COMP90051 **Statistical Machine Learning**

Semester 2, 2015

Lecturer: Ben Rubinstein

2. Statistical Schools

THE UNIVERSITY OF MELBOURNE

Statistical Schools of Thought

Remainder of deck is to provide *intuition* into how algorithms in this subject come about and inter-relate

Based on Berkeley CS 294-34 tutorial slides by Ariel Kleiner

Frequentist Statistics

- Abstract problem
  - Given: $X_1, X_2, \ldots, X_n$ drawn i.i.d. from some distribution
  - Want to: identify unknown distribution
- 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 ….)
  - Select $\hat{\theta}\,(x_1, \ldots, x_n)$ some function (or statistic) of data

Hat means estimate or estimator

- Examples
  - Given n coin flips, determine probability of landing heads
  - Building a classifier is a very related problem

How do Frequentists Evaluate Estimators?

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

Subscript θ means data *really* comes from p_θ

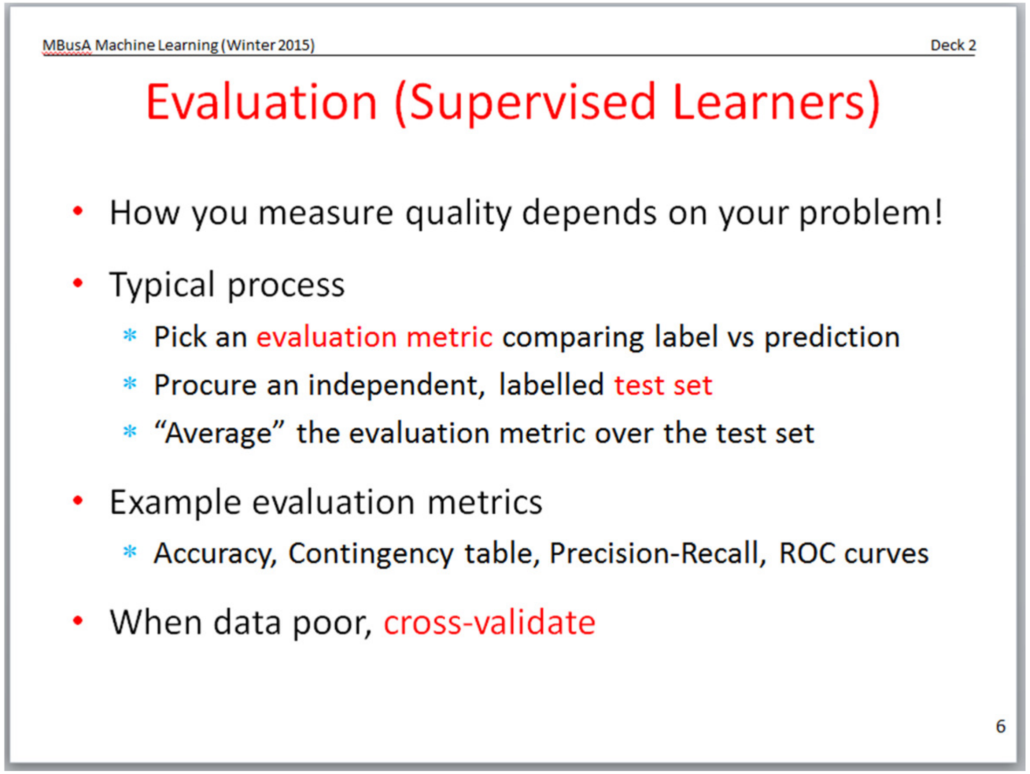
- Variance: $Var_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$
  - Efficiency: estimate has minimal variance

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

- Risk
  - Loss function $L(\theta, \hat{\theta})$ measures error of estimate
  - Risk is expected loss $R(\theta) = E_\theta[L(\theta, \hat{\theta})]$
  - Square loss vs bias-variance $E_\theta\left[(\theta - \hat{\theta})^2\right] = [B(\theta)]^2 + Var_\theta(\hat{\theta})$
- Consistency: $\hat{\theta}(X_1, \dots, X_n)$ converges to θ as *n* gets big

Is this *"Just Theoretical"*™ ?

- Recall Deck 1 →
- Those evaluation metrics? They're just estimators of a performance parameter
- Example: error



- Bias, Variance, etc. indicate quality of approximation

Maximum-Likelihood Estimation

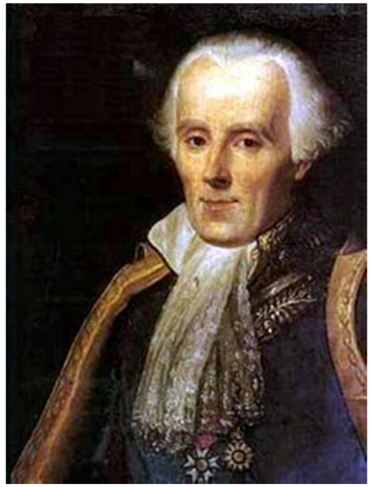
- A general principle for designing estimators; *always* has nice properties
  - Consistency
  - Asymptotic efficiency
  - Asymptotic normality
- Involves optimisation
- $\hat{\theta}(x_1, \ldots, x_n) = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(x_i)$
  - Question: Why a *product*?

Fischer

Example

- Know data comes from Normal distribution with variance 1 but unknown mean; find mean
- MLE for mean
  - $p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)$
  - Exercise: Maximising likelihood yields $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Exercise: Bias, Variance, Consistency?

Bayesian Statistics



Laplace

- Probabilities correspond to beliefs
- Parameters
  - Modeled as r.v.'s having distributions
  - Prior belief in θ encoded by prior distribution $\mathrm{P}(\theta)$
  - Write likelihood of data P(X) as conditional $\mathrm{P}(\mathrm{X}|\theta)$
  - Rather than point estimate $\hat{\theta}$, Bayesians update belief $\mathrm{P}(\theta)$ with observed data to $\mathrm{P}(\theta|X)$ the posterior distribution

More Detail (Probabilistic Inference)

Bayes

- Bayesian machine learning
  - Start with prior $\mathrm{P}(\theta)$ and likelihood $\mathrm{P}(X|\theta)$
  - Observe data $X = x$
  - Update prior to posterior $\mathrm{P}(\theta|X = x)$
- We'll later cover tools to get the posterior
  - Bayes Theorem: revereses order of conditioning

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

  - Marginalisation: eliminates unwanted variables

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

Example

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

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

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

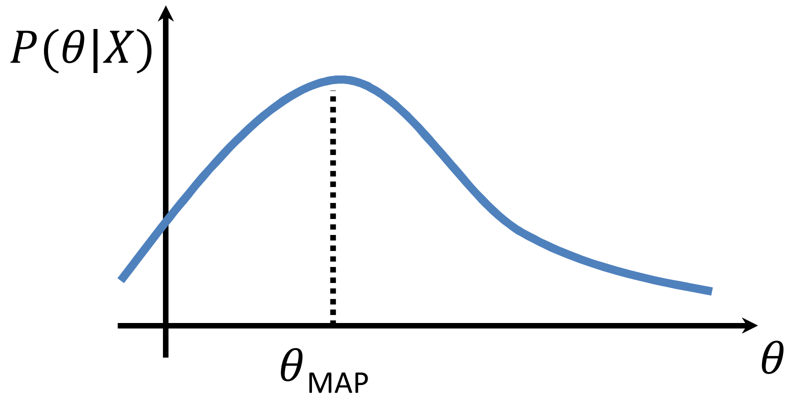
$$= \left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{(1-\theta)^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{\theta^2}{2}\right)\right]$$

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

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

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)



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

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