

Lecture 2. Statistical Schools of Thought

COMP90051 Statistical Machine Learning

Semester 2, 2019
Lecturer: Ben Rubinstein



THE UNIVERSITY OF
MELBOURNE

Copyright: University of Melbourne

This lecture

- How do learning algorithms come about?
 - * Frequentist statistics
 - * Statistical decision theory
 - * Bayesian statistics
- Types of probabilistic models
 - * Parametric vs. Non-parametric
 - * Generative vs. Discriminative

Extending Berkeley CS 294-34 tutorial slides by Ariel Kleiner

Frequentist Statistics

Wherein unknown model parameters are treated as having fixed but unknown values.

Frequentist statistics

- Abstract problem

- * Given: X_1, X_2, \dots, X_n drawn i.i.d. from some distribution
- * Want to: identify unknown distribution, or a property of it

Independent and
identically distributed

- Parametric approach (“**parameter estimation**”)

- * Class of **models** $\{p_\theta(x): \theta \in \Theta\}$ indexed by **parameters** Θ (could be a real number, or vector, or)
- * **Point estimate** $\hat{\theta}(x_1, \dots, x_n)$ a function (or **statistic**) of data

Hat means estimate
or estimator

- Examples

- * Given n coin flips, determine probability of landing heads
- * Choosing a classifier

Bias, variance and asymptotic versions

Frequentists seek good behaviour in ideal conditions

- **Bias:** $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$
- **Variance:** $Var_{\theta}(\hat{\theta}) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2]$

Subscript θ
means data really
comes from p_{θ}

$\hat{\theta}$ still function of data

Asymptotic properties

- **Consistency:** $\hat{\theta}(X_1, \dots, X_n)$ converges to θ as $n \rightarrow \infty$
- **Efficiency:** asymptotic variance is as small as possible

Maximum-Likelihood Estimation

- A **general principle** for designing estimators
- Involves **optimisation**
- $\hat{\theta}(x_1, \dots, x_n) \in \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(x_i)$
- MLE estimators are consistent (under technical conditions)



Fischer

$$\text{MLE: } \hat{\theta}(x_1, \dots, x_n) \in \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(x_i)$$

Question: Why a product?

The more likelihoods,
the higher the **A**
product.

So that we can weigh
each data point **B**
separately.

The data points are
assumed **C**
independent.

Example I: Bernoulli

- Know data comes from Bernoulli distribution with unknown parameter (e.g., biased coin); find mean
- MLE for mean

- * $p_{\theta}(x) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0 \end{cases} = \theta^x (1 - \theta)^{1-x}$
(note: $p_{\theta}(x) = 0$ for all other x)

- * Maximising likelihood yields $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$

Example II: Normal

- Know data comes from Normal distribution with variance 1 but unknown mean; find mean

- MLE for mean

- * $p_{\theta}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \theta)^2\right)$

- * Maximising likelihood yields $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$

- Exercise: derive MLE for *variance* σ^2 based on

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \text{ with } \theta = (\mu, \sigma^2)$$

MLE 'algorithm'

1. Given data X_1, \dots, X_n **define** probability distribution, p_θ , assumed to have **generated the data**
2. Express likelihood of data, $\prod_{i=1}^n p_\theta(X_i)$
(usually its ***logarithm***... ***why?***)
3. Optimise to find *best* (most likely) parameters $\hat{\theta}$
 1. take partial derivatives of log likelihood wrt θ
 2. set to 0 and solve
(failing that, use **iterative gradient method**)

Statistical Decision Theory

Branch within statistics, optimisation, economics, control, emphasising utility maximisation.

Decision theory



Wald

- Act to maximise utility - connected to economics and operations research
- **Decision rule** $\delta(\mathbf{x}) \in A$ an action space
 - * E.g. Point estimate $\hat{\theta}(x_1, \dots, x_n)$
 - * E.g. Out-of-sample prediction $\hat{Y}_{n+1} | X_1, Y_1, \dots, X_n, Y_n, X_{n+1}$
- **Loss function** $l(a, \theta)$: economic cost, error metric
 - * E.g. square loss of estimate $(\hat{\theta} - \theta)^2$
 - * E.g. 0-1 loss of classifier predictions $1[y \neq \hat{y}]$

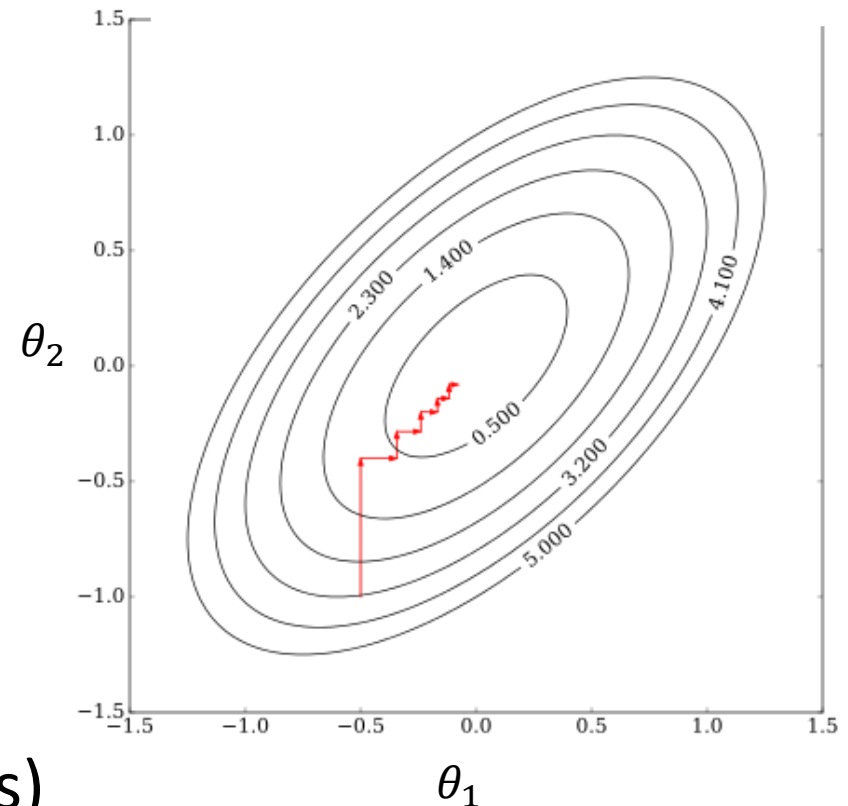
Risk & Empirical Risk Minimisation (ERM)

- In decision theory, really care about *expected* loss
- **Risk** $R_\theta[\delta] = E_{\mathbf{X} \sim \theta}[l(\delta(\mathbf{X}), \theta)]$
 - * E.g. true test error
 - * aka generalization error
- Want: Choose δ to minimise $R_\theta[\delta]$
- Can't directly! Why?
- **ERM**: Use training set \mathbf{X} to approximate p_θ
 - * Minimise **empirical risk** $\hat{R}_\theta[\delta] = \frac{1}{n} \sum_{i=1}^n l(\delta(X_i), \theta)$



Looking ahead to L3

- Optimisation and ML
 - * Max likelihood estimation
 - * Empirical risk minimisation
 - * ... many others
- Cannot do ML without it
- We will cover a little (requires multivariate/vector calculus)



Wikimedia Commons.
Author: Nicoguardo (CC4)

Is this “*Just Theoretical*”™ ?

- Recall Lecture 1 →
- Those evaluation metrics? They’re just estimators of a performance parameter
- Example: error
- Bias, Variance, etc. indicate quality of approximation

COMP90051 Statistical Machine Learning

Evaluation (supervised learners)

- How you measure quality depends on your problem!
- Typical process
 - * Pick an **evaluation metric** comparing label vs prediction
 - * Procure an independent, labelled **test set**
 - * “Average” the evaluation metric over the test set
- Example evaluation metrics
 - * Accuracy, Contingency table, Precision-Recall, ROC curves
- When data poor, **cross-validate**

22

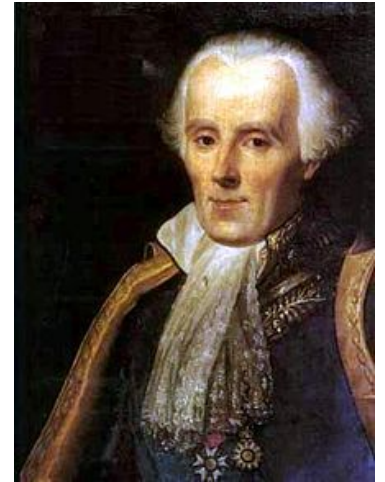
Bias-variance decomposition

- Bias: $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$
- Variance: $Var_{\theta}(\hat{\theta}) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2]$
- Bias-variance decomposition of **square-loss risk**
$$E_{\theta}[(\theta - \hat{\theta})^2] = [B(\hat{\theta})]^2 + Var_{\theta}(\hat{\theta})$$

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 $P(X|\theta)$
 - * 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

- Bayesian probabilistic inference

- * Start with prior $P(\theta)$ and likelihood $P(X|\theta)$
- * Observe data $X = x$
- * Update prior to posterior $P(\theta|X = x)$



Bayes

- 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

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

This quantity
is called the
evidence

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

Example

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

$$\begin{aligned}
 P(\theta|X = 1) &= \frac{P(X = 1|\theta)P(\theta)}{P(X=1)} \\
 &\propto P(X = 1|\theta)P(\theta) \\
 &= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\theta)^2}{2}\right) \right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) \right] \\
 &\propto N(0.5, 0.5)
 \end{aligned}$$

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

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

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

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

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

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

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

Factorise

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

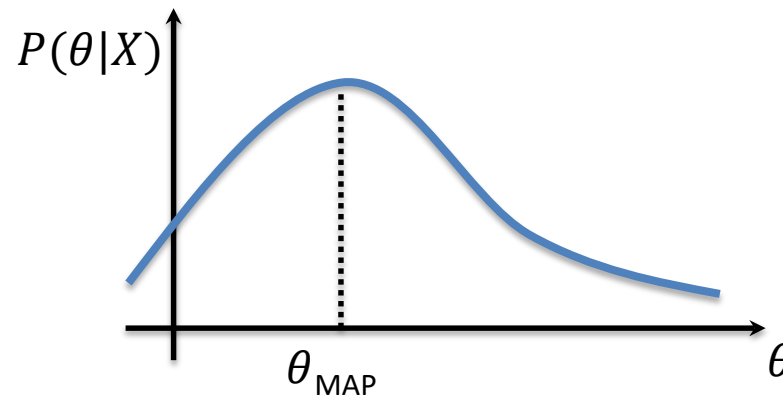
Recognise as (unnormalized) Normal!

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

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 $\operatorname{argmax}_{\theta} P(\theta|X)$ (**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* $P(\theta) \propto 1$

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

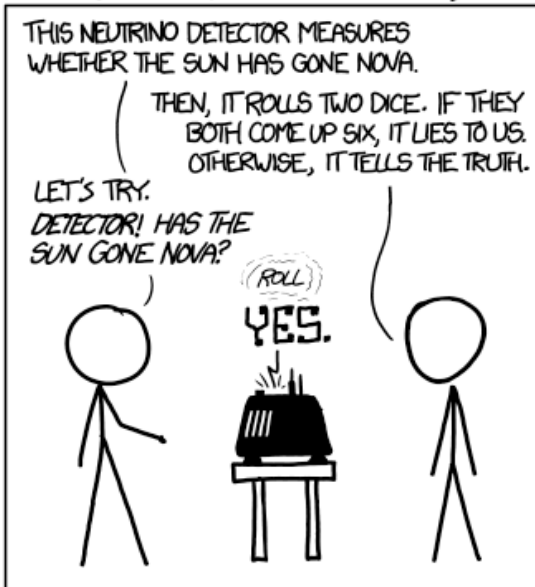
© Rafael Irizarry 2014/10/13

Are You a Bayesian or a Frequentist?

Michael I. Jordan
 Department of EECS
 Department of Statistics
 University of California, Berkeley

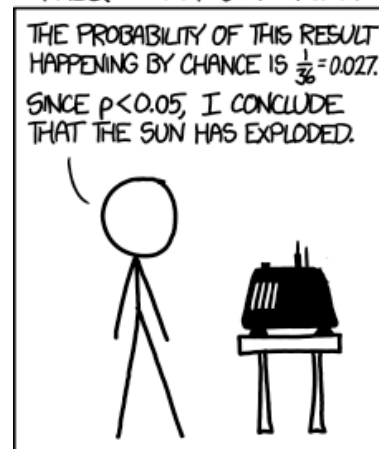
<http://www.cs.berkeley.edu/~jordan>

DID THE SUN JUST EXPLODE?
 (IT'S NIGHT, SO WE'RE NOT SURE.)

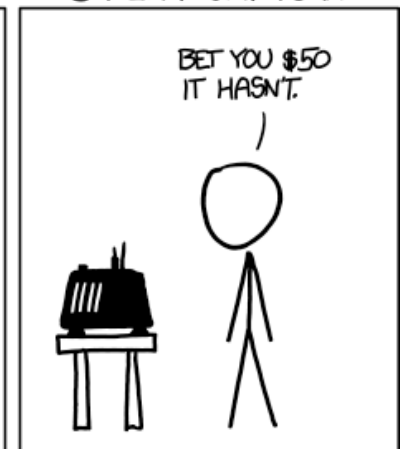


<https://xkcd.com/1132/> CC-NC2.5

FREQUENTIST STATISTICIAN:



BAYESIAN STATISTICIAN:



(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 $(X_1, Y_1), \dots, (X_n, Y_n)$
 - * 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

- Philosophies: frequentist vs Bayesian
- Principles behind many learners:
 - * MLE
 - * Risk minimisation
 - * Probabilistic inference, MAP
- Parametric vs Non-parametric models
- Discriminative vs. Generative models

Next time: Linear regression (demo's ideas) and Optimisation (needed for MLE, ERM, etc.)

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