# COMP90051 **Statistical Machine Learning**

Semester 2, 2017

Lecturer: Trevor Cohn

## 2. Statistical Schools

THE UNIVERSITY OF MELBOURNE

Adapted from slides by Ben Rubinstein

# Statistical Schools of Thought

Remainder of lecture 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 $\Theta$ (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: $B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}(X_1, \ldots, X_n)] - \theta$
- Variance: $Var_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$
  - Efficiency: estimate has minimal varianc
- Square loss vs bias-variance

$$E_\theta\left[\left(\theta - \hat{\theta}\right)^2\right] = [B(\theta)]^2 + Var_\theta(\hat{\theta})$$

- Consistency: $\hat{\theta}(X_1, \ldots, X_n)$ converges to $\theta$ as *n* gets big

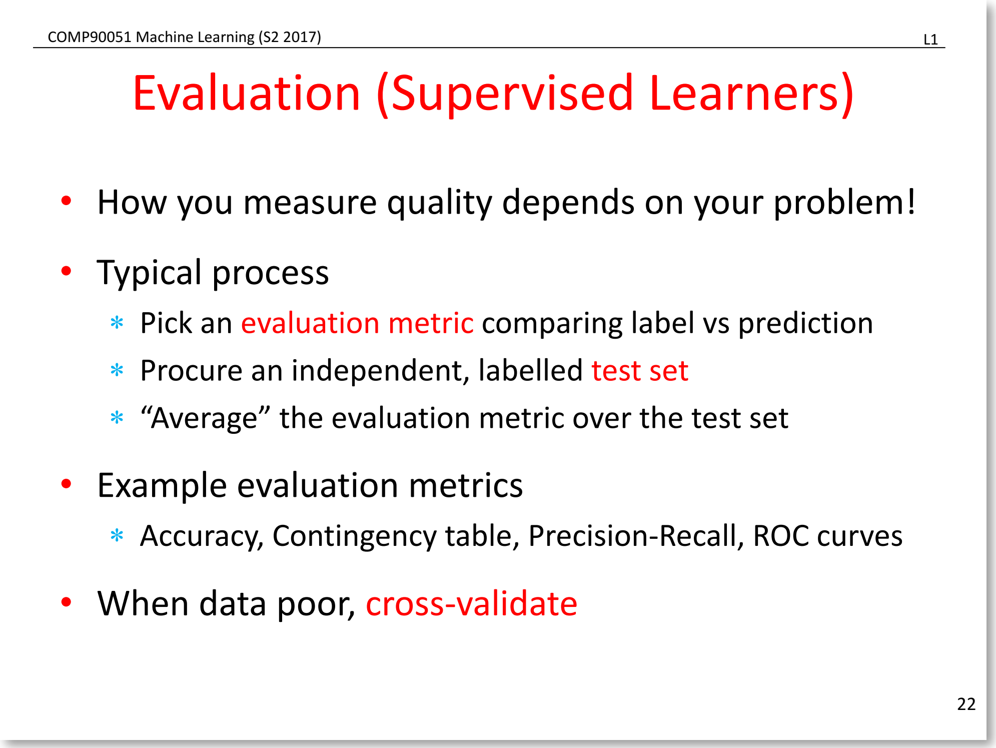
… more on this later in the subject …

Subscript θ means data *really* comes from $p_\theta$

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

# Is this “*Just Theoretical*”™ ?

- Recall Lecture 1 →



COMP90051 Machine Learning (S2 2017) L1

## 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

- 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
- Involves optimisation
- $\hat{\theta}(x_1, \dots, x_n) = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(x_i)$
  - Question: Why a *product*?

Fischer

# Example I: 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* $\sigma^2$ based on $p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$ with $\theta = (\mu, \sigma^2)$

# Example II: Bernoulli

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

    (note: $p_\theta(x)$ = 0 for all other $x$)
  - Maximising likelihood yields $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Corrected typo after lecture, 27/7/17

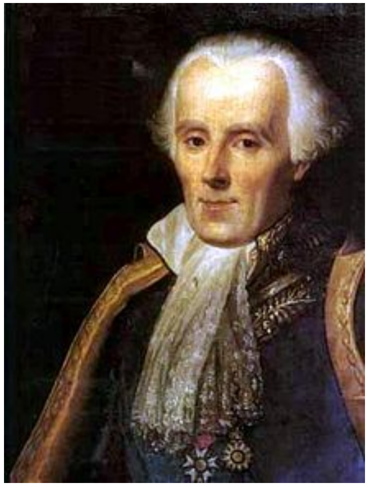
# MLE 'algorithm'

1. given data $x_1, \dots, x_n$ define probability distribution, $p_\theta$, 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 $\theta$
   2. set to 0 and solve
      (failing that, use iterative gradient method)

# Bayesian Statistics



Laplace

- Probabilities correspond to beliefs
- Parameters
  - Modeled as r.v.'s having distributions
  - Prior belief in $\theta$ 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: reverses order of conditioning

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

  - Marginalisation: eliminates unwanted variables

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

# Example

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

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

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

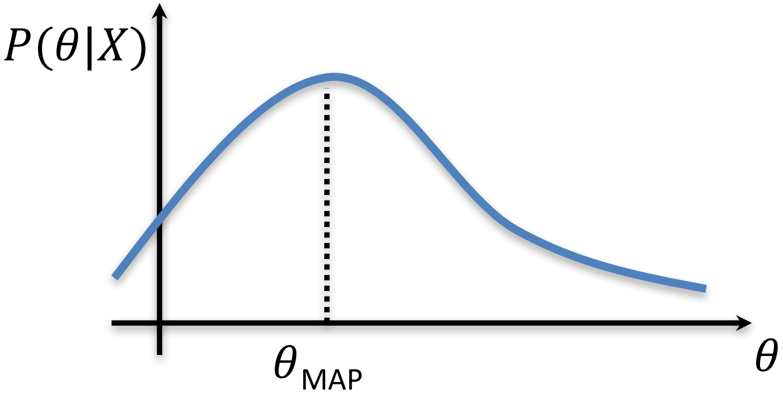
$$= \left[\boxed{\frac{1}{\sqrt{2\pi}}} exp\left(-\frac{(1-\theta)^2}{2}\right)\right]\left[\boxed{\frac{1}{\sqrt{2\pi}}} exp\left(-\frac{\theta^2}{2}\right)\right]$$

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

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

# 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)



# MLE in Bayesian context

- MLE formulation: find parameters that best fit data

$$\hat{\theta} = \mathrm{argmax}_{\theta}\, P(X = x|\theta)$$

- Consider the MAP under a Bayesian formulation

$$\begin{aligned}\hat{\theta} &= P(\theta|X = x)\\ &= \mathrm{argmax}_{\theta} \frac{P(X = x|\theta)P(\theta)}{P(X = x)}\\ &= \mathrm{argmax}_{\theta}\, P(X = x|\theta)P(\theta)\end{aligned}$$

- Difference is **prior** $P(\theta)$; assumed *uniform* for MLE

# 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 $(X_1, Y_1)$,… ,$(X_n, Y_{\mathrm{n}})$
  - 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
- Discriminative vs. Generative models