Lecture 2c. Statistical Schools of Thought: The Bayesian Paradigm

COMP90051 Statistical Machine Learning

Semester 2, 2020
Lecturer: Ben Rubinstein

THE UNIVERSITY OF MELBOURNE

This lecture

How do learning algorithms come about?

- Frequentist statistics
- Statistical decision theory
- Extremum estimators
- **Bayesian statistics**

Types of probabilistic models

- **Parametric vs. Non-parametric**
- **Generative vs. Discriminative**

Bayesian Statistics

Wherein unknown model parameters have associated distributions reflecting prior belief.

Bayesian statistics

Laplace

- Probabilities correspond to beliefs
- Parameters
  - Modeled as r.v.'s having distributions
  - Prior belief in θ encoded by prior distribution $P(\theta)$
    - Parameters are modeled like r.v.'s (even if not really random)
    - Thus: data likelihood $P_\theta(X)$ written as conditional $P(\mathrm{X}|\theta)$
  - Rather than point estimate $\hat{\theta}$, Bayesians update belief $P(\theta)$ with observed data to $P(\theta|X)$ the posterior distribution)

Tools of probabilistic inference

Bayes

- Bayesian probabilistic inference
  - Start with prior $\mathrm{P}(\theta)$ and likelihood $\mathrm{P}(X|\theta)$
  - Observe data $X = x$
  - Update prior to posterior $\mathrm{P}(\theta|X = x)$
- Primary tools to obtain the posterior
  - Bayes Rule: reverses order of conditioning

$$P(\theta|X = x) = \frac{P(X = x|\theta)P(\theta)}{P(X = x)}$$

  - Marginalisation: eliminates unwanted variables

$$\mathrm{P}(X = x) = \sum_t P(X = x, \theta = t)$$

These are general tools of probability and not specific to Bayesian stats/ML

This quantity is called the evidence

Example

- We model $X|\theta$ as $N(\theta, 1)$ with prior $N(0,1)$
- Suppose we observe X=1, then update posterior

$$P(\theta|X=1) = \frac{P(X=1|\theta)P(\theta)}{P(X=1)}$$

$$\propto P(X=1|\theta)P(\theta)$$

$$= \left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{(1-\theta)^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{\theta^2}{2}\right)\right]$$

$$\propto N(0.5,0.5)$$

NB: allowed to push constants out front and "ignore" as these get taken care of by normalisation

$$P(\theta|X=1) = \frac{P(X=1|\theta)P(\theta)}{P(X=1)}$$

$$\propto P(X=1|\theta)P(\theta)$$

Name of the game is to get posterior into a recognisable form. exp of quadratic *must* be a Normal

Discard constants w.r.t θ

$$= \left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{(1-\theta)^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{\theta^2}{2}\right)\right]$$

Collect exp's

$$\propto exp\left(-\frac{(1-\theta)^2+\theta^2}{2}\right)$$

$$= exp\left(-\frac{2\theta^2-2\theta+1}{2}\right)$$

Want leading numerator term to be θ^2 by moving coefficient to denominator

$$= exp\left(-\frac{\theta^2-\theta+\frac{1}{2}}{2\cdot\frac{1}{2}}\right)$$

Complete the square in numerator: move out excess constants

$$= exp\left(-\frac{\theta^2-\theta+\frac{1}{4}}{2\cdot\frac{1}{2}}\right)\cdot \underline{exp\left(-\frac{\frac{1}{4}}{2\cdot\frac{1}{2}}\right)}$$

$$\propto exp\left(-\frac{\theta^2-\theta+\frac{1}{4}}{2\cdot\frac{1}{2}}\right)$$

Factorise

$$= exp\left(-\frac{(\theta-\frac{1}{2})^2}{2\cdot\frac{1}{2}}\right)$$

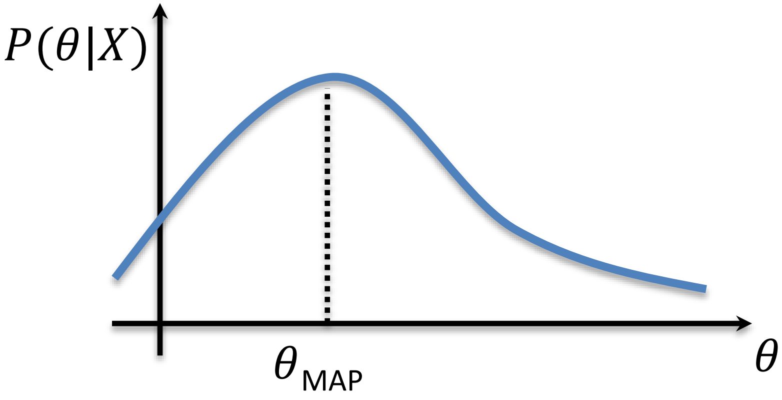
Recognise as (unnormalized) Normal!

$$\propto \mathcal{N}(0.5, 0.5)$$

Constant underlined
Variance/std deviation circled

How Bayesians make point estimates

- They don't, unless forced at gunpoint!
  - The posterior carries full information, why discard it?
- But, there are common approaches
  - Posterior mean $E_{\theta|X}[\theta] = \int \theta P(\theta|X) d\theta$
  - Posterior mode $\underset{\theta}{\operatorname{argmax}} P(\theta|X)$ (max a posteriori or MAP)
  - There're Bayesian decision-theoretic interpretations of these



MLE in Bayesian context

- MLE formulation: find parameters that best fit data
$$\hat{\theta} \in \operatorname{argmax}_\theta P(X = x|\theta)$$

- Consider the MAP under a Bayesian formulation
$$\begin{aligned}\hat{\theta} &\in \operatorname{argmax}_\theta P(\theta|X = x) \\ &= \operatorname{argmax}_\theta \frac{P(X = x|\theta)P(\theta)}{P(X = x)} \\ &= \operatorname{argmax}_\theta P(X = x|\theta)P(\theta)\end{aligned}$$

- **Prior** $P(\theta)$ weights; MLE like *uniform* $P(\theta) \propto 1$

- Extremum estimator: Max $\log P(X = x|\theta) + \log P(\theta)$

Frequentists vs Bayesians – Oh My!

- Two key schools of statistical thinking
  - Decision theory complements both
- Past: controversy; animosity; almost a 'religious' choice
- Nowadays: deeply connected

I declare the Bayesian vs. Frequentist debate over for data scientists

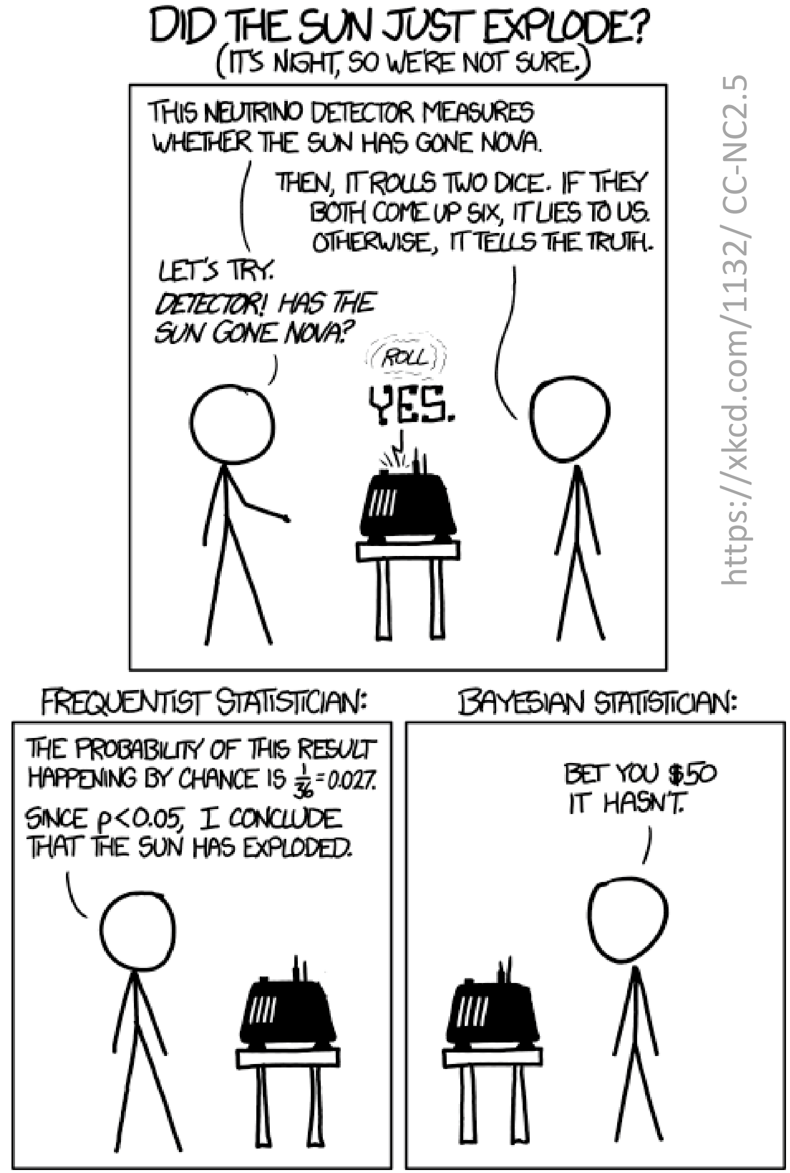
Rafael Irizarry 2014/10/13

Are You a Bayesian or a Frequentist?

Michael I. Jordan

Department of EECS
Department of Statistics
University of California, Berkeley

http://www.cs.berkeley.edu/~jordan



(Some) Categories of Probabilistic Models

Parametric vs non-parametric models

Parametric	Non-Parametric
Determined by fixed, finite number of parameters	Number of parameters grows with data, potentially infinite
Limited flexibility	More flexible
Efficient statistically and computationally	Less efficient

Examples to come! There are non/parametric models in both the frequentist and Bayesian schools.

Generative vs. discriminative models

- X's are instances, Y's are labels (supervised setting!)
  - Given: i.i.d. data $(X_1, Y_1),\ldots,(X_n, Y_n)$
  - Find model that can predict Y of new X
- Generative approach
  - Model full joint $P(X, Y)$
- Discriminative approach
  - Model conditional $P(Y|X)$ only
- Both have pro's and con's

Examples to come! There are generative/discriminative models in both the frequentist and Bayesian schools.

Summary

- Bayesian paradigm: Its all in the prior!
- Bayesian point estimate: MAP (an extremum estimator)
- Parametric vs Non-parametric models
- Discriminative vs. Generative models

Next: Logistic regression (unlike you've ever seen before)

Workshops week #2: learning Bayes one coin flip at a time!