Introduction to Bayesian learning
Lecture 2: Bayesian methods for (un)supervised problems

Anne Sabourin, Ass. Prof., Telecom ParisTech

September 2017

1. Lecture 1 Cont'd : Conjugate priors and exponential family.
   - Introduction, examples
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice
2. Lecture 1 Cont'd : A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality
3. Supervised learning example : Naive Bayes Classification
4. Bayesian linear regression
   - Regression : reminders
   - Bayesian linear regression
5. Bayesian model choice
   - Bayesian model averaging
   - Bayesian model selection
   - Automatic complexity penalty
   - Laplace approximation and BIC criterion
   - Empirical Bayes

Reminder : conjugate priors

last lecture (see lecture notes)

- Definition : hyper parameter, conjugate family
- examples :
    - normal model,known variance - normal prior
    - normal model, known mean - gamma prior on inverse variance
    - normal model , unknown mean and variance : normal-gamma prior on $(\mu, \lambda = 1/\sigma^2)$

conjugate priors for multivariate normal

- $X \sim \mathcal{N}(\mu, \Lambda^{-1})$, $\mu \in \mathbb{R}^d, \Lambda \in \mathbb{R}^{d \times d}$ positive, definite (precision matrix : inverse of covariance matrix)

1. unknown mean $\rightarrow$ conjugate prior family on μ : a multivariate Gaussian distributions

2. unknown precision $\rightarrow$ conjugate prior on Λ : Wishart distributions $\mathcal{W}(\nu, W)$ with ν degrees of freedom ($\nu \in \mathbb{N}^*$) and $W \in \mathbb{R}^{d \times d}$.

Wishart distribution

defined on the cone of positive definite matrices.

- The Wishart distribution $\mathcal{W}(\nu, W)$ has density

$$f_{\mathcal{W}}(\Lambda|\nu, W) = B \det \lambda^{(\nu-d-1)/2} \exp\left\{\frac{-1}{2}\mathrm{Tr}(W^{-1}\Lambda)\right\}$$

  w.r.t. Lebesgue on $\mathbb{R}^{\frac{d(d+1)}{2}}$: $\prod_{i\leq j} \mathrm{d}\Lambda_{(i,j)}$, restricted to the set of positive definite matrices.
- B : a normalizing constant.
- probabilistic representation : let M be a random $\nu \times d$ matrix with *i.i.d.*rows $M_{(i,\cdot)} \sim \mathcal{N}(0, W)$. Then

$$\Lambda \sim \mathcal{W}(\nu, W) \iff \Lambda \stackrel{\mathrm{d}}{=} M^\top M = \sum_{i=1}^{n} M_{(i,\cdot)}^\top M_{(i,\cdot)}$$

- More details : see *e.g.* *Eaton, Multivariate Statistics : A Vector Space Approach, 2007 (Chapter 8)*

conjugate priors for multivariate normal, Cont'd

3. Unknown mean and precision $\rightarrow$ conjugate prior family on (μ, Λ) : the Gaussian-Wishart distribution with hyper-parameters (W, ν, m, β)
$$\pi(\mu, \Lambda) = \pi_1(\Lambda)\pi_2(\mu|\lambda)$$
with
$$\boldsymbol{\pi}_2 = \mathcal{W}(W, \nu), \nu \in \mathbb{N}, W \text{ positive definite},$$
$$\boldsymbol{\pi}_2(\,\cdot\,|\Lambda) = \mathcal{N}(m, (\beta\Lambda)^{-1}), \qquad m \in \mathbb{R}^d, \beta > 0$$

Definition : exponential family

A dominated parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is an *exponential family* if the densities write

$$p_\theta(x) = C(\theta) h(x) \exp\left\{ \langle T(x), R(\theta) \rangle \right\}$$

for some functions

$$R : \Theta \to \mathbb{R}^k, \quad T : \mathcal{X} \to \mathbb{R}^k,$$
$$C : \Theta \to \mathbb{R}_+^*, \quad h : \mathcal{X} \to \mathbb{R}_+^*.$$

- $C(\theta)$: a normalizing constant
- $R(\theta)$: the *natural parameter* (R : the 'good' re-parametrization)
- If $R(\theta) = \theta$, the family is *natural*.

- Most textbook distributions are from the exponential family !

Example I : Bernoulli model

- $\theta \in \Theta =]0, 1[$, $\mathcal{X} = \{0, 1\}$
- The model is dominated by $\lambda = \delta_0 + \delta_1$

$$\begin{aligned} p_\theta(x) &= \theta^x (1-\theta)^{1-x} \\ &= \exp\{x \log \theta + (1-x) \log(1-\theta)\} \\ &= (1-\theta) \exp\left\{ \underbrace{x}_{T(x)} \underbrace{\log \frac{\theta}{1-\theta}}_{R(\theta)} \right\} \end{aligned}$$

- The model is an exponential family with
  - $T(x) = x$
  - natural parameter : $\rho = R(\theta) = \log \frac{\theta}{1-\theta}$.
  - normalizing constant $C(\theta) = (1-\theta)$

Example II : Gaussian model

- $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$
- the model is dominated by Lebesgue on $\mathcal{X} = \mathbb{R}$.

$$
\begin{aligned}
p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(x^2 - 2\mu x + \mu^2)}{2\sigma^2} \right\} \\
&= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-\mu}{2\sigma^2}\}}_{C(\theta)} \exp\left\langle \underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{T(x)}, \underbrace{\begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}}_{R(\theta)} \right\rangle
\end{aligned}
$$

- The model is an exponential family with
  - $T(x) = (x, x^2)$
  - natural parameter : $\rho = R(\theta) = (\mu/\sigma^2, -1/2\sigma^2)$.
  - normalizing constant $C(\theta) = (2\pi\sigma^2)^{-1/2}$

likelihood for *i.i.d.* samples in exponential families

$$p_\theta(x_1) = C(\theta)h(x)\exp\left\{\langle R(\theta), T(x_1)\rangle\right\}$$

$$\Rightarrow$$

$$p_\theta^{\otimes n}(x_{1:n}) = C(\theta)\underbrace{\prod_{i=1}^{n} h(x_i)}_{h_n(x_{1:n})}\exp\left\{\left\langle \underbrace{\sum_{i=1}^{n} T(x_i)}_{T_n(x_{1:n})}, R(\theta)\right\rangle\right\}.$$

Natural parameter space

- natural parametrization : $\rho = R(\theta)$.
- The density $p_\rho(x) = C(\rho)h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle$ integrates to 1 $\iff \rho \in \mathcal{E}$, the *natural parameter space*, *i.e.*

$$\mathcal{E} = \left\{\rho : \int_{\mathcal{X}} h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x) < \infty\right\}$$

- If $\mathcal{E}$ is open : the family is *regular*.

Maximum likelihood in regular exponential families

natural parametrization : $\rho = R(\theta)$. λ : reference measure.

lemma : expression for $\mathbb{E}_\rho\big[T(X)\big]$

$$\mathbb{E}_\rho\big[T(X)\big] = -\nabla_\rho\{\ln C(\rho)\}$$

Proof

$$1 \equiv C(\rho)\int_{\mathcal{X}} h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x)$$

(with regularity to exchange $\int$ and ∇) $\Rightarrow$)

$$0 = \nabla_\rho C(\rho)\underbrace{\int_{\mathcal{X}} h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x)}_{C(\rho)^{-1}} + C(\rho)\int_{\mathcal{X}} h(x)T(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x)$$

$$\Rightarrow 0 = \frac{1}{C(\rho)}\nabla_\rho C(\rho) + \mathbb{E}\big(T(X)\big)$$

□

Maximum likelihood in regular exponential families, cont'd

proposition

The MLE estimator $\widehat{\rho}$ in a regular exponential family satisfies

$$\mathbb{E}_{\widehat{\rho}}[T(X)] = \frac{1}{n}\sum_i T(x_i).$$

Proof

$$\begin{aligned}\nabla_\rho \log p_{\widehat{\rho}}(x) = 0 &\iff \nabla_\rho\{n \log C(\rho) + \langle \textstyle\sum T(x_i), \rho\rangle\} = 0 \\ &\iff \nabla_\rho \log C(\widehat{\rho}) = \frac{-1}{n}\sum_i T(x_i).\end{aligned}$$

then use the lemma. $\square$

Conjugate priors in exponential family

Proposition

A natural exponential family with densities
$p_\theta(x) = C(\theta)h(x)\exp\langle\theta, T(x)\rangle$, admits a conjugate prior family
$\mathcal{F} = \{\boldsymbol{\pi}_{\lambda,\mu}, \lambda > 0, \mu \in M_\lambda \subset \mathbb{R}^k\}$, with

$$\pi_{\lambda,\mu}(\theta) = K(\mu,\lambda)C(\theta)^\lambda \exp\left\{\langle\theta,\mu\rangle\right\}$$

and $M_\lambda = \{\mu : \int_\Theta \pi(\mu,\lambda)\,\mathrm{d}\theta < \infty\}$.
The posterior for n observation is

$$\pi_{\lambda,\mu}(\theta|x_{1:n}) \propto C(\theta)^{\lambda+n} \exp\left\{\langle\theta, \mu + \sum_i T(x_i)\rangle\right\}$$

so that $\boldsymbol{\pi}_{\lambda,\mu}(\,\cdot\,|x_{1:n}) = \boldsymbol{\pi}_{\lambda_n,\mu_n}(\,\cdot\,)$, with

$$\lambda_n = \lambda + n\,; \qquad \mu_n = \mu + \sum_i T(x_i)$$

proof exercise

Example : Poisson model

$$\begin{aligned} p_\theta(x) &= e^{-\theta}\theta^x/x!, \qquad \mathcal{X} = \mathbb{N}, \theta > 0 \\ &= \frac{1}{x!} e^{-\theta} e^{x \log \theta} \end{aligned}$$

$\rightarrow$ an exponential family with

$$T(x) = x\,, \qquad \rho = R(\theta) = \log\theta \in \mathbb{R}\,, \qquad C(\rho) = \exp\{-e^\rho\}$$

conjugate prior for ρ :

$$\pi_{a,b}(\rho) \propto \exp\{-be^\rho\}\ \exp\{a\rho\}.$$

Back to θ :

$$\pi(\theta) = \text{"}\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\rho}\frac{\mathrm{d}\rho}{\mathrm{d}\theta}\text{"} = \theta^{a-1}\exp\{-b\theta\} \quad \text{(Gamma density)}$$

$\rightarrow$ The Gamma family is a conjugate prior for θ.

About the choice of a conjugate prior

- A convenient choice only
- One must still choose hyper-parameters (λ, μ)
- This is an issue of *model choice*
- possible to do so via *empirical Bayes* methods, see lecture 2 and lab session.

Other prior choices : non informative priors

- Goal : minimize the bias induced by the prior

- If Θ compact : one can choose $\pi(\theta) = \text{Constant}$
- If Θ non compact, $\int_\theta \pi(\theta)\,\mathrm{d}\theta = \int_\Theta C\,\mathrm{d}\theta = +\infty$
  OK to do so as long as the posterior is well defined, *i.e.* when

$$\int_\Theta p_\theta(x)\,\mathrm{d}\pi(\theta) < \infty.$$

uniform only *w.r.t.* the reference measure $\rightarrow$ not invariant under re-parametrization.
e.g. Flat prior on $]0,1[$ in a $\mathcal{B}er(\theta)$ model $\rightarrow$ non flat over $\rho = \log[\theta/(1-\theta)]$

Other prior choices : Jeffreys prior

- For Θ open in $\mathbb{R}^d$. Reasonable with $d = 1$.

- Remind the Fisher information (in a regular model) :

$$I(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial \log p_\theta(X)}{\partial\theta}\Big)^2\Big] = -\mathbb{E}\Big[\frac{\partial^2 \log p_\theta(X)}{\partial\theta^2}\Big].$$

- $I(\theta)$ is the expected curvature of the likelihood around θ.
- Interpretation as a an average information carried by X about θ.
- Idea : grant more prior mass to highly informative θ's

Definition : Jeffreys prior

In a dominated model with densities $p_\theta, \theta \in \Theta$, the Jeffreys prior has densities *w.r.t.* Lebesgue on Θ :

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

- **exercise** compute the Jeffreys prior in the Bernoulli model, in the location model $\mathcal{N}(\theta, \sigma^2)$, σ^2 known and in the scale model $\mathcal{N}(\mu, \theta^2)$, μ known.

Invariance of the Jeffreys prior

- Change of variable : $h(\theta) = \eta$. Then $p_\theta = p_{h(\theta)}$.
- Let $\boldsymbol{\theta} \sim \boldsymbol{\pi}_{J,\theta}$ the Jeffreys prior. Then $\boldsymbol{\eta} \sim \boldsymbol{\pi}_{J,\theta} \circ h^{-1}$ with density

$$\pi(\eta) \overset{\text{for } \theta = h^{-1}(\eta)}{=} \pi_{J,\theta}(\theta) \frac{\mathrm{d}\theta}{\mathrm{d}\eta} = \frac{\sqrt{I(\theta)}}{h'(\theta)}$$

- On the other hand compute the Jeffreys prior on η :

$$\pi_{J,\eta}(\eta) = \sqrt{I_\eta(\eta)} = \mathbb{E}_\eta\Big[\Big(\frac{\partial \log p_\eta(X)}{\partial \eta}\Big)^2\Big]^{1/2}$$

$$\overset{\theta = h^{-1}(\eta)}{=} \mathbb{E}_\theta\Big[\Big(\frac{\partial \log p_\theta(X)}{\partial \theta} \frac{\mathrm{d}\theta}{\partial \eta}\Big)^2\Big]^{1/2} = \frac{\sqrt{I(\theta)}}{h'(\theta)}.$$

- Same result : the Jeffreys prior in the η parametrization is the image measure of the Jeffreys prior in the θ parametrization.
- In other words the Jeffreys prior is parametrization-invariant.

Rough overview

as the sample size $n \to \infty$

- The influence of the prior choice vanishes
- The posterior distribution concentrates around the true value θ_0 (almost surely)
- The posterior distribution is asymptotically normal with mean $\widehat{\theta} =$ the maximum likelihood, and variance $n^{-1} I(\theta)^{-1}$ (same as MLE's)

1. Lecture 1 Cont'd : Conjugate priors and exponential family.

2. Lecture 1 Cont'd : A glimpse at Bayesian asymptotics

   Example : Beta-Binomial model

   Posterior consistency

   Asymptotic normality

3. Supervised learning example : Naive Bayes Classification

4. Bayesian linear regression

5. Bayesian model choice

Reminder : Beta-Binomial model

- Bayesian model $\begin{cases} \boldsymbol{\theta} \sim \boldsymbol{\pi} = \mathcal{B}eta(a, b) \\ X|\theta \sim \mathcal{B}er(\theta). \end{cases}$

- ${\mathrm{P}_\theta}^\infty$: distribution over $\mathcal{X}^\infty$ of the random sequence $(X_n)_{n\geq 1} \overset{\text{i.i.d}}{\sim} \mathrm{P}_\theta$

- posterior distribution (conjugate prior) :

$$\boldsymbol{\pi}(\,\cdot\,|x_{1:n}) = \mathcal{B}er(a + s, b + n - s), \quad s = \sum_1^n x_i.$$

Posterior expectation and variance

$$\mathbb{E}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n}) = \frac{a+\sum_1^n X_i}{a+b+n}$$

$$= \frac{a/n + \frac{1}{n}\sum_1^n X_i}{(a+b)/n+1}$$

$$\xrightarrow[n\to\infty]{a.s.} \theta_0 \quad \text{under } \mathrm{P}_{\theta_0}^{\infty}$$

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n}) = \frac{\left(a+\sum_1^n X_i\right)\left(b+n-\sum_1 X_i\right)}{\left(a+b+n\right)^2\left(a+b+n+1\right)}$$

$$= \frac{1}{n}\frac{\left(a/n+\frac{1}{n}\sum_1^n X_i\right)\left(b/n+1-\frac{1}{n}\sum_1 X_i\right)}{\left((a+b)/n+1\right)^2\left((a+b+1)/n+1\right)}$$

$$\underset{\mathrm{P}_{\theta_0}^{\infty}-a.s.}{\sim} \frac{\theta_0(1-\theta_0)}{n} \underset{\text{exercise}}{=} (n\, I(\theta_0))^{-1}$$

Concentration of the posterior distribution

- Write $\boldsymbol{\theta}_n^* = \boldsymbol{\theta}_n^*(X_{1:n}) = \mathbb{E}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n})$.
- Tchebychev inequality $\Rightarrow \forall \delta > 0, \forall U = (\boldsymbol{\theta}_n^* - \delta, \boldsymbol{\theta}_n^* + \delta)$,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\pi}}\left(\boldsymbol{\theta} \notin U|X_{1:n}\right) &= \mathbb{P}_{\boldsymbol{\pi}}\left(\left(\boldsymbol{\theta} - \boldsymbol{\theta}_n^*\right)^2 > \delta^2|X_{1:n}\right) \\
&\leq \frac{\mathbb{V}\mathrm{ar}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n})}{n\delta^2} \\
&\underset{\mathrm{P}_{\theta_0}^{\infty}-a.s.}{\sim} \frac{\theta_0(1-\theta_0)}{n\delta^2} \qquad \xrightarrow[n\to\infty]{a.s.} 0.
\end{aligned}$$

- summary : $\mathrm{P}_{\theta_0}^{\infty}$ - a.s., we have that
  - The posterior distribution concentrates around the posterior expectancy $\boldsymbol{\theta}_n^*$
  - $\boldsymbol{\theta}_n^* \xrightarrow[n\to\infty]{} \theta_0$.

Posterior consistency

Definition

Let $\{\mathrm{P}_\theta, \theta \in \Theta\}, \boldsymbol{\pi}$ be a Bayesian model and let $\theta_0 \in \Theta$. The posterior is *consistent* at θ_0 if For all neighborhood U of θ_0,

$$\boldsymbol{\pi}(U|X_{1:n}) \xrightarrow[n\to\infty]{} 1, \quad \mathrm{P}_{\theta_0}^{\infty}\text{-a.s.}$$

- In general consistency holds when Θ is finite dimensional if $\boldsymbol{\pi}$ assigns positive mass to θ_0's neighborhoods.
- See *e.g.* [Ghosh and Ramamoorthi, 2003], Chapter 1.3, 1.4 for details

Doob's theorem

Theorem

If Θ and $\mathcal{X}$ are complete, separable, metric spaces endowed with their Borel σ-field, if $\theta \mapsto P_\theta$ is 1 to 1, then for any prior $\boldsymbol{\pi}$ on Θ, $\exists \Theta_0 \subset \Theta$ with $\boldsymbol{\pi}(\Theta_0) = 1$ such that for all $\theta_0 \in \Theta_0$, the posterior is consistent at θ_0.

- **issue** The $\boldsymbol{\pi}$-negligible set where consistency does not hold may be large.
- Under additional regularity conditions, consistency holds at a given θ_0.

Consistency at a given θ_0.

Theorem([Ghosh and Ramamoorthi, 2003], Th. 1.3.4)

Let Θ be compact, metric and $\theta_0 \in \Theta$. Let $T(x, \theta) = \log \frac{p_\theta(x)}{p_{\theta_0}(x)}$. Assume

1. $\forall x \in \mathcal{X}$, $\theta \mapsto T(x, \theta)$ is continuous
2. $\forall \theta \in \Theta$, $x \mapsto T(x, \theta)$ is measurable
3. $\mathbb{E}\left(\sup_{\theta \in \Theta} |T(\theta, X_1)|\right) < \infty$.

Then

1. The maximum likelihood estimator is consistent at θ_0 (CV in proba)
2. If $\theta_0 \in \mathrm{Supp}(\pi)$, then the posterior is consistent at θ_0.

Bayesian asymptotic normality : Overview

- Tells us about the rate of convergence of $\boldsymbol{\pi}(\cdot|X_{1:n})$ towards δ_{θ_0}.

- With a $\sqrt{n}$ re-scaling, a Gaussian limit centered at the MLE (under appropriate regularity conditions)

- Good references : [Van der Vaart, 1998], [Ghosh and Ramamoorthi, 2003], [Schervish, 2012]

Bernstein - Von Mises Theorem

(stated for $\Theta \subset \mathbb{R}$, similar statements for $\Theta \subset \mathbb{R}^d$).

Theorem

Under appropriate regularity conditions (detailed in [Ghosh and Ramamoorthi, 2003], Th. 1.4.2),
Let $\mathbf{s} = \sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}_n(X_{1:n}))$, with $\widehat{\theta}(X_{1:n})$ the MLE. Let $\pi^*(s|X_{1:n})$ be the posterior density of $\mathbf{s}$. Then

$$\int_{\mathbb{R}} \left| \pi^*(s|X_{1:n}) - \sqrt{\frac{I(\theta_0)}{2\pi}} e^{\frac{-s^2 I(\theta_0)}{2}} \right| \mathrm{d}s \xrightarrow[n\to\infty]{a.s.} 0 \text{ under } \mathrm{P}_{\theta_0}^{\infty}$$

- Interpretation : as $n \to \infty$,

$$\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}_n(X_{1:n}) \overset{d}{\approx} \mathcal{N}(0, I(\theta_0)^{-1}), \textit{ i.e.}$$

$$\boldsymbol{\theta} \overset{d}{\approx} \mathcal{N}\left(\widehat{\theta}_n, \frac{I(\theta_0)^{-1}}{n}\right)$$

- Multivariate case : similar result with multivariate Gaussian and Fisher information matrix.

Asymptotic normality of the posterior mean

$\theta_n^* = \mathbb{E}_{\boldsymbol{\pi}}[\boldsymbol{\theta}|X_{1:n}]$, $\quad \widehat{\theta}_n$: maximum likelihood.

Theorem

In addition to the assumptions of BVM Theorem, assume $\int_{\mathbb{R}} |\theta|\pi(\theta)\,\mathrm{d}\theta < \infty$. Then under $\mathrm{P}_{\theta_0}^{\infty}$,

1. $\sqrt{n}(\theta_n^* - \widehat{\theta}_n) \xrightarrow[n\to\infty]{} 0$ in probability
2. $\sqrt{n}(\theta_n^* - \theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I(\theta_0)^{-1})$.

Regularity conditions for BVM theorem

1. $\{x \in \mathcal{X} : p_\theta(x) > 0\}$ does not depend on θ

2. $L(\theta, x) = \log p_\theta(x)$ is three times differentiable *w.r.t.* θ in a neighborhood of θ_0.

3. $\mathbb{E}_{\theta_0}|\frac{\partial}{\partial\theta}L(\theta_0, X)| < \infty, \mathbb{E}_{\theta_0}|\frac{\partial^2}{\partial\theta^2}L(\theta_0, X)| < \infty$ and $\mathbb{E}_{\theta_0}\sup_{\theta\in(\theta_0-\delta,\theta_0+\delta)}\frac{\partial^3}{\partial\theta^3}L(\theta_0, X)| < \infty$

4. $\int_{\mathcal{X}}$ and ∂_θ may be interchanged.

5. $I(\theta_0) > 0$.

Remark : under these conditions the MLE is asymptotically normal, $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1}$ as well.

Setting

Not purely Bayesian framework : the training step is not necessarily Bayesian, only the prediction step is.

- Sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ (d features)
- some features may be categorical, some discrete, some continuous ...
- data $X_i = (X_{i,1}, \dots X_{i,d})$, $i = 1, \dots, n$.
- Classification problem : X_i may come from anyone of K classes $(\mathcal{C}_1, \dots, \mathcal{C}_K)$.
- Example $\begin{cases} X_{i,1} \in \mathbb{R}^{p \times p} : & \text{X-ray image from patient } i \\ X_{i,2} \in \{0, 1\} : & \text{result of a blood test from patient } i. \end{cases}$
- classes : {ill, healthy, healthy carrier}.
- **Goal** predict the class $c \in \{1, \dots, K\}$ of a new patient.

Naive Bayes assumption

Conditionally to the class $c(i) \in \{1, \ldots, K\}$ of observation i, the features $(X_{i,1}, \ldots, X_{i,d})$ are independent.

- Looks like a strong (and erroneous) assumption!
- In practice : produces reasonable prediction (even though the posterior probabilities of each class are not to be taken too seriously)

1. Training step

- Training set $\{(x_{i,j}, c(i)),\ i \in \{1,\ldots,n\},\ j \in \{1,\ldots,d\}\}$, $c(i) \in \{1,\ldots,K\}$.

- for $k \in \{1,\ldots,K\}$:
  - Retain observations of class $k \to i \in I_k$.
  - For $j \in \{1,\ldots,d\}$ estimate the class distribution, with density

    $$p_{j,k}(x_j) = p(x_{i,j} | c(i) = k),$$

    using data $(x_{i,j})_{i \in I_k}$, usually in a parametric model with parameter $\theta_{j,k}$: $\to$ estimated density $p_{j,k,\widehat{\theta}_{j,k}}(\,\cdot\,)$
  - **output :** the conditional distribution of X given $C = k$,

    $$p_k(x) = \prod_{j=1}^{k} p_{j,k,\widehat{\theta}_{j,k}}(x_j)$$

2. computing the predictive class probabilities

input :

- new data point $x = (x_1, \dots, x_d)$
- From step 1 : conditional distributions of X given $C = k$:
  $p_k(\cdot) = \prod p_{j,k,\widehat{\theta}_{j,k}}$ (plug-in method, neglect estimation error of $\widehat{\theta}_{j,k}$).

(a) Assign a prior probability to each class : $\pi = (\pi_1, \dots, \pi_K)$,
$\pi_k = \mathbb{P}_{\boldsymbol{\pi}}(C = k)$.
step 1 $\rightarrow$ joint density of (X, C) : $q(x, k) = \pi_k p_k(x)$.

(b) Apply the discrete Bayes formula :

$$\pi(k|x) = \frac{\pi_k p_k(x)}{\sum_{c=1}^{K} \pi_c p_c(x)} = \frac{\pi_k \prod_{j=1}^{d} p_{j,k,\widehat{\theta}_{j,k}}(x_j)}{\sum_{c=1}^{K} \pi_c \prod_{j=1}^{d} p_{j,c,\widehat{\theta}_{j,c}}(x_j)}$$

Easy to implement ! $O(kdN)$ for N testing data.

3. final step : class prediction

- Classification task : output= a predicted class $\widehat{x}$
- Naive Bayes prediction for a new point x

$$\widehat{c} = \underset{k \in \{1,\dots,k}{\operatorname{argmax}} \pi(k|x).$$

(a maximum a posteriori)

Example : text documents classification

- 2 classes : $\{1 = \text{ spam }, 2 = \text{ non spam }\}$
- vocabulary $\mathcal{V} = \{w_1, \ldots, w_V\}$.
- dataset : documents (email) $T_i = (T_{i,j}, j = 1, \ldots, N_i)$, $i \leq n$ with
  - N_i : number of words in T_i
  - $t_{i,j} \in \mathcal{V}$: j^{th} word in T_i

Conditional model (text documents)

- Naive Bayes assumption : in document T_i, conditionally to the class, words are drawn independently from each other the vocabulary $\mathcal{V}$

- T_i can be summarized by a 'bag of words' $X_i = (X_{i,1}, \ldots, X_{i,V})$:

$$X_{i,j} : \text{number of occurrences of word } j \text{ in } T_i.$$

- Conditional model for X_i given its class $k \in \{1, 2\}$:

$$\mathcal{L}(X_i | C = k) = \mathcal{M}ulti\big(\theta_k = (\theta_{1,k}, \ldots, \theta_{V,k}), N_i\big), \quad i.e.$$

$$p_{k,\theta_k}(x) = \frac{N_i!}{\prod_{j=1}^{V} x_{i,j}!} \prod_{j=1}^{V} \theta_{j,k}^{x_{i,j}}$$

1. training step (text documents)

Fit separately 2 Multinomial models on spam and non-spam

- Here : the Dirichlet prior $\mathcal{D}iri(a_1 \dots, a_v)$, $a_j > 0$ is conjugate for the Multinomial model, with density

$$diri(\theta | a_1, \dots, a_V) = \frac{\Gamma(\sum_{j=1}^{V} a_j)}{\prod_{j=1}^{V} \Gamma(a_j)} \prod_{j=1}^{V} \theta_j^{a_j - 1}$$

on $\mathcal{S}_V = \{\theta \in \mathbb{R}_+^V : \sum_{j=1}^{V} \theta_j = 1\}$ the $V-1$-simplex.

- Mean of $\boldsymbol{\theta}$ under $\boldsymbol{\pi} = \mathcal{D}iri(a_1, \dots, a_V)$:

$$\mathbb{E}_{\boldsymbol{\pi}}(\boldsymbol{\theta}) = \left(\frac{a_1}{\sum_j a_j}, \dots, \frac{a_V}{\sum_j a_j} \right)$$

- The posterior for $x_{1:n} = (x_{i,1}, \dots, x_{i,V})_{i \in \{1, \dots, n\}}$ is

$$\mathcal{D}iri\Big((a_1 + \sum_{i=1}^{n} x_{i,1}),\ \dots,\ (a_V + \sum_{i=1}^{n} x_{i,V}) \Big).$$

1. training step (text documents) Cont'd

- Concatenate documents of each class separately

$$\rightarrow \quad x^{(k)} = (x_j^{(k)})_{j=1,\dots,V} \;, \quad k = 1, 2$$

with $x_{k,j}$ = total # occurrences of word j in documents of class k.

- $\theta_k = (\theta_{k,1}, \dots, \theta_{k,V})$ multinomial parameter for class k.

- Flat priors on $\boldsymbol{\theta}_k$: $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \mathcal{D}iri(1, \dots, 1)$

- Posterior mean estimates

$$\widehat{\theta}_k = \mathbb{E}_{\boldsymbol{\pi}_k}[\boldsymbol{\theta}|x^{(k)}] = \left(\frac{x_1^{(k)} + 1}{V + \sum_{j=1}^{V} x_j^{(k)}}, \dots, \frac{x_V^{(k)} + 1}{V + \sum_{j=1}^{V} x_j^{(k)}} \right)$$

(the prior acts as regularizer : '+1' term avoids 0 probabilities.

2. Prediction step

- For a new document x^{new} the predictive probabilities of each class are :

$$\pi(C = k|x^{new}) = \frac{p(x^{new}|C = k)\pi_1}{p(x^{new}|C = k)\pi_1 + p(x^{new}|C = 2)\pi_2}$$

with

$$p(x^{new}|C = k) \propto \prod_{j=1}^{V} \widehat{\theta_{k,j}}^{x_j^{new}}$$

- The class prediction is

$$k^*(x^{new}) = \underset{k=1,2}{\operatorname{argmax}}\, p(x^{new}|C = k)$$

The regression problem

- Supervised learning : training dataset (x_i, Y_i), $i \leq n$, with
  - $x_i \in \mathcal{X}$ the features for observation i (considered non random)
  - $Y_i \in \mathbb{R}$ the label (random variable).
- **goal** : for a new observation with features x_{new}, predict Y_{new}, *i.e.* construct a *regression function* $h \in \mathcal{H}$, so that $h(x)$ is our best prediction of Y at point x.
- h should
  - be simple (avoid over-fitting) $\rightarrow$ simple class $\mathcal{H}$.
  - fit the data well : measured through a loss function $L(x, y, h)$.
    example : squared error loss $L(x, y, h) = (y - h(x))^2$.

Multiple classical strategies

- Statistical learning approach : empirical risk minimization

$$R_n(x_{1:n}, y_{1:n}, h) = \frac{1}{n} L(x_i, y_i, h)$$
$$\to \underset{h \in \mathcal{H}}{minimize} \qquad R_n(x_{1:n}, y_{1:n}, h)$$

- Probabilistic modeling approach (likelihood based) : assume *e.g.*

$$Y_i = h_0(x_i) + \epsilon_i \ ,$$

$\epsilon_i \sim P_\epsilon$ independent noises, *e.g.* $P_\epsilon = \mathcal{N}(0, \sigma^2)$, σ^2 known or not.

$\to$ likelihood of h, $p_h(x_{1:n}, y_{1:n}) = \prod_{i=1}^n p_\epsilon(y_i - h(x_i))$.

$$\to \underset{h \in \mathcal{H}}{minimize} \quad -\sum_{i=1}^{n} \log p_\epsilon(y_i - h(x_i))$$

- With Gaussian noises, both strategies coincide.

Linear regression

- h : a linear combination of basis functions $\phi_j : \mathcal{X} \mapsto \mathbb{R}$ (feature maps), $j \in \{1, \ldots, p\}$

$$h(x) = \sum_{j=1}^{p} \theta_j \phi_j(x), \quad \theta_j \text{ unknown}, \quad \phi_j \text{ known}, \qquad i.e.$$

$$\mathcal{H} = \left\{ \sum_{j=1}^{p} \theta_j \phi_j : \quad \theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p \right\}$$

- Examples
  - $\mathcal{X} = \mathbb{R}^p, \quad \phi_j(x) = x_j$: canonical feature map
  - $\mathcal{X} = \mathbb{R}, \quad \phi_j(x) = x^{j-1}$: polynomial basis function
  - $\mathcal{X} = \mathbb{R}^d, \quad \phi_j(x) = \frac{1}{(2\pi)^{d/2} \det \Sigma_j} \exp -\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j)$, Gaussian basis function

Empirical risk minimization for linear regression

- Empirical risk :

$$R_n(x_{1:n}, y_{1:n}, \theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle \theta, \phi(x_i) \rangle)^2 = \frac{1}{2} \| y_{1:n} - \Phi\theta \|^2,$$

  with $\Phi \in \mathbb{R}^{n \times p}$: design matrix, $\Phi_{i,j} = \phi_j(x_i)$.
- Minimizer of R_n : the *least squares* estimator
- explicit solution when $\Phi^\top \Phi$ is of rank p (invertible)

$$\widehat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top y_{1:n}$$

Regularization

- **goals** : prevent
  - over-fitting
  - numerical instabilities (inversion of $(\Phi^\top \Phi)$.
- Add a complexity penalty (function of θ) to the empirical risk
- penalty :$\lambda \|\theta\|_2^2 \rightarrow$ ridge regression
- penalty :$\lambda \|\theta\|_1 \rightarrow$ Lasso regression
- *e.g.* with L_2 penalty, the optimization problem becomes

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \|y_{1:n} - \Phi\theta\|^2 + \lambda \|\theta\|_2^2 \quad \text{for some } \lambda > 0.$$

$\rightarrow$ solution $\widehat{\theta} = \left[\Phi^\top \Phi + \lambda I_p\right]^{-1} \Phi^\top y_{1:n}$.

Bayesian linear model

- Again, $Y_i = \langle \theta, \Phi(x_i) \rangle + \epsilon_i$
- Assume $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$, $\beta > 0$ noise precision viewed as a constant (known or not)
- Prior distribution on $\boldsymbol{\theta} \in \mathbb{R}^p$: $\boldsymbol{\pi} = \mathcal{N}(m_0, S_0)$.
- independence assumption : $\epsilon_1 \perp\!\!\!\perp \epsilon_2 \perp\!\!\!\perp \cdots \perp\!\!\!\perp \boldsymbol{\theta}$.
- $Y = Y_{1:n} = \Phi\theta + \epsilon_{1:n}$, with $\Phi \in \mathbb{R}^{n \times p}, \Phi_{i,j} = \phi_j(x_i)$.

Bayesian model

$$\begin{cases} \boldsymbol{\theta} \sim \boldsymbol{\pi} = \mathcal{N}(m_0, S_0) \\ \mathcal{L}\big[Y|\theta\big] = \mathcal{N}(\Phi\theta, \frac{1}{\beta} I_n) \end{cases}$$

- Natural Bayesian estimator : $\widehat{\theta} = \mathbb{E}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|Y_{1:n})$.
  $\rightarrow$ posterior distribution ?

Conditioning and augmenting Gaussian vectors

Lemma

Let

$$\begin{cases} W \sim \mathcal{N}(\mu, \Lambda^{-1}) \\ \mathcal{L}[Y|w] = \mathcal{N}(Aw + b, L^{-1}) \end{cases}$$

i.e. $Y = AW + b + \epsilon$ with $\epsilon \sim \mathcal{N}(0, L^{-1}) \perp\!\!\!\perp W$.

Then $\mathcal{L}[W|y] = \mathcal{N}(m_y, S)$ with

$$S = (\Lambda + A^\top \Lambda A)^{-1}$$

$$m_y = S[A^\top L(y - b) + \Lambda\mu.]$$

proof : homework (see exercises sheet online)

Application to posterior computation

Using the lemma with

$$A = \Phi, \quad b = 0, \quad W = \boldsymbol{\theta}, \quad \Lambda = S_0^{-1}, \quad \mu = m_0, \quad L = \beta I_p,$$

we obtain immediately the posterior distribution

$$\boldsymbol{\pi}(\,\cdot\,|Y_{1:n}) = \mathcal{L}[\boldsymbol{\theta}|y_{1:n}] = \mathcal{N}(m_n, S_n)$$

with

$$\begin{cases} S_n = \left(S_0^{-1} + \beta \Phi^\top \Phi\right)^{-1} \\ m_n = S_n\left(\beta \Phi^\top y_{1:n} + S_0^{-1} m_0\right) \end{cases} \tag{1}$$

Posterior mean estimate

$$\widehat{\theta} = \mathbb{E}_{\boldsymbol{\pi}}[\boldsymbol{\theta}|y_{1:n}] = m_n$$

Special case : diagonal, centered prior

- choose $m_0 = 0$, $S_0 = \alpha^{-1} I_p$, with α : prior precision (it makes sense !)
- Then (1) becomes

$$\begin{cases} S_n = \left(\alpha I_p + \beta \Phi^\top \Phi\right)^{-1} & = & \beta^{-1}\left(\frac{\alpha}{\beta} + \Phi^\top \Phi\right)^{-1} \\ m_n = S_n\left(\beta \Phi^\top y_{1:n}\right) & = & \underbrace{\left(\frac{\alpha}{\beta} + \Phi^\top \Phi\right)^{-1} \Phi^\top y_{1:n}}_{\text{penalized least squares solution}} \end{cases} \tag{2}$$

Adding a prior $\mathcal{N}(0, \alpha^{-1} I_p)$

$$\Longleftrightarrow$$

Adding a L_2 regularization with parameter $\lambda = \alpha/\beta$.

remark : Narrow prior $\Longleftrightarrow$ large α $\Longleftrightarrow$ large penalty

Predictive distribution

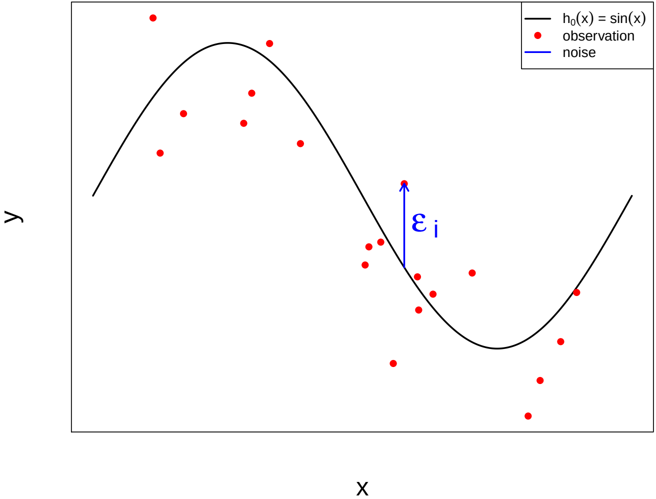
New data point (x_{new}, Y_{new}), with Y_{new} not observed and x_{new} known :

- **goal** : obtain the posterior distribution of Y_{new} (mean and variance$\rightarrow$ credible intervals).
- We still have $Y_{new} = \langle \boldsymbol{\theta}, \phi(x_{new}) \rangle + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \beta^{-1})$ and $\epsilon \perp\!\!\!\perp \boldsymbol{\theta}$.
- Now (after training step) $\boldsymbol{\theta} \sim \boldsymbol{\pi}(\,\cdot\,|y_{1:n}) = \mathcal{N}(m_n, S_n)$
- Thus $Y_{new} \overset{\mathrm{d}}{=}$ linear transform of Gaussian vector $(\epsilon, \boldsymbol{\theta})$

$$\mathcal{L}[Y_{new}|y_{1:n}] = \mathcal{N}\left(\phi(x_{new})^\top m_n, \;\; \phi(x_{new})^\top S_n \phi(x_{new}) + \beta^{-1}\right)$$

Example : polynomial basis functions

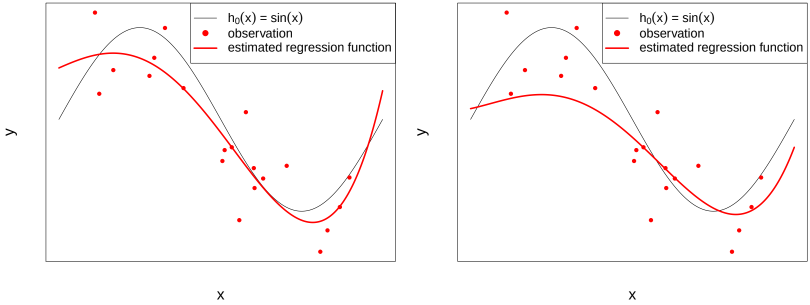
- True regression functions : $h_0(x) = \sin(x)$
- Polynomial basis functions : $\phi(x) = (1, x, x^2, x^3, x^4)$ $(p = 5)$.



Estimated regression function

- $\widehat{h}(x) = \langle \widehat{\theta}, \Phi(x) \rangle = \widehat{\theta}_1 + \sum_{j=2}^{5} \widehat{\theta}_j x^{j-1}$

- With the previous dataset

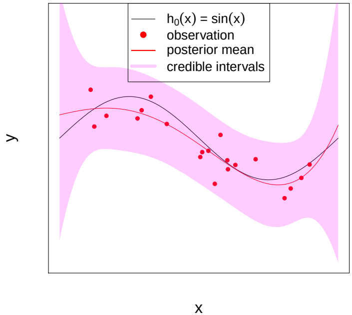


$\alpha = 0.01$ $\alpha = 100$

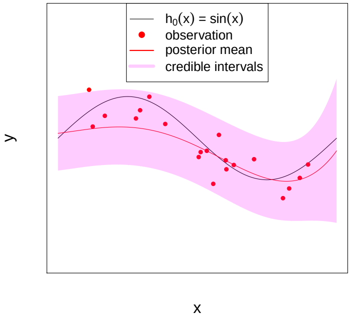
Predictive distribution

- $\widehat{h}(x)$: the mean of $\mathcal{L}(Y_{\text{new}}|y_{1:n})$ for $x_{\text{new}} = x$
- Remind $\mathcal{L}(Y_{\text{new}}|y_{1:n}) = \mathcal{N}(\widehat{h}(x), \sigma^2_{new} = \phi(x)^\top S_n \phi(x) + \beta^{-1})$
- $\rightarrow$ posterior credible interval for Y ,

$$I_x = \left[\widehat{h}(x) - 1/96\sqrt{\sigma^2_{new}}, \widehat{h}(x) + 1/96\sqrt{\sigma^2_{new}}\right]$$



$\alpha = 0.01$



$\alpha = 100$

Model choice problem

- What if several model in competition $\{M_k, k \in \{1, \dots, K\}\}$, with $M_k = \{\Theta_k, \boldsymbol{\pi}_k\}$?
- Continuous case : family of models $\{M_\alpha, \alpha \in \mathcal{A}\}$
- $\rightarrow$ How to choose k or α ?
- Examples :
  - $M_1 = \{\Theta, \boldsymbol{\pi}_1\}$, $M_2 = \{\Theta, \boldsymbol{\pi}_2\}$ with $\boldsymbol{\pi}_1$ a flat prior and $\boldsymbol{\pi}_2$ the Jeffreys prior
  - M_α linear model with normal prior on the noise $\mathcal{N}(0, \alpha^{-1})$

Hierarchical models

- Bayesian view : put a prior on unknown quantities, then condition upon data.
- Model choice problem : put a 'hyper-prior' on $\alpha \in \mathcal{A}$ (or $k \in \{1, \ldots, K\}$) $\rightarrow$ hierarchical Bayesian model
- Convenient when dealing with parallel experiments

Example of hierarchical model

Example : 2 rivers with fishes.

- $X_i \in \{0, 1\}$: fished fish ill or sound.
- $X_i \sim \mathcal{B}er(\theta)$, with $\theta = \theta_1$ in river 1 and $\theta = \theta_2$ in river 2.
- θ_1 and θ_2 are 2 realizations of $\boldsymbol{\theta} \sim \mathcal{B}eta(a, b)$
- $\alpha = (a, b)$: hyper-parameter for the prior
- hierarchical Bayes : put a prior on α (*e.g.* product of 2 independent Gammas).

Posterior mean estimates in a BMA framework

- denote $\boldsymbol{\pi}^h$ the hyper-prior on k (or α)
- Let us stick to the discrete case , $k \in \{1, \dots, M\}$.
- The prior is a mixture distribution $\boldsymbol{\pi} = \sum_{k=1}^{K} \pi^h(k)\boldsymbol{\pi}_k(\,\cdot\,)$, *i.e.* for all $\boldsymbol{\pi}$-integrable function $g(\boldsymbol{\theta})$,

$$\mathbb{E}_{\boldsymbol{\pi}}[g(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\pi}^h}\Big[\mathbb{E}_{\boldsymbol{\pi}}[g(\boldsymbol{\theta})|k]\Big] = \sum_{k=1}^{K} \pi^h(k) \int_{\Theta_k} g(\theta)\,\mathrm{d}\pi_k(\theta)$$

- So is the posterior distribution, thus the posterior mean is a weighted average

$$\begin{aligned}\widehat{g} = \mathbb{E}_{\boldsymbol{\pi}}[g(\boldsymbol{\theta})|X_{1:n}] &= \mathbb{E}_{\boldsymbol{\pi}^h}\Big[\mathbb{E}_{\boldsymbol{\pi}}[g(\boldsymbol{\theta})|k, X_{1:n}]|X_{1:n}\Big] \\ &= \sum_{k=1}^{K} \pi^h(k|X_{1:n}) \underbrace{\int_{\Theta_k} g(\theta)\,\mathrm{d}\pi_k(\theta|X_{1:n})}_{\widehat{g}_k\text{:posterior mean in model } k}\end{aligned}$$

Model evidence

Computing the posterior mean in the BMA framework requires

- Computing the posterior means in each individual model
  $\rightarrow k$ 'moderate' tasks
- Averaging them with weights $\pi^h(k|X_{1:n})$, *posterior weight of model* k
- Bayes formula

$$\pi^h(k|X_{1:n}) = \frac{\pi^h(k)p(X_{1:n}|k)}{\sum_{j=1}^K \pi^h(j)p(X_{1:n}|j)}$$

with

$$\begin{aligned} p(X_{1:n}|k) &= \text{ evidence of model } k \\ &= \int_{\Theta_k} p(X_{1:n}|\theta)\,\mathrm{d}\pi_k(\theta) \\ &= m_k(X_{1:n}) \text{ marginal likelihood of } X_{1:n} \text{ in model } k \end{aligned}$$

hard to compute (integral)

Shortcomings of BMA

- Inference has to be done in each individual model
- Usually one weight (say $\pi(k^*|X_{1:n})$) $\gg$ all others (reason : concentration of the posterior around the true $\theta_0 \in \Theta_{k_0}$ and $k^* = k_0$ $\implies$ final estimate $\widehat{g} \approx \widehat{g}_{k_0}$. Other $\widehat{g}_k$'s are almost useless

Bottleneck : compute k^*.
model choice problem.

Posterior weights, model evidence and Bayes factor

Recall $$k^* = \operatorname*{argmax}_k \pi(k|X_{1:n}) = \operatorname*{argmax}_k \underbrace{p(X_{1:n}|k)}_{\text{evidence of model } k} \pi^h(k)$$

- Uniform prior on $k \implies$ only the evidence $p(X_{1:n}|k)$ matters.
- in any case : prior influence vanishes with n.
- Relevant quantity to compare model k and j :

$$B_{kj} = \frac{p(X_{1:n}|k)}{p(X_{1:n}|j)} : \qquad \text{Bayes factor (Jeffreys, 61)}$$

- Suggested scale for decision making :

$\log_{10} B_{kj}$	B_{kj}	evidence against B_j
$0 \to 1/2$	$1 \to 3.2$	not significant
$1/2 \to 1$	$3.2 \to 10$	substantial
$1 \to 2$	$10 \to 100$	strong
> 2	> 100	decisive

Occam's razor principle

Between 2 models explaining the data equally well,
one ought to choose the simplest one.

$\rightarrow$ Avoid over-fitting

$\rightarrow$ Better generalization properties.

Occam's razor and model evidence

- When selecting k^* according to the model evidences $p(X_{1:n}|k)$, the Occam's razor is automatically implemented.

- Reason : the prior plays the role of a regularizer.

automatic complexity penalty : intuition 1

$$
\begin{aligned}
\text{Complex model} &\implies \text{large } \Theta_k \\
&\implies \text{small } \pi_k(\theta) \text{ (if uniform over } \Theta_k) \\
&\implies \int_{\Theta_k} p_\theta(x_{1:n})\pi_k(\theta)\,\mathrm{d}\theta \text{ small}
\end{aligned}
$$

(average over large regions where $p_\theta(x_{1:n})$ small)

automatic complexity penalty : intuition 2

- if $\Theta_k \subset \mathbb{R}$: assume
  - π_k flat over interval of length $\Delta\theta_k^{prior}$
  - $p_{\theta_k}(X_{1:n})$ peaked around $p_{\widehat{\theta}_{MAP,k}}(X_{1:n})$ with 'width' $\Delta_k^{posterior}$.
- then $\pi_k(\theta) \approx 1/\Delta_k^{prior}$ and

$$p(X_{1:n}|k) = \int_{\Theta_k} p_\theta(x)\pi_k(\theta)\,\mathrm{d}\theta \approx p_{\widehat{\theta}_{MAP,k}}(X_{1:n}) \qquad \underbrace{\frac{\Delta\theta_k^{posterior}}{\Delta\theta_k^{prior}}}_{\text{complexity penalty}}$$

- If $\Theta_k \subset \mathbb{R}^d$ and same approximation in each dimension

$$\log p(X_{1:n}|k) \approx \log p_{\widehat{\theta}_{MAP,k}}(X_{1:n}) + \qquad \underbrace{d \log \frac{\Delta\theta_k^{posterior}}{\Delta\theta_k^{prior}}}_{\text{dimension + complexity penalty}}$$

1. Lecture 1 Cont'd : Conjugate priors and exponential family.

2. Lecture 1 Cont'd : A glimpse at Bayesian asymptotics

3. Supervised learning example : Naive Bayes Classification

4. Bayesian linear regression

5. Bayesian model choice
   - Bayesian model averaging
   - Bayesian model selection
   - Automatic complexity penalty
   - Laplace approximation and BIC criterion
   - Empirical Bayes

Bibliography

[Ghosh and Ramamoorthi, 2003] Ghosh, J. and Ramamoorthi, R. (2003).
Bayesian nonparametrics. 2003.

[Schervish, 2012] Schervish, M. J. (2012).
Theory of statistics.
Springer Science & Business Media.

[Van der Vaart, 1998] Van der Vaart, A. W. (1998).
Asymptotic statistics, volume 3.
Cambridge university press.