Lecture 2. Statistical Schools of Thought

COMP90051 Statistical Machine Learning

Semester 1, 2021
Lecturer: Trevor Cohn

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

Frequentist Statistics

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

Frequentist statistics

Independent and identically distributed

- Abstract problem
  - Given: $X_1, X_2, \ldots, X_n$ drawn i.i.d. from some distribution
  - Want to: identify unknown distribution, or a property of it
- 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 $\hat{\theta}\ (X_1, \ldots, X_n)$ a function (or statistic) of data

Hat means estimate or estimator

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

Estimator Bias

Frequentists seek good behaviour, in ideal conditions

- Bias: $\mathrm{B}_\theta(\hat{\theta}) = \mathrm{E}_\theta[\hat{\theta}(X_1, \dots, X_n)] - \theta$

Subscript θ means data *really* comes from p_θ

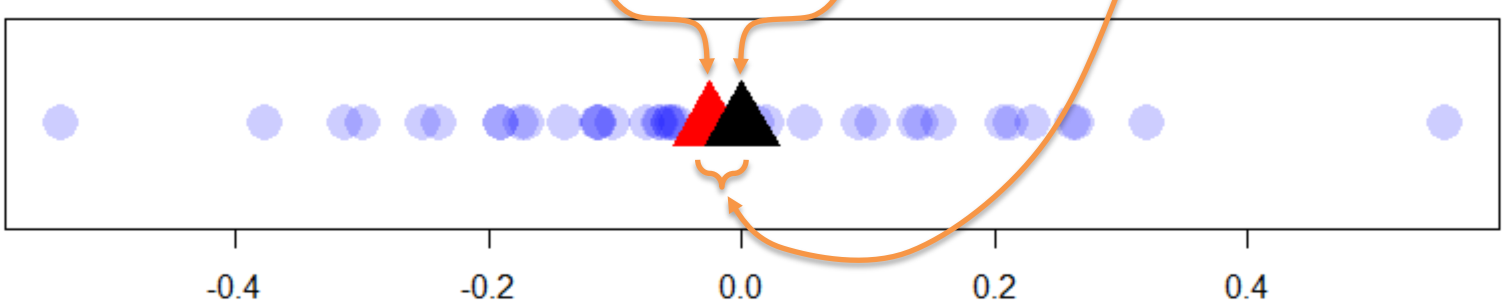
Example: for *i*=1…40

- $X_{i,1}, \dots, X_{i,20} \sim p_\theta = Normal(\theta = 0, \sigma^2 = 1)$
- $\hat{\theta}_i = \frac{1}{20}\sum_{j=1}^{20} X_{i,j}$ the sample mean, plot as ●

Average of $\hat{\theta}_1, \dots, \hat{\theta}_{40}$

True $\theta = 0$

Gap converges to bias (in probability)!



Estimator Variance

Frequentists seek good behaviour, in ideal conditions

- Variance: $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[\left(\hat{\theta} - \mathrm{E}_\theta[\hat{\theta}]\right)^2\right]$

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

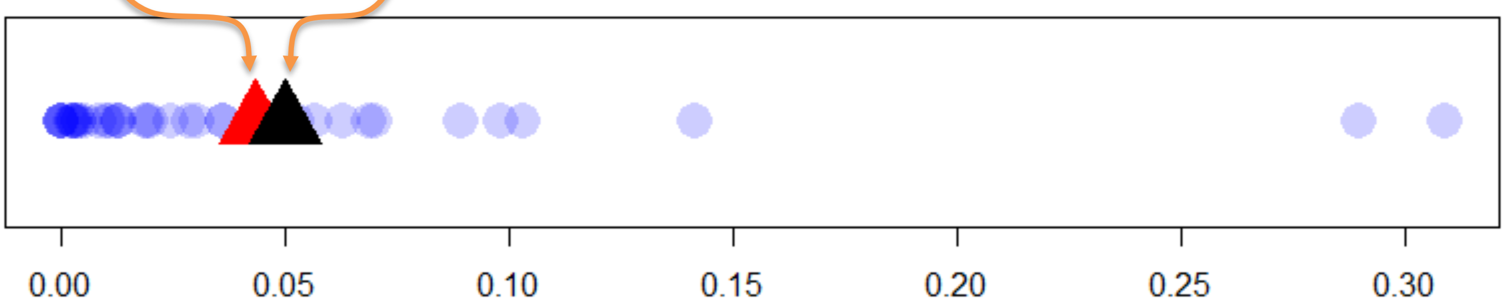
Example cont.

- Plot each $\left(\hat{\theta}_i - \mathrm{E}_\theta[\hat{\theta}_i]\right)^2 = \hat{\theta}_i{}^2$ as

Average of $\hat{\theta}_1{}^2, \ldots, \hat{\theta}_{40}{}^2$

True $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \frac{\sigma^2}{20} = 0.05$

Once again, average converges to true (in probability)!



Asymptotically Well Behaved

For our example estimator (sample mean), we could calculate its exact bias (zero) and variance (σ^2). Usually can't guarantee low bias/variance exactly ☹

Asymptotic properties often hold! ☺

- Consistency: $\hat{\theta}(X_1, \ldots, X_n) \to \theta$ in probability
- Asymptotic efficiency: $\mathrm{Var}_\theta\left(\hat{\theta}(X_1, \ldots, X_n)\right)$ converges to the smallest possible variance of any estimator of θ

Bias closer and closer to zero

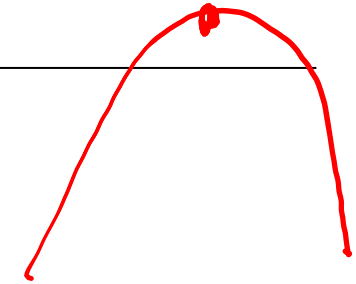
Variance closer & closer to optimal

Maximum-Likelihood Estimation

- A general principle for designing estimators
- Involves optimisation
- $\hat{\theta}(x_1, \ldots, x_n) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(x_i)$
- *"The best estimate is one under which observed data is most likely"*

Fischer

Later: MLE estimators usually well-behaved asymptotically



Example I: Bernoulli

- Know data comes from Bernoulli distribution with unknown parameter (e.g., biased coin); find mean
  - $p_\theta(x) = \begin{cases} \theta, & \text{if } x = 1 \\ 1-\theta, & \text{if } x = 0 \end{cases} = \underline{\theta^x (1-\theta)^{1-x}}$

    (note: $p_\theta(x) = 0$ for all other x)

$\mathcal{L}(\theta) = \log \prod_i p_\theta(x_i) = \sum_i \log p_\theta(x_i) = \sum_i x_i \log \theta + (1-x_i) \log (1-\theta)$

$= \bar{x} \log \theta + (n-\bar{x}) \log (1-\theta)$ $\qquad \bar{x} = \sum_i x_i$

$\frac{\partial \mathcal{L}(\hat{\theta})}{\partial \theta} = \frac{\bar{x}}{\theta} - \frac{(n-\bar{x})}{1-\theta} = 0$

$\theta = \bar{x}/n$

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

Derivation: https://stats.stackexchange.com/questions/275380/maximum-likelihood-estimation-for-bernoulli-distribution

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

Derivation: https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood

MLE 'algorithm'

1. Given data $X_1, \ldots, X_n$ 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 gradient descent)

Mini Summary

- Frequentist school of thought
- Point estimates
- Quality: bias, variance, consistency, asymptotic efficiency
- Maximum-likelihood estimation (MLE)

Next: Statistical Decision Theory, Extremum estimators

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 $\delta(\boldsymbol{x}) \in A$ an action space
  - E.g. Point estimate $\hat{\theta}(x_1, \ldots, x_n)$
  - E.g. Out-of-sample prediction $\hat{Y}_{n+1} | X_1, Y_1, \ldots, X_n, Y_n, X_{n+1}$
- Loss function $l(a, \theta)$: economic cost, error metric
  - E.g. square loss of estimate $\left(\hat{\theta} - \theta\right)^2$
  - 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 $R_\theta[\delta] = \mathrm{E}_{\boldsymbol{X}\sim\theta}[l(\delta(\boldsymbol{X}),\theta)]$
  - E.g. true test error
  - aka generalization error
- Want: Choose δ to minimise $R_\theta[\delta]$
- Can't directly! Why?
- ERM: Use training set ***X*** to approximate p_θ
  - Minimise empirical risk $\hat{R}_\theta[\delta] = \frac{1}{n}\sum_{i=1}^{n} l(\delta(X_i),\theta)$

Decision theory vs. Bias-variance

We've already seen

- Bias: $\mathrm{B}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[\hat{\theta}(X_1, \ldots, X_n)\right] - \theta$
- Variance: $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[(\hat{\theta} - \mathrm{E}_\theta[\hat{\theta}])^2\right]$

But are they equally important? How related?

- Bias-variance decomposition of square-loss risk

$$\mathrm{E}_\theta\left[\left(\theta - \hat{\theta}\right)^2\right] = [\mathrm{B}(\hat{\theta})]^2 + \mathrm{Var}_\theta(\hat{\theta})$$

Mini Summary

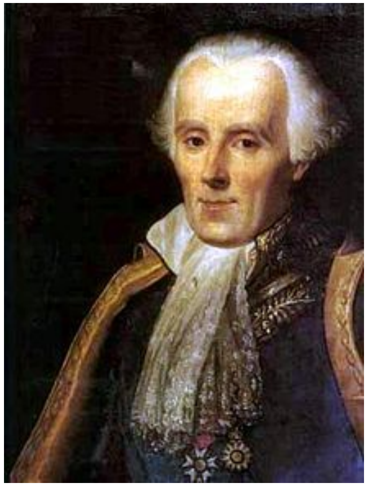
- Decision theory: Utility-based, Minimise risk
- Many familiar learners minimise loss over data (ERM)

Next: Last but not least, the Bayesian paradigm

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

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

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

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

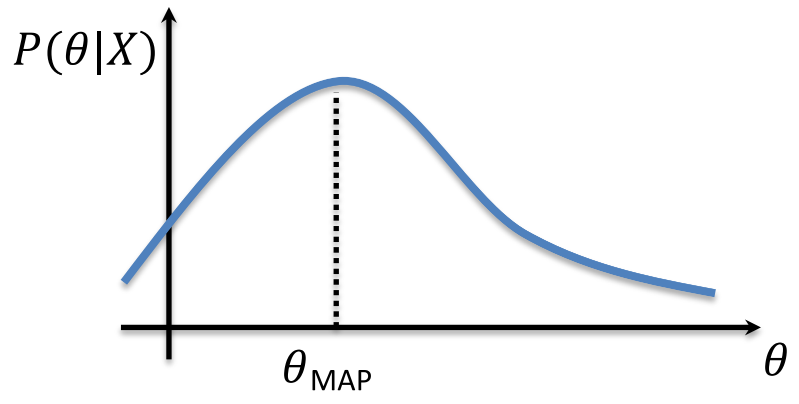
Recognise as (unnormalized) Normal!

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

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
$$\hat{\theta} \in \operatorname{argmax}_\theta P(\theta|X = x)$$
$$= \operatorname{argmax}_\theta \frac{P(X = x|\theta)P(\theta)}{P(X = x)}$$
$$= \operatorname{argmax}_\theta P(X = x|\theta)P(\theta)$$

- **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

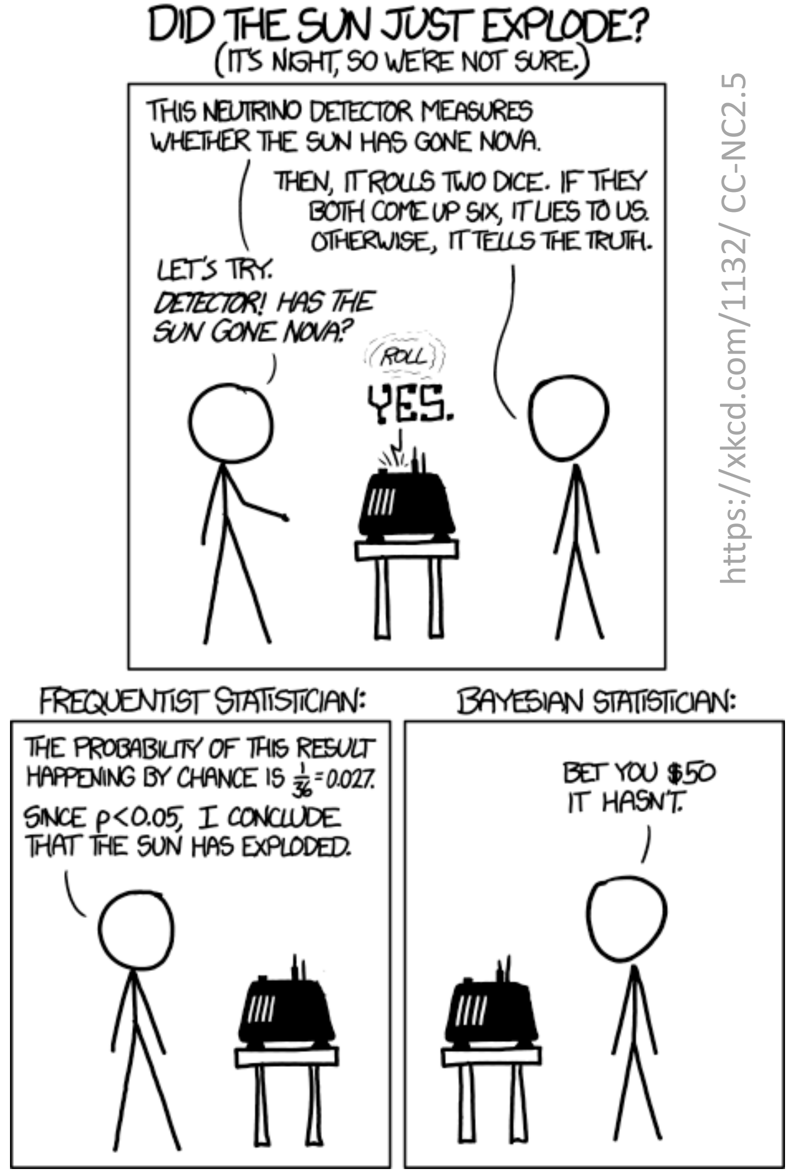
Rafael Irizarry 2014/10/13

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



(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 pros and cons

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

Mini Summary

- Bayesian paradigm: Its all in the prior!
- Bayesian point estimate: MAP
- 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!