BML lecture #1: Bayesics

http://github.com/rbardenet/bml-course

Rémi Bardenet
remi.bardenet@gmail.com

CNRS & CRIStAL, Univ. Lille, France

Instructors

Besides me, the lecturers are

- Gabriel Victorino Cardoso (Mines Paris, PSL Univ.),
- Julyan Arbel (Inria Grenoble).


Course outline

I. What is Bayesian ML?
  1. Warmup: linear regression
  2. Maximizing expected utility
II. How do perform Bayesian ML?
  1. Monte Carlo methods
  2. Variational Bayes
III. Why do you want to apply BML? } Adv.
  Foundations
IV. Nonparametrics
V. Generative models in BML. } Adv.
VI. Student seminar.

Homework: exercise sheet, practicals

Rémi 4x3

← projects start

Julyan 3hrs

Gabriel 2x3

1x 4hrs

Outline

Quotes from Gelman et al., 2013 on Bayesian methods

- *[...] practical methods for making inferences from data, using probability models for quantities we observe and for quantities about which we wish to learn.*

- *The essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis.*

- *Three steps:*
  1. *Setting up a full probability model,*
  2. *Conditioning on observed data, calculating and interpreting the appropriate "posterior distribution",*
  3. *Evaluating the fit of the model and the implications of the resulting posterior distribution. In response, one can alter or expand the model and repeat the three steps.*

Notation that I will try to stick to

- $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ denote observable data/labels.
- $x_{1:n} \in \mathcal{X}^n$ denote covariates/features/hidden states.
- $z_{1:n} \in \mathcal{Z}^n$ denote hidden variables.
- $\theta \in \Theta$ denote parameters.
- X denotes an $\mathcal{X}$-valued random variable. Lowercase x denotes either a point in $\mathcal{X}$ or an $\mathcal{X}$-valued random variable.

More notation

- Whenever it can easily be made formal, we write densities for our random variables and let the context indicate what is meant. So if $X \sim \mathcal{N}(0, \sigma^2)$, we write

$$\mathbb{E}h(X) = \int h(x) \frac{e^{-x^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \mathrm{d}x = \int h(x) p(x) \mathrm{d}x.$$

  Similarly, for $X \sim \mathcal{P}(\lambda)$, we write

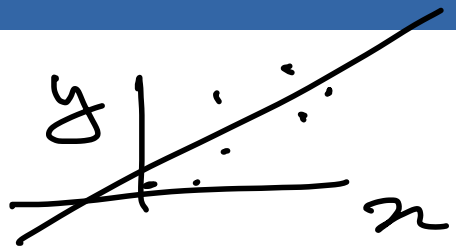
$$\mathbb{E}h(X) = \sum_{k=0}^{\infty} h(k) e^{-\lambda} \frac{\lambda^k}{k!} = \int h(x) p(x) \mathrm{d}x$$

- All pdfs are denoted by p, so that, e. g.

$$\begin{aligned}
\mathbb{E}h(Y, \theta) &= \int h(y, \theta) p(y, \theta) \, \mathrm{d}y \mathrm{d}\theta \\
&= \int h(y, \theta) p(y, x, \theta) \, \mathrm{d}x \mathrm{d}y \mathrm{d}\theta \\
&= \int h(y, \theta) p(y, \theta | x) p(x) \, \mathrm{d}x \mathrm{d}y \mathrm{d}\theta
\end{aligned}$$

Outline

Outline



$y_i = x_i^T \theta + \varepsilon_i$ iid
$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

$y = X\theta + \varepsilon$

Thm: $\hat{\theta}_{MLE} \in \operatorname{argmin} \mathcal{N}(y \mid X\theta, \sigma^2 I)$

$\hat{\theta}_{MLE} \sim \mathcal{N}(\theta, \sigma^2 (X^T X)^{-1})$

· Fisher (1st half of 20th century)

Proof: Assume $X^T X$ is full-rank

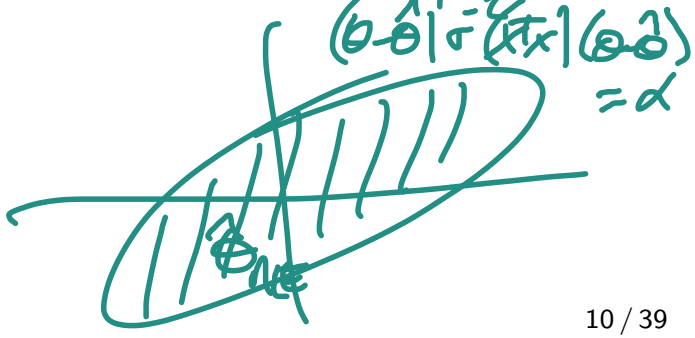
Then $\hat{\theta}_{MLE} = (X^T X)^{-1} X^T y$. So

$\mathbb{E}\hat{\theta}_{MLE} = (X^T X)^{-1} X^T \mathbb{E}[y \mid \theta, X]$

$= (X^T X)^{-1} X^T X \theta = \theta$.

$\mathbb{V}\hat{\theta}_{MLE} = (X^T X)^{-1} X^T \underbrace{\mathbb{V}[y \mid \theta, X]}_{\sigma^2 I} X (X^T X)^{-1}$

Rk: $P_{y|X\theta}(\theta \in A_\alpha(X, y)) \geq 1 - \alpha$



* $L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$
* $R(\theta, \hat{\theta}) = \mathbb{E}_{y|x,\theta} L(\theta, \hat{\theta}(x,y))$.

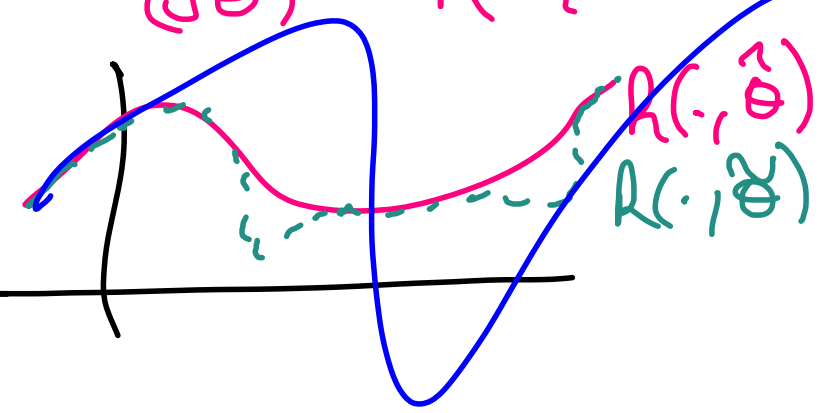
Def: $\hat{\theta}_{\text{minimax}} \in \operatorname{argmin}_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$

$\hat{\theta}_{\text{Bayes}} \in \operatorname{argmin}_{\hat{\theta}} \mathbb{E}_{\theta \sim \pi} R(\theta, \hat{\theta})$

Def: $\hat{\theta}$ is inadmissible if $\exists \tilde{\theta}$

$(\forall \theta)$ $R(\theta, \tilde{\theta}) \leq R(\theta, \hat{\theta})$

$(\exists \theta)$ $R(\theta, \tilde{\theta}) < R(\theta, \hat{\theta})$



Wald ↑1950

Thm (ext. of James-Stein estimator)

$\hat{\theta}_{MLE}$ is inadmissible in linear regression.

Let $\lambda > 0$, $\hat{\theta}_\lambda \in \operatorname{argmin} \|y - X\theta\|^2 + \lambda \|\theta\|^2$

$$= (X^T X + \lambda I)^{-1} X^T y. \quad (*)$$

Proposition: Let $B > 0$, $(\exists \lambda)$ s.t. $(\forall \theta : \|\theta\| \le B)$

$$R(\theta, \hat{\theta}_\lambda) < R(\theta, \hat{\theta}_{MCE}).$$

Proof: * $R(\theta, \hat{\theta}_\lambda) = \mathbb{E}_{y|X,\theta} \|\theta - \hat{\theta}_\lambda\|^2$

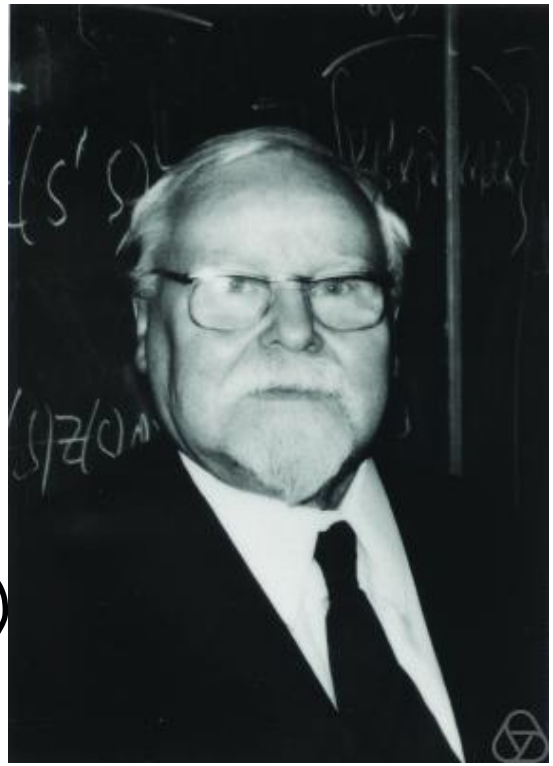
$$= \|\theta - \mathbb{E}_{y|X,\theta} \hat{\theta}_\lambda\|^2 + \operatorname{Tr} \mathbb{V} \hat{\theta}_\lambda \quad (*)$$

* Assume for simplicity $X^T X = I$,

$$\mathbb{E} \hat{\theta}_\lambda = \frac{1}{1+\lambda} X^T \mathbb{E} y = \frac{1}{1+\lambda} \theta.$$

$$\operatorname{Tr} \mathbb{V} \hat{\theta}_\lambda = \frac{1}{(1+\lambda)^2} \operatorname{Tr}\left[X^T \mathbb{V} y \, X\right] = \frac{1}{(1+\lambda)^2} \sigma^2 d$$

So $R(\theta, \hat{\theta}_\lambda) = \|\theta\|^2 \left(1 - \frac{1}{1+\lambda}\right)^2 + \frac{\sigma^2 d}{(1+\lambda)^2} \le B^2 \frac{\lambda^2}{(1+\lambda)^2} + \frac{\sigma^2 d}{(1+\lambda)^2}$



Tikhonov
Hoerl & Kennard

$\hat{\theta}_{Bayes} \in \arg\min_{\hat{\theta}} \mathbb{E}_{\theta \sim p} \mathbb{E}_{y|x,\theta} L(\theta, \hat{\theta})$

Proposition: Let $\theta \sim \mathcal{N}(0, \tau_0^2 I)$ (*)

Let $y|X,\theta \sim \mathcal{N}(X\theta, \sigma^2 I)$

Then $\hat{\theta}_{Bayes} = \hat{\theta}_\lambda$ for $\lambda = \dots$

Proof: $\mathbb{E}_\theta \mathbb{E}_{y|x,\theta} L(\theta, \hat{\theta}) =$

$$\int \|\theta - \hat{\theta}\|^2 \underbrace{p(y|X,\theta)p(\theta)}_{p(y,\theta|X)} dy\, d\theta =$$

$$\int \left[\int \|\theta - \hat{\theta}\|^2 p(\theta|X,y) d\theta \right] p(y|X) dy$$

Thus $\hat{\theta}_{Bayes} := \mathbb{E}(\theta|X,y) \in \arg\min_{\hat{\theta}} \mathbb{E}_{\theta \sim p} R(\theta, \hat{\theta})$.

Now, note that $p(\theta|X,y) = \frac{p(y,\theta|X)}{p(y|X)} = \mathcal{N}\left(\theta \middle| \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T y, \dots\right)$.

Bayes (18th century)

Inference in regression models

Outline

Describing a decision problem under uncertainty

- A state space $\mathcal{S}$,
  Every quantity you need to consider to make your decision.
- Actions $\mathcal{A} \subset \mathcal{F}(\mathcal{S}, \mathcal{Z})$,
  Making a decision means picking one of the available actions.
- A reward space $\mathcal{Z}$,
  Encodes how you feel about having picked a particular action.
- A loss function $L : \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$.
  How much you would suffer from picking action a in state s.

Classification as a decision problem

- $\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \times \mathcal{Y}$, i.e. $s = (x_{1:n}, y_{1:n}, x, y)$.
- $\mathcal{Z} = \{0, 1\}$.
- $\mathcal{A} = \{a_g : s \mapsto 1_{y \neq g(x; x_{1:n}, y_{1:n})}, \quad g \in \mathcal{G}\}$.
- $L(a_g, s) = 1_{y \neq g(x; x_{1:n}, y_{1:n})}$.

PAC bounds; see e.g. (Shalev-Shwartz and Ben-David, 2014)

Let $(x_{1:n}, y_{1:n}) \sim \mathbb{P}^{\otimes n}$, and independently $(x, y) \sim \mathbb{P}$, we want an algorithm $g(\cdot; x_{1:n}, y_{1:n}) \in \mathcal{G}$ such that if $n \geqslant n(\delta, \varepsilon)$,

$$\mathbb{P}^{\otimes n}\left[\mathbb{E}_{(x,y)\sim\mathbb{P}} L(a_g, s) \leqslant \varepsilon\right] \geqslant 1 - \delta.$$

In particular, $\mathbb{E}_{x_{1:n}, y_{1:n}, x, y} L(a_g, s) \leq (1-\delta)\varepsilon + \delta$

Regression as a decision problem

- $\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \times \mathcal{Y}$
- $\mathcal{Z} = \mathbb{R}$
- $\mathcal{A} = \{ a_g : s \mapsto y - g(x, x_{1:n}, y_{1:n}) \,,\ g \in \mathcal{G} \}$
- $L : a_g, s \mapsto \| y - g(x, x_{1:n}, y_{1:n}) \|^2$ $\quad (= -u(a_g(s)))$.

"Statistics" Reason: ~~for~~ $n \geq n(\varepsilon, \delta)$,

$$P^{\otimes n}\left[\mathbb{E}_P L(a_g, S) \leq \varepsilon \right] \geq 1 - \delta$$

Estimation as a decision problem (credible intervals).

- $\mathcal{S} = \mathcal{Y}^n \times \Theta$, $\Theta = \mathbb{R}$ for simplicity.
- $\mathcal{Z} = \{0,1\} \times \mathbb{R}_+$
- $\mathcal{A} = \{a_I : s \mapsto (\mathbb{1}_{\theta \notin I}, |I|)\}$
- $L : a, s \mapsto \mathbb{1}_{\theta \notin I} + \gamma |I|$ for some $\gamma > 0$.

$$\mathbb{E}_{y_{1:n}|\theta} L(a_I, s) = \mathbb{E}_{y_{1:n}|\theta} \mathbb{1}_{\theta \notin I(y_{1:n})} + \gamma \, \mathbb{E}_{y_{1:n}|\theta} |I(y_{1:n})|$$

Clustering as a decision problem

Exo

- $\mathcal{S} =$
- $\mathcal{Z} =$
- $\mathcal{A} =$
-

Density estimation as a decision problem

- $\mathcal{S} =$ $\mathcal{Y}^n \times \mathcal{Y}$
- $\mathcal{Z} =$
- $\mathcal{A} =$ $\{a_{\pi,\ell(\cdot|\cdot)}(y_{1:n}, y) = \int \pi(z)\ell(y|z)dz \,,\ \pi, \ell \in \mathcal{P}\mathcal{G}\}$
- $L : a, s \mapsto -\log a_{\pi,\ell}(y_{1:n}, y)$

$$\mathbb{E}_{Y_{1:n},Y}\left[-\log \int \pi(z|Y_{1:n})\ell(Y|z,Y_{1:n})dz\right]$$

is the risk.

Outline

SEU is what defines the Bayesian approach

The subjective expected utility principle

1. Choose $\mathcal{S}, \mathcal{Z}, \mathcal{A}$ and a loss function $L(a, s)$,
2. Choose a distribution p over $\mathcal{S}$,
3. Take the the corresponding Bayes action

$$a^\star \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s \sim p} L(a, s). \tag{1}$$

Corollary: minimize the posterior expected loss

Now partition $s = (s_{\text{obs}}, s_{\text{u}})$, then

$$a^\star \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s_{\text{obs}}} \mathbb{E}_{s_{\text{u}} | s_{\text{obs}}} L(a, s).$$

In ML, $\mathcal{A} = \{a_g\}$, with $g = g(s_{\text{obs}})$, so that (1) is equivalent to $a^\star = a_{g^\star}$, with

$$g^\star(s_{\text{obs}}) \triangleq \arg\min_{g} \mathbb{E}_{s_{\text{u}} | s_{\text{obs}}} L(a, s).$$

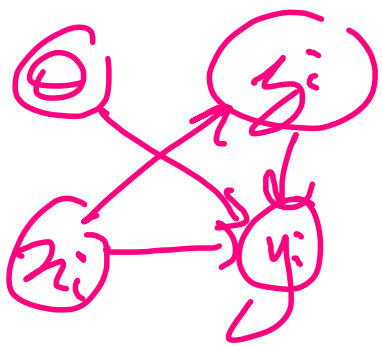
Outline

Pause: A recap on probabilistic graphical models

- Often we specify joint distributions through their conditionals.
- PGMs (aka "Bayesian" networks) represent the dependencies in a joint distribution $p(s)$ by a directed acyclic graph $G = (E, V)$,

$$p(s) = \prod_{v \in V} p(s_v | s_{\mathrm{pa}(v)}).$$

$$p(\theta, y_{1:N}) = \left(\prod_{i=1}^{N} p(y_i|\theta)\right) p(\theta)$$



$$p(\theta, z_{1:N}, y_{1:N}, x_{1:N}) = p(\theta) \prod_{i} p(x_i)\, p(z_i|x_i)\, p(y_i|z_i, x_i, \theta)$$

Pause: A recap on probabilistic graphical models

Check (Murphy, 2012, Section 10.5) for a refresher.

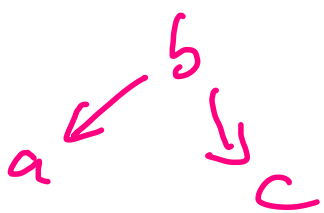
Theorem

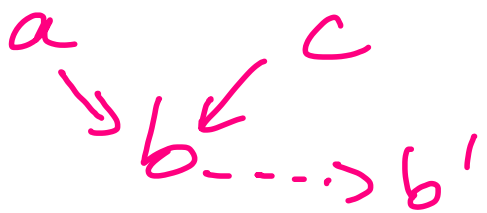
Given a PGM $G = (V, E)$, and $A, B, C \subset V$, then $A \perp B | C$ iff every path from a node in A to a node in B is d-blocked by C.

d-blocking

An undirected path P in G is d-blocked by $E \subset V$ if at least one of the following conditions hold.

- P contains a "chain" $a \to b \to c$ and $b \in E$.
- P contains a "tent" $a \leftarrow b \to c$ and $b \in E$.
- P contains a "v-structure" $a \to b \leftarrow c$ and neither b nor any of its descendants are in E.





Pause: A recap on probabilistic graphical models

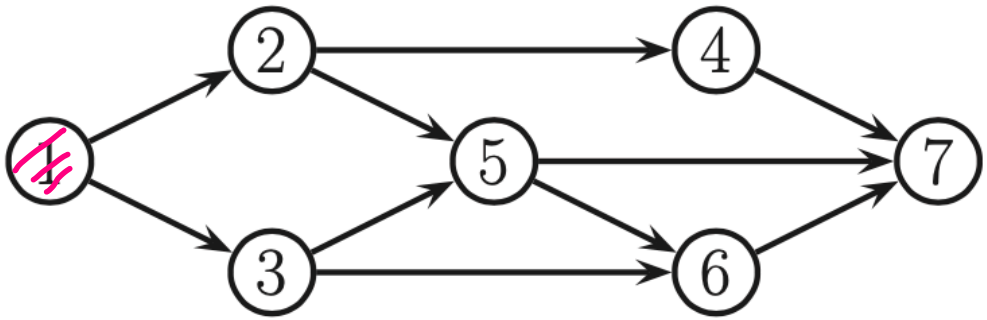


Figure 10.11 A DGM.

- Does $x_2 \perp x_6 | x_5, x_1$?
- Does $x_2 \perp x_6 | x_1$?
- Write the joint distribution as factorized over the graph.

$$p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_2,x_3)$$
$$p(x_6|x_5,x_3)p(x_7|x_4,x_6,x_5).$$

Estimation as a decision problem: point estimates

Estimation as a decision problem: credible intervals

Choosing priors (see Exercises)

Classification as a decision problem

Regression as a decision problem 1/2

Regression as a decision problem 2/2

Dimensionality reduction as a decision problem

Clustering as a decision problem

Topic modelling as a decision problem

Outline

An issue (or is it?)

Depending on how they interpret and how they implement SEU, you will meet many types of Bayesians (46656, according to Good).

A few divisive questions

- Using data or the likelihood to choose your prior; see Lecture 4.
- Using MAP estimators for their computational tractability, like in inverse problems

$$\hat{x}_\lambda \in \arg\min \|y - Ax\| + \lambda\Omega(x).$$

- When and how should you revise your model (likelihood or prior)?
- MCMC vs variational Bayes (more in Lectures #2 and #3)

References I

[1] A. Gelman et al. *Bayesian data analysis*. 3rd. CRC press, 2013.

[2] K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[3] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.