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

Anne Sabourin, As. Prof., Telecom ParisTech

September 2019

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 $X_i = (X_{i,1}, \ldots X_{i,d})$, $i = 1, \ldots, n$.
- Classification problem : X_i may come from anyone of K classes $(\mathcal{C}_1, \ldots, \mathcal{C}_K)$.
- 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 $c \in \{1, \ldots, K\}$ 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 $\big\{(x_{i,j}, c(i)),\ i \in \{1,\dots,n\},\ j \in \{1,\dots,d\}\big\}$, $c(i) \in \{1,\dots,K\}$.

- for $k \in \{1,\dots,K\}$:
  - Retain observations of class $k \to i \in I_k$.
  - For $j \in \{1,\dots,d\}$ 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 $x = (x_1, \ldots, x_d)$
- 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 : $\pi = (\pi_1, \ldots, \pi_K)$,
$\pi_k = \mathbb{P}_{\boldsymbol{\pi}}(C = k)$.
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

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

(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 in 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 $\mathcal{D}iri(a_1 \ldots, a_V)$, $a_j > 0$ is conjugate for the Multinomial model, with density

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

  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 $\boldsymbol{\pi} = \mathcal{D}iri(a_1, \ldots, a_V)$:

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

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

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

1. training step (text documents) Cont'd

- Concatenate documents of each class separately

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

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

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

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

- Posterior mean estimates

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

(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

$$R_n(x_{1:n}, y_{1:n}, h) = \frac{1}{n} \sum_{1}^{n} L(x_i, y_i, h)$$

$$\rightarrow \underset{h \in \mathcal{H}}{minimize} \qquad R_n(x_{1:n}, y_{1:n}, h)$$

- 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.

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

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

- With Gaussian noises, both strategies coincide.

Linear regression

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

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

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

- 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

$$S = (\Lambda + A^\top L A)^{-1}$$
$$m_y = S[A^\top L(y - b) + \Lambda\mu.]$$

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

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

penalized least squares solution

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

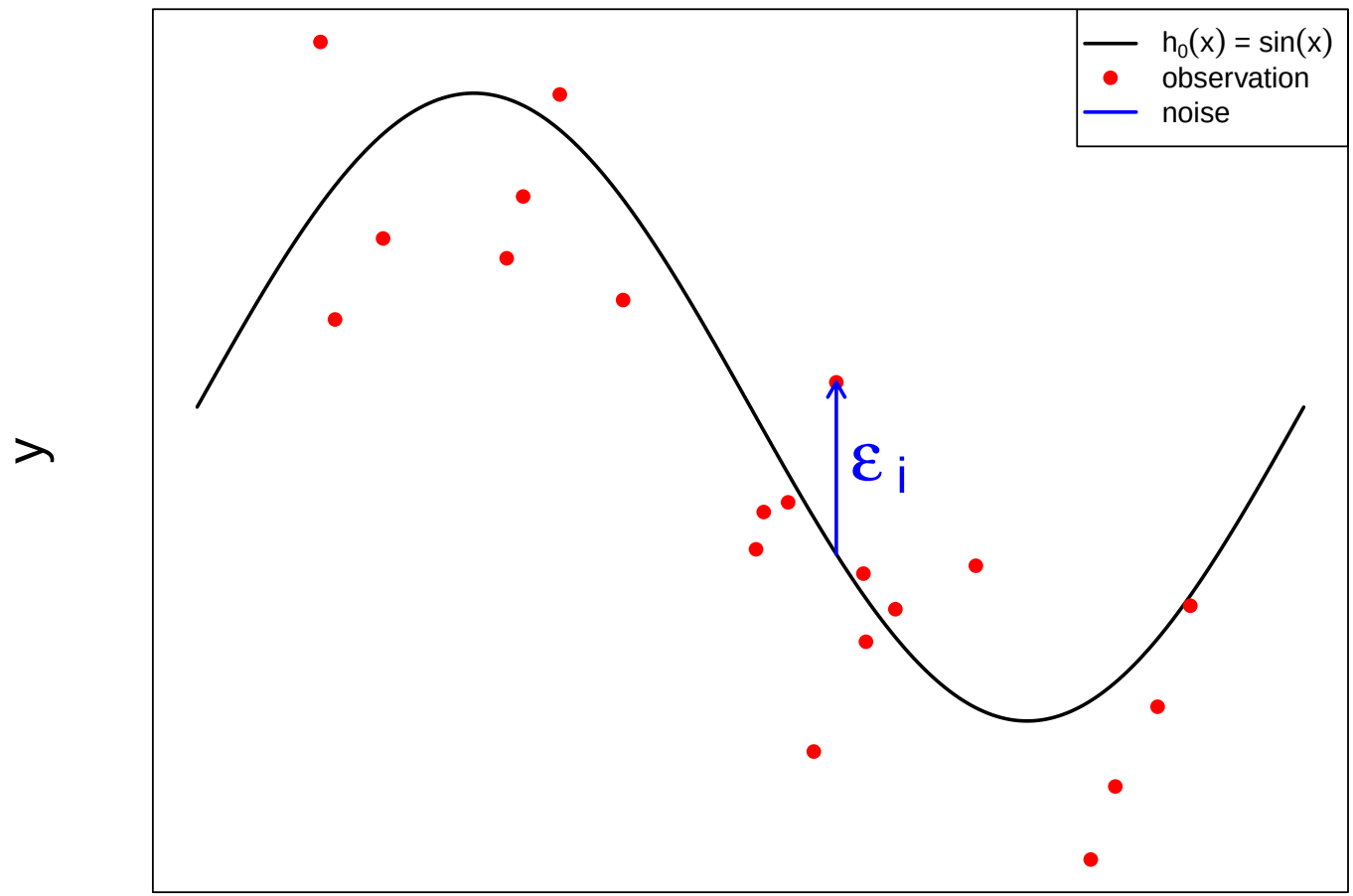
New data point (x_{new}, Y_{new}), with Y_{new} not observed and x_{new} known :

- **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

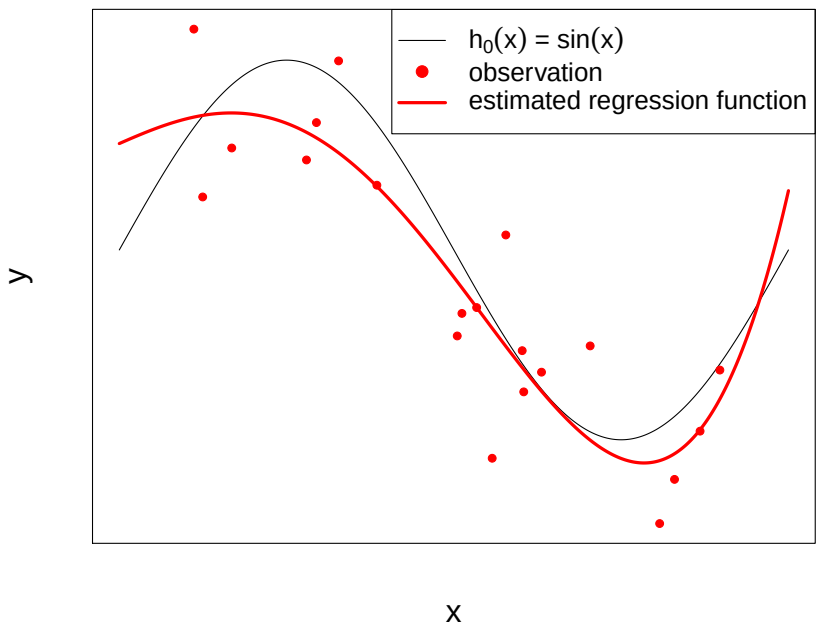
- True regression functions : $h_0(x) = \sin(x)$
- Polynomial basis functions : $\phi(x) = (1, x, x^2, x^3, x^4)$ $(p = 5)$.



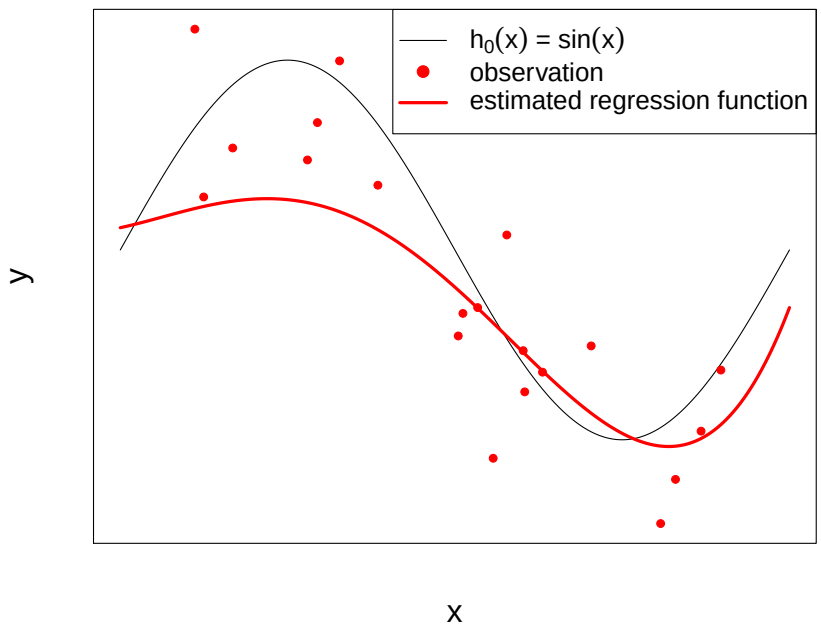
Estimated regression function

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

- With the previous dataset



$\alpha = 0.01$

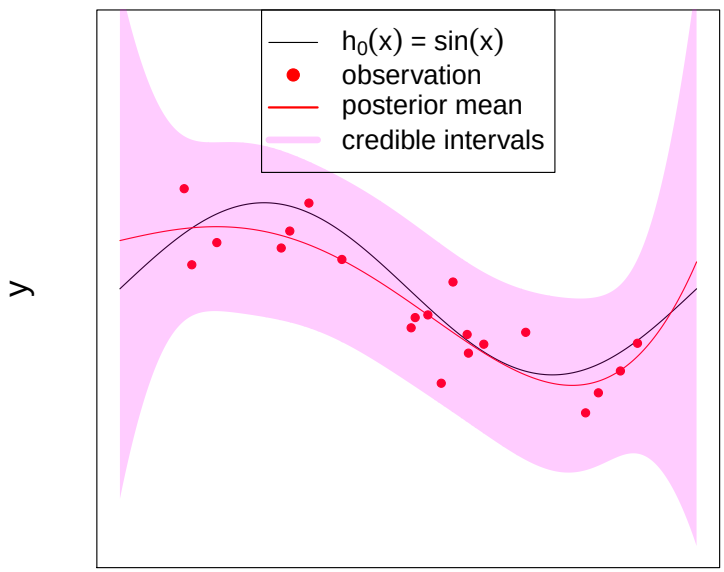


$\alpha = 100$

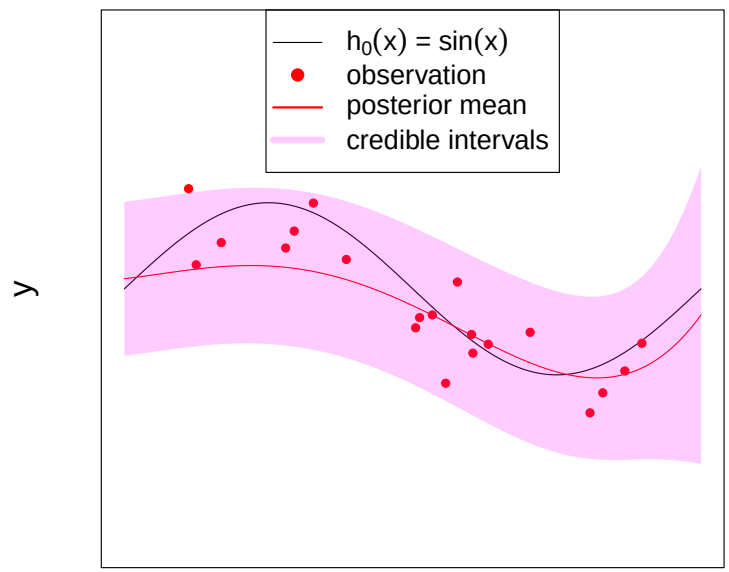
Predictive distribution

- $\widehat{h}(x)$: 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 $\{M_k, k \in \{1, \ldots, K\}\}$, with $M_k = \{\Theta_k, \pi_k\}$?
- Continuous case : family of models $\{M_\alpha, \alpha \in \mathcal{A}\}$
- $\rightarrow$ How to choose k or α ?
- Examples :
  - $M_1 = \{\Theta, \pi_1\}$, $M_2 = \{\Theta, \pi_2\}$ with π_1 a flat prior and π_2 the Jeffreys prior
  - M_α linear model with normal prior on the noise $\mathcal{N}(0, \alpha^{-1})$

1. Supervised learning example : Naive Bayes Classification

2. Bayesian linear regression

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

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 , $k \in \{1, \ldots, M\}$.
- 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 $g(\boldsymbol{\theta})$,

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

- by the tower rule for conditional expectations, the posterior mean is a weighted average

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

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 $k^* = k_0$ $\Longrightarrow$ 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

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

- 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 :

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

- 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 Δ_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

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

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

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