right)^k \left|\frac{\widehat{\Sigma}^{-1}}{n}\right|}$

Thus:

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

The main reason why presenting the BIC is that a theorem prove the consistency of the BIC. In other words, when the number of observations is sufficient, thanks to this criterion we choose with a probability that converges to 0, a model that satisfies:

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

To bring a quick clarification about the notations used in this part (model selection), please read below. The notation is a bit confusing (it was used for example in Bishop's book, but is a bit sloppy).

From the Bayesian perspective, we could treat the model choice as a random variable M. In the M_1 vs. M_2 vs. M_3 example, there are only 3 models, and thus M is a discrete variable with 3 possible values ($M = M_1$, $M = M_2$ or $M = M_3$).

Therefore, when we were writing quantities like the Bayes factor $p(M_1|D)/p(M_2|D)$, It really meant $p(M = M_1|D)/p(M = M_2|D)$. It did not mean that M_1 and M_2 were two different random variables which can take complicated values (someone asked what space M_1 was in and it seemed very complicated – what is meant is just that M is an index in possible (few) models).

D was the data random variable as usual. The mixing of random variables (here M) vs. their possible values ($M = 1, 2$ etc) in the same notation (like $p(M_1|D)$) is usual but confusing; better to use the explicit $p(M = M_1|D)$ notation to distinguish a value vs. a generic random variable. . . .

However, in general, M could be as complicated as we want. For example, it could be a vector of hyper-parameters for the prior distributions. Or it could also have binary component indicating the absence or presence of an edge in graphical model, etc. It does not have to just be an index. It could even be a continuous objects !

It is also fine to have infinite dimensional objects[1]. For example, consider the latent variable model: x is observed, θ and α are latent variables; and M decides the prior over α. I.e. suppose $p(x|\theta, \alpha, M) = Multi(\theta, 1)$, $p(\theta|\alpha, M) = Dir(\theta|\alpha)$, and $p(\alpha|M) = M(\alpha)$ i.e. M ranges over possible distributions over the positive vector α. M here is quite a complicated object, but this is fine. . .

[1] This would be in the "non-parametric setting" – non-parametric = infinite dimensional.

Appendix A

A.1 Example of model

A.1.1 Bernoulli variable

Let's consider random variables $X_i \in \{0, 1\}$. We'll assume that the X_i are i.i.d. conditionally to θ. Then they follow a Bernoulli law:

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

A.1.2 Priors

Let's introduce the *distribution* Beta whose density on $[0, 1]$ is

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

Where $B(\alpha, \beta)$ is a short-name of the Beta *function*:

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

And the Gamma function:

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

We can show that $B\left(\alpha, \beta\right)$ is symmetric and satisfies:

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

We choose as the prior distribution on θ the Beta distribution:

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

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

A.1.3 A posteriori

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

But:

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

Hence:

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

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

Thus, if instead of considering a unique variable , we observe an i.i.d. sample of data, the joint distribution can be written as:

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

Let's introduce:

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

Then we get:

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

A.2 Special case of the Beta distribution

We remind that:

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

For $\alpha = \beta = 1$, we get a uniform prior.
For $\alpha = \beta > 1$, we get a bell curve.
For $\alpha = \beta < 1$, we get a U curve.

$\mathbb{E}\left[\theta\right] = \frac{\alpha}{\alpha+\beta}$
$\mathbb{V}\left[\theta\right] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{(\alpha+\beta)} \times \frac{\beta}{(\alpha+\beta)} \times \frac{1}{(\alpha+\beta+1)}$
For $\alpha > 1$ and $\beta > 1$, we get the mode: $\frac{\alpha-1}{\alpha+\beta-2}$.
In the case, let's write D for the data:

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

We can see that the a posteriori expectation of the parameter is a convex combination of the maximum likelihood estimator and the prior expectation. It converges asymptotically to the maximum likelihood estimator .

If we use a uniform prior distribution, $\mathbb{E}\left[\theta|D\right] = \frac{k+1}{n+2}$. Laplace proposed to correct the frequentist estimator, it seemed odd to him that he was not defined in the absence of data. He proposed to add two virtual observation (0 and 1) such that in the absence of data the estimator equals $\frac{1}{2}$. This correction is known as *Laplace's correction.*

The variance of the a posteriori distribution decrease in $\frac{1}{n}$.

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

We have chosen a sharper distribution around θ_M, in the same way than in a frequentist approach, the confidence intervals narrow around the estimator when the number of observations increase.

A.2.1 Playful propriety

$$p\left(x_1, x_2, \ldots, x_n\right) = \frac{B\left(k+\alpha, n-k+\beta\right)}{B\left(\alpha,\beta\right)} = \frac{\Gamma\left(\alpha+k\right)\Gamma\left(\beta+n-k\right)\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha+\beta+n\right)\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \tag{A.1}$$

Let's use this well-known property of the Gamma function:

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

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

such that

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

let's write $\alpha^{[k]} = \alpha\left(\alpha+1\right)\ldots\left(\alpha+k-1\right)$ and simplify the expression A.1:

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

We shall note the analogy with the Polya urn model: let us consider $(\alpha+\beta)$ balls of colour: α are black, β are white. When drawing a first black ball, the probability of the event is:

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

After the drawing, we put back the ball in the urn and we add a ball of the same colour. Let's imagine that we draw again a black ball then the probability of this event is:

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

However:

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

In more general case , we show by recurrence that the marginal probability of obtaining some sequence of colours by drawing from a Polya urn is exactly the marginal probability of obtaining the same result from the marginal model, obtained by integrating on a priori $theta$. First, this show that drawings from a Polya urn are exchangeable; Secondly, the mechanism of this type of urn, and its exchangeability, we'll be useful for the Gibbs sampling and for the same type of Bayesian models.

A.2.2 Conjugate priors

Let $\mathbb{F}$ be a set. We assume that $p\left(x|\theta\right)$ known, we deduce from that: $p\left(\theta\right)\in\mathbb{F}$ such that $p\left(\theta|x\right)\in\mathbb{F}$. We say that $p\left(\theta\right)$ is conjugated to the model $p\left(x|\theta\right)$.

Exponential model

Let's consider:

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

For $p\left(x|\theta\right)$, θ is the canonical parameter. For $p\left(\theta\right)$, α is the canonical parameter and θ is the sufficient statistic. Let us note that B do not stand for the Beta distribution.

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

Let us define:

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

Then:

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

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

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

Since the family is an exponential one,

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

θ_{MAP} results from:

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

Thus we get $\mu_{MAP}=\mu\left(\theta\right)$ in the previous equation. Consequently:

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

Univariate Gaussian

With and a priori on μ but not on σ^2

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

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

Thus:

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

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

Where:

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

Thus:

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

And:

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

Indeed, the variance decreases in $\frac{1}{n}$.

With an a priori on σ^2 but not onμ We get $p\left(\sigma^2\right)$ as an Inverse Gamma form.

With an a priori on μ and σ^2 Gaussian a priori on x and μ, Inverse Gamma a priori on σ^2. Please refer to the chapter 9 of the course handout (Jordan's polycopié).

Appendix B

B.1 A posteriori Maximum (MAP)

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

Because, with the Bayes rule:

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

The a posteriori maximum is not really Bayesian, it's rather a slight modification brought to the frequentist estimator.

B.1.1 Predictive probability

In the Bayesian paradigm, the probability of a future observation x^* will be estimated by the *Predictive probability*:

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

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

A sequential calculus is possible since:

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

Vocabulary:

- a priori information: $p\left(\theta|x_1,x_2,\ldots,x_{n-1}\right)$
- likelihood: $p\left(x_n|\theta\right)$
- a posteriori information: $p\left(\theta|x_1,x_2,\ldots,x_n\right)$

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

B.2 Naive Bayes

B.2.1 Introduction

Remarque: Contrary to its name, “Naive Bayes” is *not* a Bayesian method.

Let’s Consider the following problem of classification $x\in\mathbb{X}^p\longmapsto y\in\{1,2,\ldots,M\}$.

Here, $x=\left(x_1,x_2,\ldots,x_p\right)$ is a vector of descriptors (or features): $\forall i\in\{1,2,\ldots,p\},x_i\in\mathbb{X}$, with $\mathbb{X}=\{1,2,\ldots,K\}$ (or $\mathbb{X}=\mathbb{R}$).

Goal: Learn $p\left(y|x\right)$.

A very naive method will trigger off a combinatorial explosion: $\theta\in\mathbb{R}^{K^p}$.

Bayes formula gets us:

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

The Naive Bayes method consists in assuming that the features x_i are all conditionally independent from the class, hence:

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

Then, the Bayes formula gives us:

$$p\left(y|x\right)=\frac{p\left(y\right)\prod\limits_{i=1}^{p}p\left(x_i|y\right)}{p\left(x\right)}=\frac{p\left(y\right)\prod\limits_{i=1}^{p}p\left(x_i|y\right)}{\sum\limits_{y'}p\left(y'\right)\prod\limits_{i=1}^{p}p\left(x_i|y'\right)}$$

We consider the case where the features take discrete values. Consequently the new graphical model contains only discrete random variables. Then, we can write a discrete model as an exponential family. Indeed we can write:

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

and

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

We can see that the dummy functions $\delta(x_i = k, y = k')$ and $\delta(y = k')$ are the *sufficient statistics* of the joint distribution model for y and the variables x_i, where $\theta_{ikk'}$ and $\theta_{k'}$ are *canonical parameters*. Thus , we can write:

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

Where $A((\theta_{ikk'})_{i,k,k'}, (\theta_{k'})_{k'})$ is the log-partition function.

We have rewritten the joint distribution model of $(y, x_1, \ldots, x_p)$ as an exponential family. Given that the maximum of likelihood estimator of an exponential family, where the canonical parameters are not combined, is also the maximum entropy estimator; as seen in a previous course and provided that the statistical moments of the sufficient statistics equal their empirical moments.

Thus, if we introduce

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

The maximum likelihood estimator must satisfy the moment constraints

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

which define them completely.

Then, we can write the estimators of the canonical parameters as:

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

However, our goal is to obtain a classification model, that is to say, a model of only the conditional probability law. From the approximated generative model and applying the Bayes rule we can get:

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

We can re write the conditional model as an exponential family

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

Its sufficient statistics and canonical parameters are equal to those of the generative model, but seen as functions of the random variable y, given that x is fixed (we could write $\phi_{x,i,k,k'}(y) = \delta(x_i = k, y = k')$). As for the log-partition function, it is now equal to $\log p(x)$.

Warning: $\widehat{\theta}_{ikk'}$ is the maximum likelihood estimator in the generative model which, usually, is not equal to the maximum likelihood estimator in the conditional model.

B.2.2 Advantages and Drawbacks

Advantages:

- Doable in line.
- Computationally tractable solution.

Drawbacks:

- Generative: generative models produce good estimator whenever the model is “true”, or in statistical words *well specified*, which means that the process that generate the real data induce a distribution equal to the one of the generative model. When the model is not *well specified* (which is the most common case) we’d better use a discriminative method.

B.2.3 Discriminative method

The problem that we have considered in the previous section is the generative model for classification in K classes. How to learn, in a discriminatory way , a classifier in K classes? Is it possible to use an exponential family?

We have already seen the logistic regression for 2 classes classification:

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

Let’s study the K-multiclass logistic regression:

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

Although we have built the model from different staring consideration, the resulting modelling (that is the set of possible distribution) is of the same exponential family than the Naive Bayes model.

Nonetheless, the fitted model in a discriminatory approach will be different from the one fitted in a generative approach: the fitting of the K-multiclass logistic regression results from the maximisation of the likelihood of the classes $y^{(j)}$ of a set of learning, given that $x^{(j)}$ are fixed. In other words, the fitting is obtained by computing the maximum likelihood estimator in the conditional model. Unlike what happens in the generative model, the estimator can't be obtained in a analytical form and the learning requires solving a numerical optimisation problem.