

# ECE 5424: Introduction to Machine Learning

Topics:

- Statistical Estimation (MLE, MAP, Bayesian)

Readings: Barber 8.6, 8.7

Stefan Lee  
Virginia Tech

# Administrative

- HW1
  - Due on Wed 9/14, 11:55pm
  - Problem 2.2 : Two cases (in ball, out of ball)
- Project Proposal
  - Due: Tue 09/21, 11:55 pm
  - $\leq 2$  pages, NIPS format

# Recap from last time

# Procedural View

- Training Stage:
  - Raw Data  $\rightarrow x$  (Feature Extraction)
  - Training Data  $\{ (x,y) \} \rightarrow f$  (Learning)
- Testing Stage
  - Raw Data  $\rightarrow x$  (Feature Extraction)
  - Test Data  $x \rightarrow f(x)$  (Apply function, Evaluate error)

# Statistical Estimation View

- Probabilities to rescue:
  - $x$  and  $y$  are *random variables*
  - $D = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) \sim P(X, Y)$
- IID: Independent Identically Distributed
  - Both training & testing data sampled IID from  $P(X, Y)$
  - Learn on training set
  - Have some hope of *generalizing* to test set

# Interpreting Probabilities

- What does  $P(A)$  mean?
- Frequentist View
  - $\lim_{N \rightarrow \infty} \#(A \text{ is true})/N$
  - limiting frequency of a repeating non-deterministic event
- Bayesian View
  - $P(A)$  is your “belief” about  $A$
- Market Design View
  - $P(A)$  tells you how much you would bet

# Concepts

- Likelihood
  - How much does a certain hypothesis explain the data?
- Prior
  - What do you believe before seeing any data?
- Posterior
  - What do we believe after seeing the data?

# KL-Divergence / Relative Entropy

- An asymmetric measure of the distance between two distributions:

$$KL[p||q] = \sum_x p(x) [\log p(x) - \log q(x)]$$

- $KL > 0$  unless  $p = q$  then  $KL = 0$
- Tells you the extra cost if events were generated by  $p(x)$  but instead of charging under  $p(x)$  you charged under  $q(x)$ .

# Plan for Today

- Statistical Learning
  - Frequentist Tool
    - Maximum Likelihood
  - Bayesian Tools
    - Maximum A Posteriori
    - Bayesian Estimation
- Simple examples (like coin toss)
  - But SAME concepts will apply to sophisticated problems.

# Your first probabilistic learning algorithm

- After taking this ML class, you drop out of VT and join an illegal betting company.
  - Specializing in mascot fist fights.
- Your new boss asks you:
  - If the VT and UVA mascots fight tomorrow, will Hokiebird win or lose?
- You say: what is the record?
  - W, L, L, W, W
- You say:  $P(\text{Hokiebird Wins}) = \dots$



**UNKNOWN TARGET FUNCTION**

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

*(ideal credit approval function)*

**TRAINING EXAMPLES**

$$(x_1, y_1), \dots, (x_N, y_N)$$

*(historical records of credit customers)*

**LEARNING  
ALGORITHM**

$$\mathcal{A}$$

**FINAL  
HYPOTHESIS**

$$g \approx f$$

*(final credit approval formula)*

**HYPOTHESIS SET**

$$\mathcal{H}$$

*(set of candidate formulas)*

# Maximum Likelihood Estimation

- Goal: Find a good  $\theta$
- What's a good  $\theta$ ?
  - One that makes it likely for us to have seen this data
  - Quality of  $\theta$  =  $\text{Likelihood}(\theta; D) = P(\text{data} \mid \theta)$

# Why Max-Likelihood?

- Leads to “natural” estimators
- MLE is OPT if model-class is correct

# Sufficient Statistic

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- $D_1 = \{1, 1, 1, 0, 0, 0\}$
- $D_2 = \{1, 0, 1, 0, 1, 0\}$
- A function of the data  $\phi(Y)$  is a sufficient statistic, if the following is true

$$\sum_{i \in D_1} \phi(y_i) = \sum_{i \in D_2} \phi(y_i) \quad \Rightarrow \quad L(\theta; D_1) = L(\theta; D_2)$$

# How many flips do I need?

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Boss says: Last year:
  - 3 heads/wins-for-Hokiebird
  - 2 tails/losses-for-Hokiebird.
- You say:  $\theta = 3/5$ , I can prove it!
- He says: What if
  - 30 heads/wins-for-Hokiebird
  - 20 tails/losses-for-Hokiebird
- You say: Same answer, I can prove it!
- **He says: What's better?**
- You say: Humm... The more the merrier???
- He says: Is this why I am paying you the big bucks???

# Bayesian Estimation

- Boss says: What is I know the Hokiebird is a better fighter on closer to Thanksgiving?
  - (fighting for his life)
- You say: Bayesian it is then..

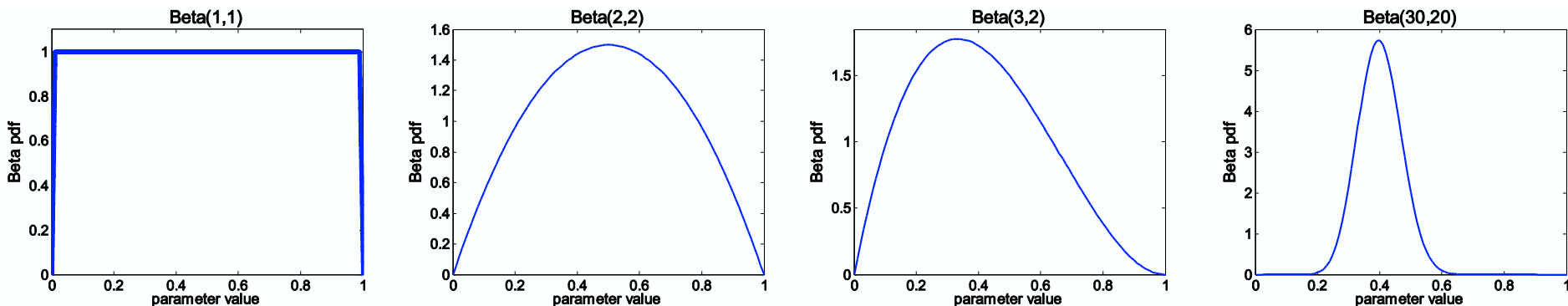
# Priors

- What are priors?
  - Express beliefs before experiments are conducted
  - Computational ease: lead to “good” posteriors
  - Help deal with unseen data
  - Regularizers: More about this in later lectures
- Conjugate Priors
  - Prior is conjugate to likelihood if it leads to itself as posterior
  - Closed form representation of posterior

# Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

- Demo:
  - <http://demonstrations.wolfram.com/BetaDistribution/>



- Benefits of conjugate priors

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

# MAP for Beta distribution

$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1} (1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

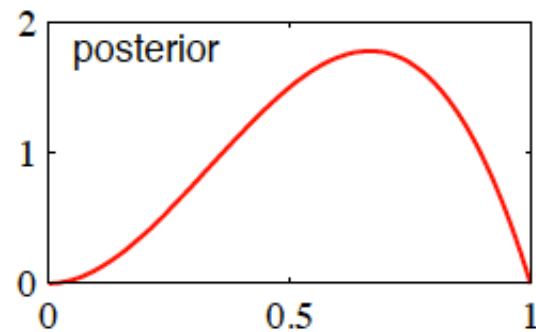
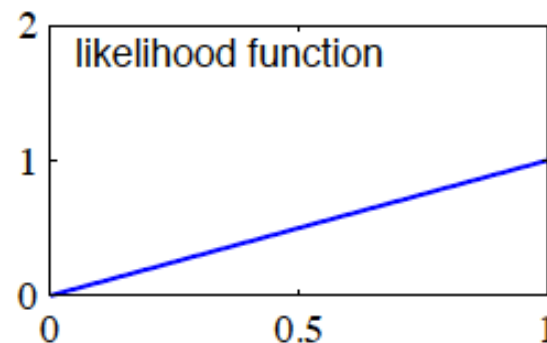
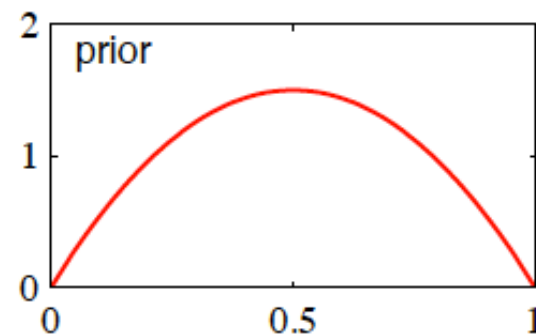
- MAP: use most likely parameter:

$$\hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) =$$

- Beta prior equivalent to extra W/L matches
- As  $N \rightarrow \infty$ , prior is “forgotten”
- **But, for small sample size, prior is important!**

# Effect of Prior

- Prior = Beta(2,2)
  - $\theta_{\text{prior}} = 0.5$
- Dataset = {H}
  - $L(\theta) = \theta$
  - $\theta_{\text{MLE}} = 1$
- Posterior = Beta(3,2)
  - $\theta_{\text{MAP}} = (3-1)/(3+2-2) = 2/3$



# What you need to know

- Statistical Learning:
  - Maximum likelihood
    - Why MLE?
  - Sufficient statistics
  - Maximum a posteriori
  - Bayesian estimation (return an entire distribution)
  - Priors, posteriors, conjugate priors
  - Beta distribution (conjugate of bernoulli)