# BML lecture #1: Bayesics

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

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France

Centre de Recherche en Informatique, Signal et Automatique de Lille

# Outline

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

## A quick motivating example before we go formal 1/2

- Let $N$ individuals evolve from Susceptible to Infected to Recovered, $x_n(t) \in \{S, I, R\}$, $1 \leqslant n \leqslant N$, $t \in [0, T]$.
- Each susceptible individual $n$ moves to $R$ according to a Poisson process with intensity

$$\sum_{k:x_k(t)=I} \lambda_{nk}(\theta_{SI}).$$

- Each infected person recovers after a Gamma$(a, b)$ time.
- This allows to express

$$p(x_1(t_{1,1}), \dots, x_1(t_{1,T_1}), \cdots, x_n(t_{n,1}), \dots, x_1(t_{n,T_n})|\theta).$$

where $\theta = (\theta_{SI}, a, b)$.

- Now, consider $p(\theta|\text{data}) \propto p(\text{data}|\theta) p(\theta)$.

## A quick motivating example before we go formal 2/2

- If asked to report an interval on a particular function of $\theta$, say $R_0$, I would output a small interval $I$ such that

$$\int_I p(\theta|\text{data})\,\mathrm{d}\theta \geqslant 0.95.$$

- If asked whether we should close universities, I would ask for
  - the cost $\alpha$ of closing unis when $R_0 < 1$,
  - the cost $\beta$ of keeping unis open while $R_0 > 1$.
- Then I would recommend closing if and only if

$$p(R_0 > 1|\text{data}) > \frac{\alpha}{\alpha + \beta}.$$

- Additionally, I would check that the decision doesn't change if I change my prior $p(\theta)$ a little.
- If it did, then I would refine my likelihood and/or wait for more data.

# 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.

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

# Abraham Wald (1902–1950)

# 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 $s$. It is also customary to first define a utility $u : \mathcal{Z} \to \mathbb{R}_+$, and then let

$$L(a, s) = \sup_{a' \in \mathcal{A}} u\big(a'(s)\big) - u\big(a(s)\big) \in \mathbb{R}_+.$$

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

**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

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

## 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

If we partition $s = (s_o, s_\mathrm{u})$, then

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

Equivalently to (1), given $s_o$, we choose

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

# A recap on probabilistic graphical models

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

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

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

# 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 2/2

## Dimensionality reduction as a decision problem

# Clustering as a decision problem

# Image denoising as a decision problem

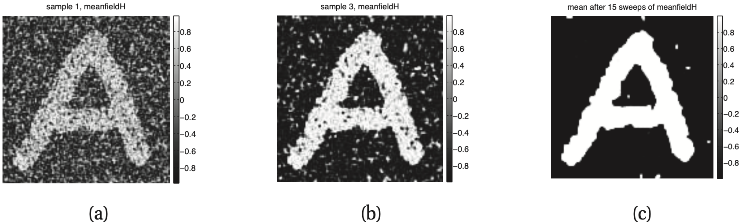


**Figure:** Taken from (Murphy, 2012, Chapter 21)

# 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)

# 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.