Lecture 2a. Statistical Schools of Thought: Frequentist

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

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

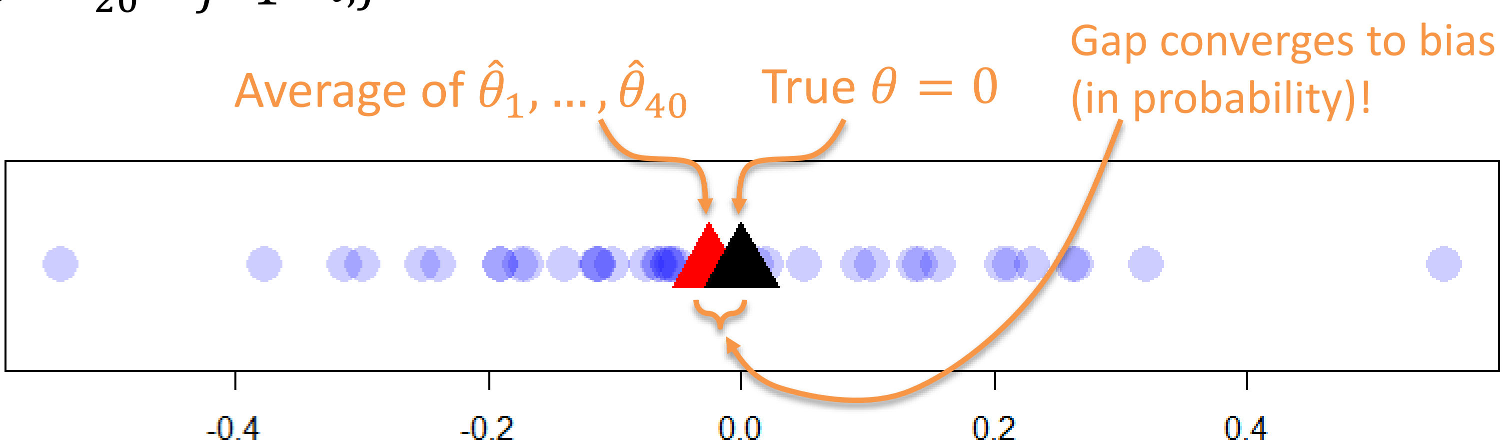
Frequentists seek good behaviour, in ideal conditions

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

Subscript θ means data *really* comes from p_θ

Example: for *i*=1…40

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



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

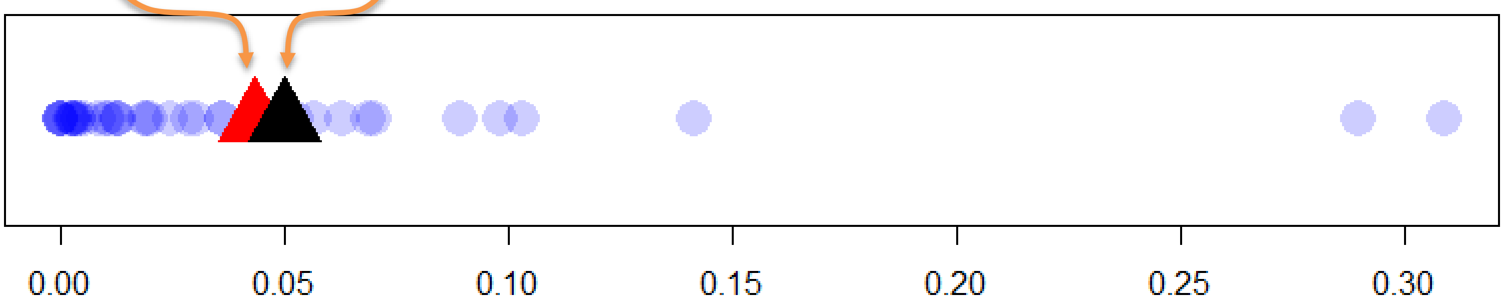
Example cont.

- 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! ☺

Bias closer and closer to zero

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

Variance closer & closer to optimal

Amazing Cramér-Rao lower bound (**outside subject scope**): $\mathrm{Var}_\theta\left(\hat{\theta}\right) \geq \frac{1}{I(\theta)}$ with $I(\theta)$ the Fisher information of p_θ for any $\hat{\theta}$

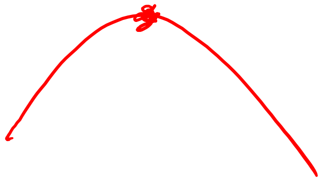
Maximum-Likelihood Estimation

- 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

$\mathcal{L}(\theta) = \log \prod_{i=1}^{n} p_\theta(x_i) = \boxed{\bar{X}\log\theta + (n-\bar{X})\log(1-\theta)}$

$\bar{X} = \sum_{i=1}^{n} X_i$

- 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 $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\frac{d}{d\theta}\mathcal{L}(\hat{\theta}) = \frac{\bar{X}}{\hat{\theta}} - \frac{n-\bar{X}}{1-\hat{\theta}} = 0 \Rightarrow \hat{\theta} = \frac{\bar{X}}{n}$

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)

Summary

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

Next time: Statistical Decision Theory, Extremum estimators

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