Introduction to Bayesian learning
Lecture 1: Bayesian learning: basics

Anne Sabourin, Ass. Prof., Telecom ParisTech

September 2017

Course mechanism

- 7 sessions of 3 hours each
- 2 lab session (sessions 3 and 7)
- Evaluation : 50% homework (2^{nd} lab report), 50% written exam.
- Course Software : `R`.

Syllabus

1. Bayesian learning : basics
2. Bayesian methods for unsupervised and supervised problems, Bayesian decision theory
3. Lab session I : `R` tutorial, Naive Bayes, Bayesian regression
4. Variational methods I
5. Variational methods II and Sampling methods I (Monte-Carlo)
6. Sampling methods II : Monte-Carlo Markov Chain (and sequential methods if some time left)
7. Lab session II : variational and sampling methods

Lecture 1, Basics of Bayesian learning : Outline

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   Point estimation
   Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

1. When is a Bayesian approach needed ?

2. The Bayesian framework

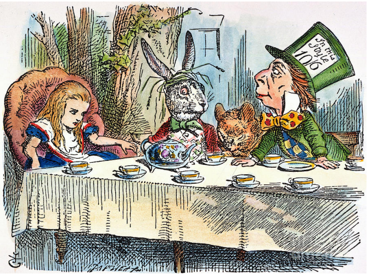
3. Construction of estimators
    - Point estimation
    - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

The English lady, the music lover and the drunkard

The English Lady claims that she can tell whether the milk was poured before or after the tea, after one sip.



Ten trials are made. At each trial the milk is randomly poured before or after the tea. The lady's gess is true 9 times over 10.

What is your verdict : can she really tell ?

The English lady, the music lover and the drunkard

A music lover claims that he can tell if a piece is from Haydn or Mozart after listening only ten seconds of it.

Ten trials are made. At each trial a music piece is randomly chosen from Haydn or Mozart. The music lover's guess is true 9 times over 10.

What is your verdict : can he really tell ?

The English lady, the music lover and the drunkard

Your drunken friend claims that he can predict the outcome of a flip of a fair coin.

Ten trials are made. At each trial a coin is flipped. The drunkard's guess is true 9 times over 10.

What is your verdict : can he really tell ?

Issues

- The 3 datasets are the same and the task is similar, however would you give the same answer in the 3 situations ?

- What level of confidence would you have concerning your answer ? Are the asymptotic confidence intervals from the Central Limit Theorem reliable ?

Bayesian statistics provide a formalism to

- Include prior beliefs in the analysis of data.
- Quantify the uncertainty by providing ‘credible intervals’ ($\neq$ classical confidence intervals)

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

Probabilistic modeling

- **Dataset** $X = X_{1:n} = (X_1, \dots, X_n)$, $X_i \in \mathcal{X} = \{0, 1\}$
  $X_i = 1$ if right guess, $1 \leq i \leq n$.
  $\mathcal{X}$ is the **sample space**, n is the **sample size**.

- **Statistical model :**
  $X_i \sim \mathrm{P}_\theta = \mathcal{B}er(\theta)$ (Bernoulli distribution) : $\mathrm{P}_\theta\{1\} = \theta$.
  $\theta \in \Theta = [0, 1]$ unknown **parameter.**
  Θ is the **parameter space**.

- *i.i.d.* (**i**ndependent, **i**dentically **d**istributed) data :
  $X = (X_1, \dots, X_n) \sim \mathrm{P}_\theta^{\otimes n}$: product distribution.

- Underlying probability space $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$. $X_i : \Omega \to \mathcal{X}$,
  $\mathrm{P}_\theta = \mathbb{P}_\theta \circ X_i^{-1}$.
  Here $\mathrm{P}_\theta\{1\} = \mathbb{P}_\theta \circ X_i^{-1}(\{1\}) = \mathbb{P}_\theta\{X_i = 1\}$.

Statistical model

Definition : statistical model

A family $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ of probability distributions indexed by $\theta \in \Theta$, over a sample space $\mathcal{X}$. Θ is the parameter space.

- 'parametric model' : when $\Theta \subset \mathbb{R}^d$.
- 'non parametric model' : when Θ is infinite-dimensional (example : mixture model with infinitely many components)

Goal

Learn about θ_0 using a dataset $X = X_{1:n} = (X_1, \dots, X_n)$, assuming that $X_i \sim P_{\theta_0}$, $1 \leq i \leq n$ for some $\theta_0 \in \Theta$.

- $X_i \in \mathbb{R}^p$: unsupervised learning,
  versus
- $X_i = (z_i, Y_i)$: supervised learning (Y_i : label)

What is a Bayesian model ?

- ‘Prior knowledge’ about θ represented by a probability distribution $\boldsymbol{\pi}$: the **prior distribution**.

- One can define a random variable $\boldsymbol{\theta}$, with $\boldsymbol{\theta} \sim \boldsymbol{\pi}$.

- The X_i’s are independent conditionnally to $\boldsymbol{\theta}$.

- We assume that *a single* θ_0 which is a realisation of $\boldsymbol{\theta}$ produces the data, *i.e.*$X_{1:n}$ is distributed according to $\mathrm{P}_{\theta_0}^{\otimes n}$, for some $\theta_0 \in \Theta$.

Definition : Bayesian model

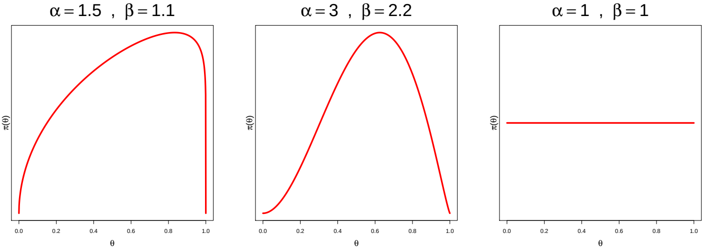
A statistical model $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ together with a prior distribution $\boldsymbol{\pi}$ on Θ.

Example : the English lady

- $\theta \in [0,1]$: probability of a right guess.
- Prior knowledge : The true θ_0 is 'probably' close to 0.5, maybe higher.
- Prior distribution : a Beta distribution $\mathcal{B}eta(\alpha, \beta)$ on $(0,1)$,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- $\mathbb{E}(\boldsymbol{\theta}) = \frac{\alpha}{\alpha+\beta}$, $\mathbb{V}\mathrm{ar}(\boldsymbol{\theta}) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



3 examples of Beta density

Doing Bayesian inference = conditioning upon the data

- The bayesian model results in a **Joint distribution** over the product space $\Theta \times \mathcal{X}$:

$$Q(A \times B) = \int_{\theta \in A} P_\theta(B) \mathrm{d}\boldsymbol{\pi}(\theta)\,, \qquad A \subset \Theta,\, B \subset \mathcal{X}.$$

  P_θ is viewed as a **conditional distribution** of X_i given θ.

- Learning = conditioning prior knowledge about $\boldsymbol{\theta}$ upon data X.

Definition : posterior distribution

conditional distribution of $\boldsymbol{\theta}$ given X

- All the inference (estimation, prediction, ...) is derived from the posterior distribution.

i.i.d. samples : notational conventions

When $X = X_{1:n} = (X_1, \ldots, X_n)$, $X_i \overset{\text{i.i.d}}{\sim} P_\theta$, $1 \le i \le n$.

- Then $X : \Omega \to \mathcal{X}^n$ and $X \sim P_\theta^{\otimes n}$ (product measure)
- Joint distribution over $\Theta \times \mathcal{X}^n$,

$$Q(A \times B) = \int_{\theta \in A} P_\theta^{\otimes n}(B) \mathrm{d}\boldsymbol{\pi}(\theta), \quad B \subset \mathcal{X}^n$$

- If P_θ has a density $p_\theta(x)$, $x \in \mathcal{X}$, then $P_\theta^{\otimes n}$ has density $p_\theta^{\otimes n}(x) = \prod_{i=1}^n p_\theta(x_i)$, $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$.
- For convenience we omit the ' $^{\otimes n}$' sign.

Computing the posterior distribution : Assumptions

- $\boldsymbol{\pi}$ has density π *w.r.t.* reference measure μ (σ-finite), $\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\mu} = \pi$.
- Dominated model : $\exists$ reference measure λ on $\mathcal{X}$ such that each P_θ has density p_θ *w.r.t.* λ : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda} = p_\theta$.
- For a given x, $\theta \mapsto p_\theta(x)$ is the **likelihood function**
- Notation : $p(x|\theta) := p_\theta(x)$.
- x : realisation of X : a single r.v. or an i.i.d. sample $(X_1, \dots, X_n)$

Computing the posterior distribution : Bayes theorem

Under the previous assumptions :

Bayes Theorem

The posterior distribution has a density *w.r.t.* μ given by

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|t)\pi(t)\mathrm{d}\mu(t)}$$

$$\text{posterior} \ \propto \ \text{likelihood} \ \times \ \text{prior}$$

For any $x \in \mathcal{X}$ such that the denominator is > 0.

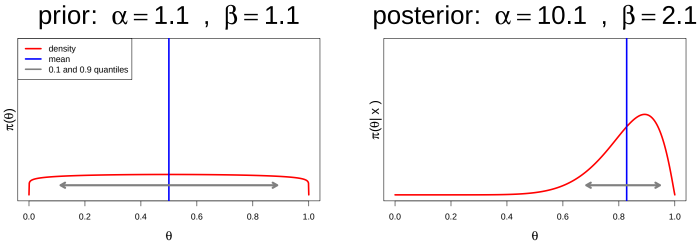
- Denominator : $m(x) = \int_\Theta p(x|t)\pi(t)\mathrm{d}\mu(t)$, marginal density of X
- remind $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ when $P(B) \neq 0$.

Example : The English lady

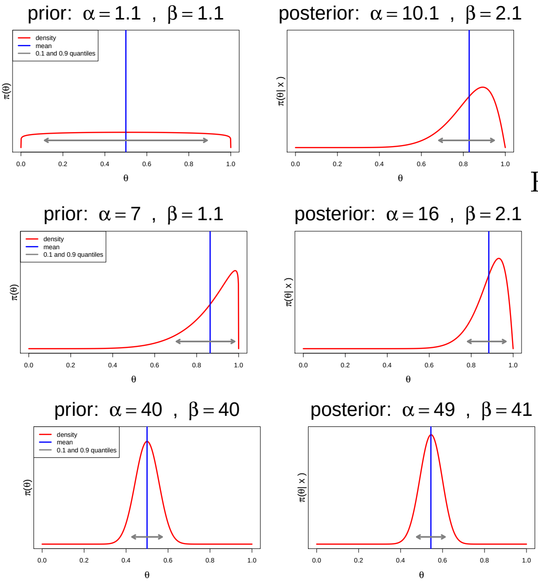
- Assumptions are met with
  - λ : counting measure on $\mathcal{X} = \{0, 1\}$, $\lambda\{0\} = \lambda\{1\} = 1$.
  - $p_\theta(x) = \theta^x(1-\theta)^{1-x}$
  - μ : Lebesgue measure on $(0, 1)$
  - π : Beta density $\mathcal{B}eta(\alpha, \beta)$
- Computing the posterior density :

$$\begin{aligned}\pi(\theta|x) &= \frac{p_\theta(x)\pi(\theta)}{\underbrace{m(x)}_{\text{does not depend on } \theta}} \\ &\propto p_\theta(x)\pi(\theta) \quad (\propto\text{: proportional to}) \\ &= \theta^{\alpha+\sum_{i=1}^n x_i - 1}(1-\theta)^{\beta+n-\sum_{i=1}^n x_i - 1} \\ &\propto \text{density of } \mathcal{B}eta(\alpha + \sum x_i, \beta + n - \sum x_i)\end{aligned}$$

Posterior density for the English lady



Influence of the prior



English lady

music lover

drunkard

Sequential nature of Bayesian learning

Posterior after n i.i.d. obs $x_{1:n}$ starting from prior π

=

Posterior after the latest obs x_n starting from prior $\pi(\theta|x_{1:n-1})$

Proof

$$\pi(\theta|x_{1:n}) = \frac{\pi(\theta)p(x_{1:n-1}|\theta)p(x_n|\theta)}{\int \pi(t)p(x_{1:n-1}|t)\mathrm{d}\mu(t)} \times \frac{\int \pi(t)p(x_{1:n-1}|t)\mathrm{d}\mu(t)}{\int \pi(t)p(x_{1:n}|t)\mathrm{d}\mu(t)}$$
$$= \frac{\pi(\theta|x_{1:n-1})p(x_n|\theta)}{\tilde{m}(x_{1:n})}$$

with

$$\tilde{m}(x_{1:n}) = \frac{\int \pi(t)p(x_{1:n-1}|t)p(x_n|t)\mathrm{d}\mu(t)}{\underbrace{\int \pi(t|x_{1:n-1})\mathrm{d}\mu(t)}_{=1} m(x_{1:n-1})}$$
$$= \int \pi(t|x_{1:n-1})p(x_n|t)\mathrm{d}t$$

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

   Point estimation

   Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

From posterior probability to estimation

- Raw output of Bayesian analysis : a posterior distribution (represented as a density or as a sample $(\theta_1, \ldots, \theta_n) \sim \pi(\theta|X_{1:n})$
- In practice : one wants to answer questions of the kind
  - Does $\theta \in \Theta_0 \subset \Theta$?
  - What is your best guess $\widehat{\theta}(X)$ for θ, given data X ? (point estimation)
  - Can you give a region $R \subset \Theta)$ such that $\mathbb{P}(\theta \in R|X_{1:n}) \geq 1 - \alpha$?

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

   Point estimation

   Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

Bayesian point estimation

most popular estimators of θ : posterior mode and posterior mean.

- Posterior mode $\widetilde{\theta} = \mathrm{argmax}_t\, \pi(t|X_{1:n})$
- Posterior mean $\theta^* = \mathbb{E}_\pi(\theta|X_{1:n}) = \int_\Theta \theta\, \pi(\mathrm{d}\theta|X_{1:n})$.
- generalisation of posterior mean for a quantity of interest $g(\theta)$:

$$g^* = \mathbb{E}_\pi(g(\theta)|X_{1:n}) = \int_\Theta g(\theta)\, \pi(\mathrm{d}\theta|X_{1:n}).$$

- Warning : posterior mode depends on the reference measure

Remark : all three estimators are 'statistics' : functions of $X_{1:n}$.

Discussion : posterior mode

- Intuition : $\widetilde{\theta}$ is the 'center' of the region $\delta\theta$ of measure $\mu(\delta\theta)$ for which the posterior mass $\approx \pi(\widetilde{\theta})\mu(\delta\theta)$ is the highest.

- Warning (main criticism) : $\widetilde{\theta}$ depends on the reference measure.

Discussion : posterior mean

Main justification (for $g : \Theta \to \mathbb{R}$)

$$g^* := \mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n}) = \underset{\gamma\in\mathbb{R}}{\operatorname{argmin}} \underbrace{\mathbb{E}_\pi\left[\left(g(\theta)-\gamma\right)^2|X_{1:n}\right]}_{\varphi(\gamma)}.$$

g^* minimizes he posterior expectancy of the quadratic risk. Indeed,

$$\begin{aligned}\varphi(\gamma) &= \int_\Theta \left(g(\theta)^2 - 2\gamma g(\theta) + \gamma^2\right)\pi(\mathrm{d}\theta|X_{1:n}) \\ &= Cste\ - 2\gamma \underbrace{\int g(\theta)\pi(\mathrm{d}\theta|X_{1:n})}_{\mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n})} + \gamma^2\end{aligned}$$

$$\varphi'(\gamma) = 2(-\mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n}) + \gamma)$$

$$\varphi'(\gamma) = 0 \iff \gamma = \mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n})$$

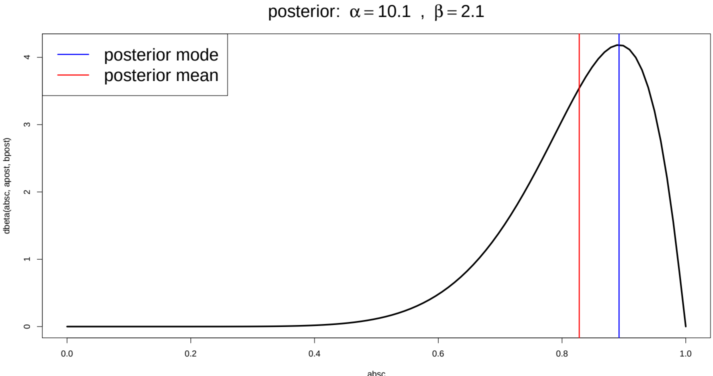
(then check that this solution is indeed a minimum : $\varphi''(g^*) > 0$)

Example : Tea Lady

- prior over $]0,1[$: $\boldsymbol{\pi} = \mathcal{B}eta(\alpha = 1.1, \beta = 1.1)$.
- posterior distribution : $\mathcal{B}eta(\alpha' = 10.1, \beta' = 2.1)$.

$$\text{mode} : \tilde{\theta} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{9.1}{10.2}$$

$$\text{mean} : \theta^* = \frac{\alpha'}{\alpha' + \beta'} = \frac{10.1}{12.2}$$



1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

   Point estimation

   Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

Interval / Set estimation

In what reasonable interval/region of Theta do you believe θ to belong ?

Goal : find a region $R \subset \Theta$ with

- High posterior mass
- Moderate 'size' (*w.r.t.* the reference measure)

defintion : poterior credible set

Given the data x, A posterior credible set of level α for a quantity of interes $g(\theta)$ is any (measurable) region $R \subset g(\Theta)$ such that

$$\mathbb{P}_\pi(g(\theta) \in R|x) \geq \alpha.$$

- ⚠ credible sets $\neq$ confidence regions $R_{classic}$ for an estimator (classical setting), such that $\mathbb{P}_\theta(R_{classic} \ni g(\theta)) \geq \alpha$, $\forall \theta$.
- in fact : confidence and credible sets 'approximately' coïncide for large sample sizes (due to Bernstein-Von-Mises theorem, see last section).

Posterior quantiles

- similarly to confidence interval, there is no unique way to define credible sets.

- easy way (for $g(\Theta) \subset \mathbb{R}$) : use posterior quantiles

- remind : if Q is a probability on $\mathbb{R}$, an α-quantile relative to Q is any $q_\alpha \in \mathbb{R}$ s.t. $Q[-\infty, q_\alpha] = \alpha$.

- When $(1-\alpha)/2$ and $(1+\alpha)/2$ quantiles for $\boldsymbol{\pi}(\cdot|x)$ exist, a credible interval of level alpha is $(q_{\frac{1-\alpha}{2}}, q_{\frac{1+\alpha}{2}}]$.

Minimum volume sets

- It is the solution to the initial requirements (large posterior mass, small reference measure)

- define $R(k) = \{\theta \ : \ \pi(\theta|x) \geq k\}$ ('interior of a density level set)

- The minimum volume set of level α is R_{k_α}, where

$$k_\alpha = \inf\{k \geq 0 : \boldsymbol{\pi}(R(k)|x) \geq \alpha\}$$

- in practice (in general) : hard to compute. Need Monte-Carlo methods, computationally intensive in high dimension.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

Point estimation

Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. Exercises

Simpson Paradox

A university is accused of sexual disrimination beacuse 45% of male applicants are accepted versus only 35% for female applicants.

However, each department (art department and engineering department) accepts more female applicants than male applicants.

How do you explain this ?

hint : use Bayes theorem and the fact that the art department is smaller than the engineering department (fewer applicants) and has a lower overall acceptance rate.

Normalizing constant for the Beta distribution ([Bishop, 2006], ex. 2.5)

prove that

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Bibliography

[Berger, 2013] Berger, J. O. (2013).
Statistical decision theory and Bayesian analysis.
Springer Science & Business Media.

[Bishop, 2006] Bishop, C. M. (2006).
Pattern recognition and machine learning.
springer.

[Robert, 2007] Robert, C. (2007).
The Bayesian choice : from decision-theoretic foundations to computational implementation.
Springer Science & Business Media.