BML lecture #1: Bayesics

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

Rémi Bardenet
remi.bardenet@gmail.com

CNRS & CRIStAL, Univ. Lille, France

Centre de Recherche en Informatique, Signal et Automatique de Lille

What comes to *your* mind when you hear "Bayesian ML"?

Course outline

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.

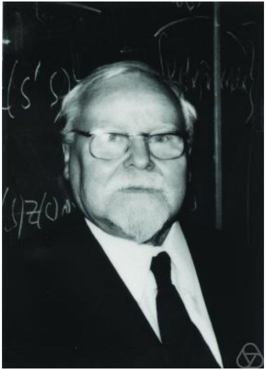
$$\begin{aligned}
\mathbb{E}h(Y, \theta) &= \int h(y, \theta) p(y, \theta) \, \mathrm{d}y\mathrm{d}\theta \\
&= \int h(y, \theta) p(y, x, \theta) \, \mathrm{d}x\mathrm{d}y\mathrm{d}\theta \\
&= \int h(y, \theta) p(y, \theta | x) p(x) \, \mathrm{d}x\mathrm{d}y\mathrm{d}\theta
\end{aligned}$$

Outline

Outline

Inference in regression models

Inference in regression models



Inference in regression models

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 $n \geqslant n(\delta, \varepsilon)$,

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

Regression as a decision problem

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

Estimation as a decision problem

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

Clustering as a decision problem

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

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

- PGMs (aka "Bayesian" networks) represent the dependencies in a joint distribution $p(s)$ by a directed graph $G = (E, V)$.
- Two important properties:

$$p(s) = \prod_{v \in V} p(s_v | s_{\mathrm{pa(v)}}) \qquad \text{and} \qquad y_v \perp y_{nd(v)} | y_{pa(v)}.$$

Also good to know how to determine whether $A \perp B | C$; see (Murphy, 2012, Section 10.5).

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

Theorem

Exercise

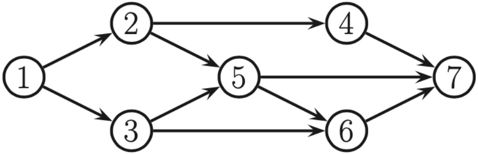


Figure 10.11 A DGM.

- 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.

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 #5.
- 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)

[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.