

CS 461: Machine Learning

Lecture 6

Dr. Kiri Wagstaff
kiri.wagstaff@calstatela.edu

Plan for Today

- Note: Room change for 2/16: E&T A129
- Solution to Midterm
- Solution to Homework 3
- Parametric methods
 - Data comes from distribution
 - Bernoulli, Gaussian, and their parameters
 - How good is a parameter estimate? (bias, variance)
- Bayes estimation
 - ML: use the data
 - MAP: use the prior and the data
 - Bayes estimator: integrated estimate (weighted)
- Parametric classification
 - Maximize the posterior probability

Review from Lecture 5

- Probability
 - Axioms
- Bayesian Learning
 - Classification
 - Bayes's Rule
 - Bayesian Networks
 - Naïve Bayes Classifier
 - Association Rules

Parametric Methods

Chapter 4

2/9/08

CS 461, Winter 2008

4

Parametric Learning

- Assume: data x comes from a distribution $p(x)$
- Model this distribution by selecting parameters θ
 - e.g., $\mathcal{N}(\mu, \sigma^2)$ where $\theta = \{\mu, \sigma^2\}$

Maximum Likelihood Model

- (Last time: Max Likelihood Classification)
- Likelihood of θ given the sample $\mathcal{X}$

$$l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$$

- Log likelihood

$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$$

- Maximum likelihood estimator (MLE)

$$\theta^* = \operatorname{argmax}_{\theta} \mathcal{L}(\theta|\mathcal{X})$$

Example: do you wear glasses?

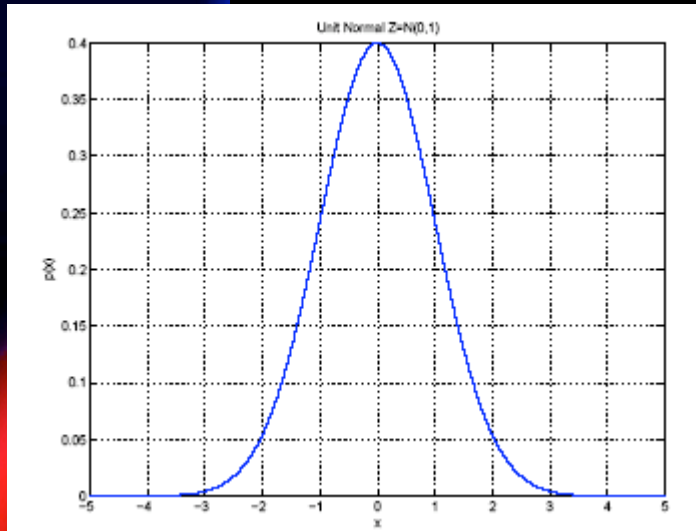
- **Bernoulli:** Two states, x in $\{0,1\}$

$$P(x) = p_o^x (1 - p_o)^{(1-x)}$$

$$\mathcal{L}(p_o | \mathcal{X}) = \log \prod_t p_o^{x^t} (1 - p_o)^{(1-x^t)}$$

$$\text{MLE: } p_o = \sum_t x^t / N$$

Gaussian (Normal) Distribution



μ σ

- $p(x) = \mathcal{N}(\mu, \sigma^2)$

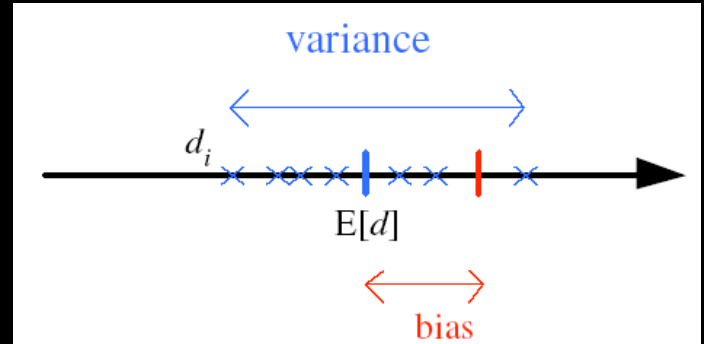
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

- MLE for μ and σ^2 :

$$m = \frac{\sum_t x^t}{N}$$
$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

How good is that estimate?

- Let d be the estimate of θ
 - It is also a random variable
 - Technically, it's $d(X)$ since it depends on X
 - $E[d]$ is the expected value of d (regardless of X)
- Bias: $b_{\theta}(d) = E[d] - \theta$
 - How far off the correct value is it?
- Variance: $Var(d) = E[(d - E[d])^2]$
 - How much does it change with different X ?
- Mean square error:
$$r(d, \theta) = E[(d - \theta)^2]$$
$$= (E[d] - \theta)^2 + E[(d - E[d])^2]$$
$$= \text{Bias}^2 + \text{Variance}$$



Bayes Estimator:

Using what we already know

- Prior information $p(\theta)$
- Bayes's rule (get posterior):
$$p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) p(\theta) / p(\mathcal{X})$$
- Maximum a Posteriori (MAP):
$$\theta_{\text{MAP}} = \operatorname{argmax}_{\theta} p(\theta|\mathcal{X})$$
- Maximum Likelihood (ML):
$$\theta_{\text{ML}} = \operatorname{argmax}_{\theta} p(\mathcal{X}|\theta)$$
- Bayes Estimator:
$$\theta_{\text{Bayes}} = E[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X}) d\theta$$
 - In a sense, it is the “weighted mean” for θ
 - For our purposes, $\theta_{\text{MAP}} = \theta_{\text{Bayes}}$

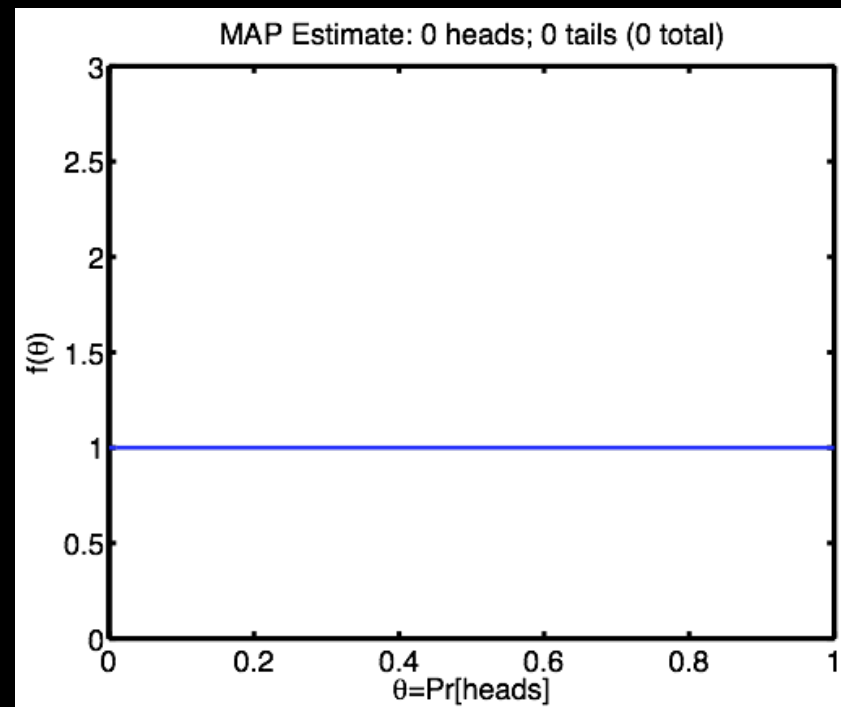
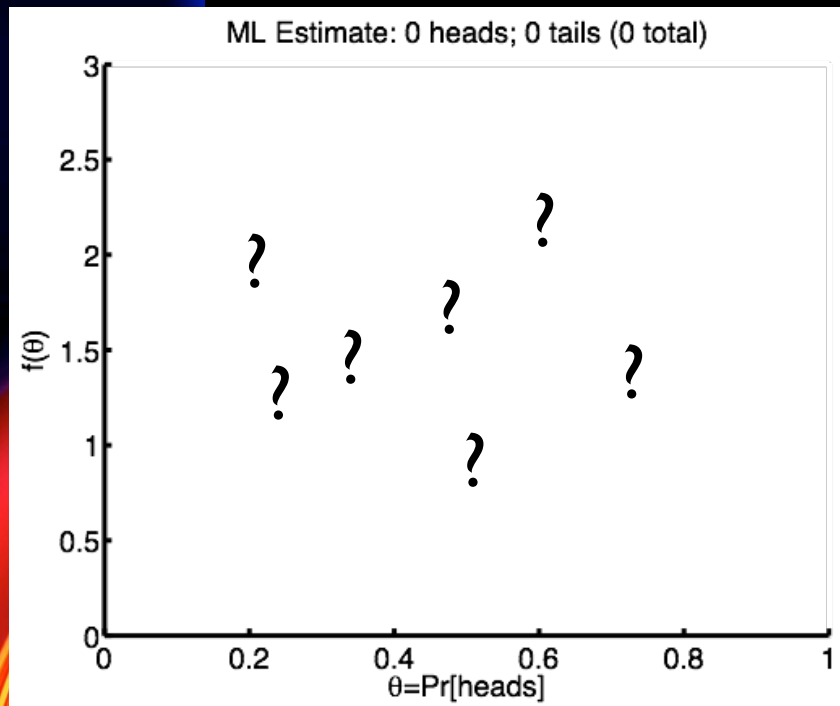
Bayes Estimator: Continuous Example

- Assume $x^t \sim \mathcal{N}(\theta, \sigma_0^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta_{\text{ML}} = m$ (sample mean)
- $\theta_{\text{MAP}} = \theta_{\text{Bayes}} =$

$$E[\theta | \mathcal{X}] = \frac{N / \sigma_0^2}{N / \sigma_0^2 + 1 / \sigma^2} m + \frac{1 / \sigma^2}{N / \sigma_0^2 + 1 / \sigma^2} \mu$$

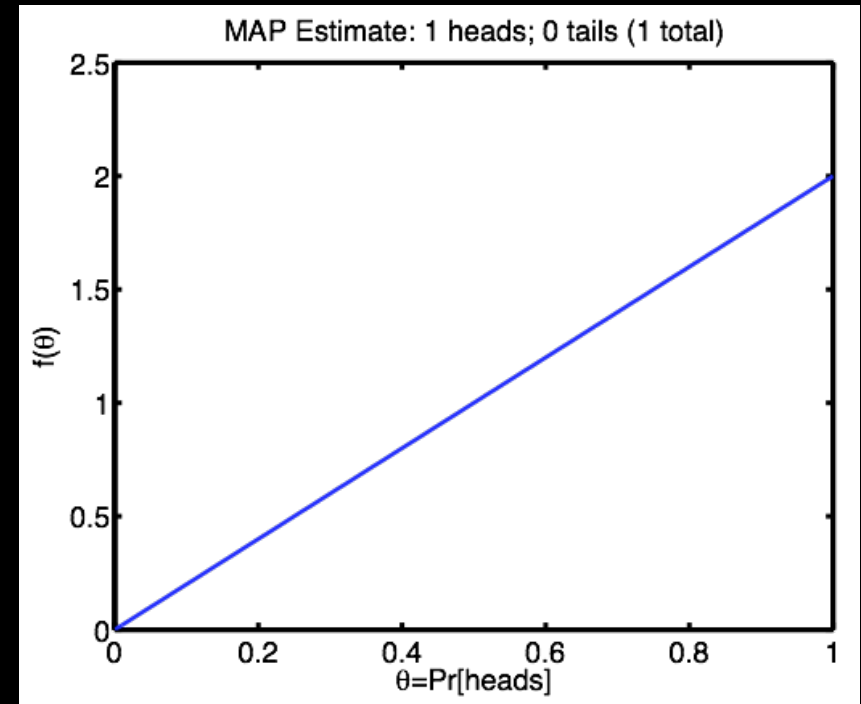
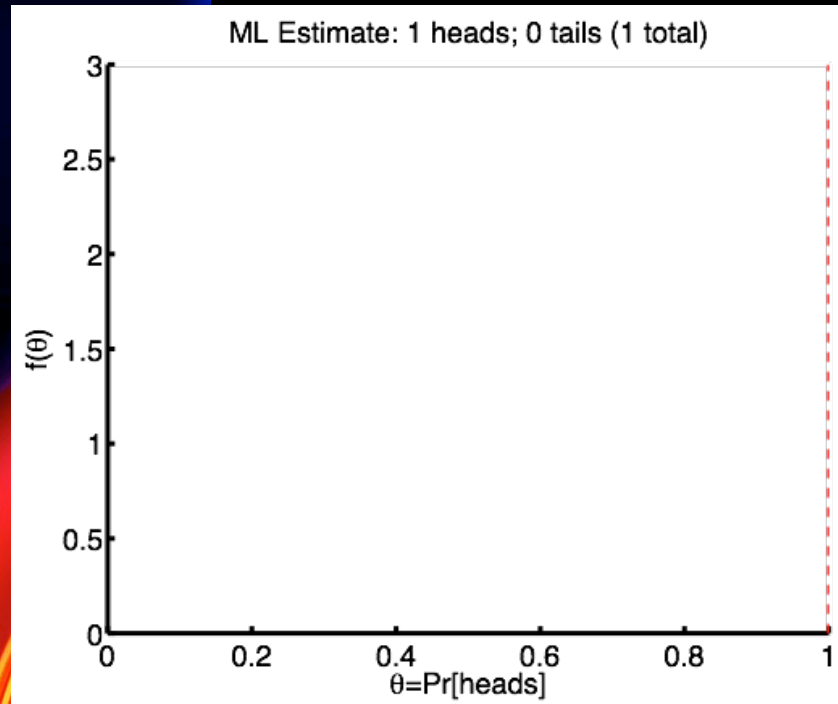
- Estimated mean = weighted average of sample mean m and prior mean μ
 - Weights indicate how much you trust the sample

Example: Coin flipping



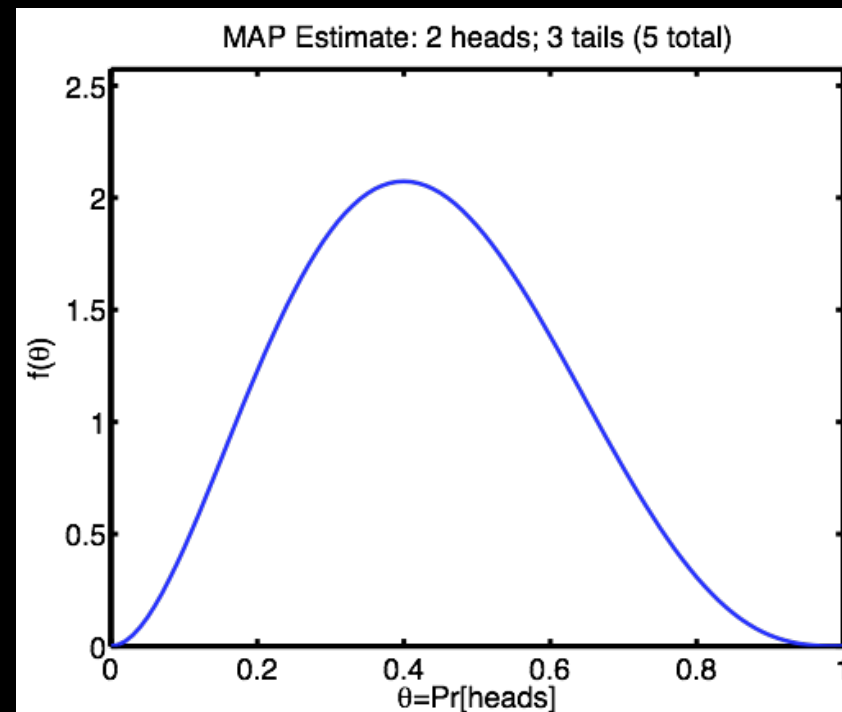
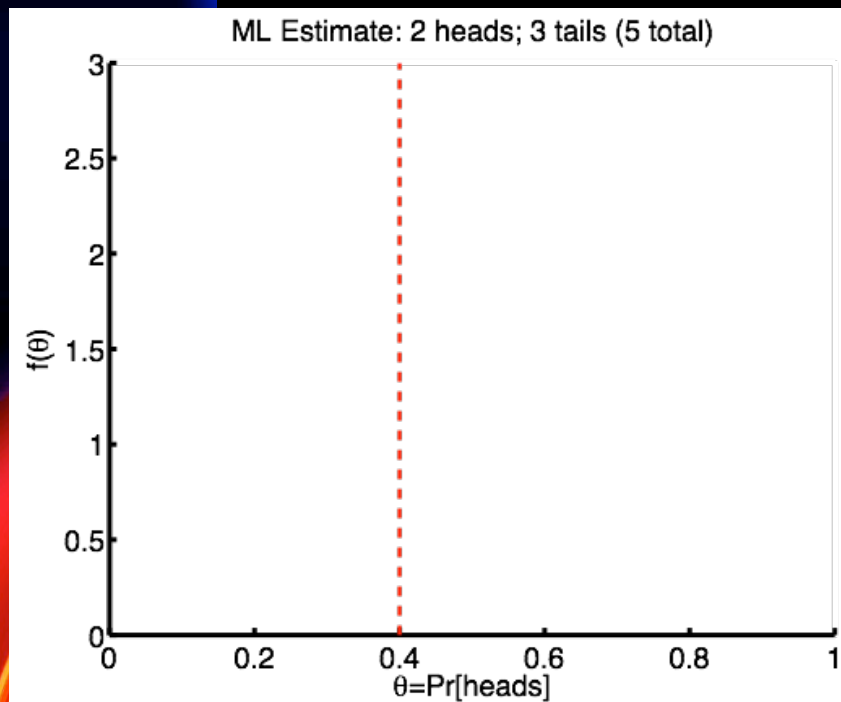
0 flips total

Example: Coin flipping



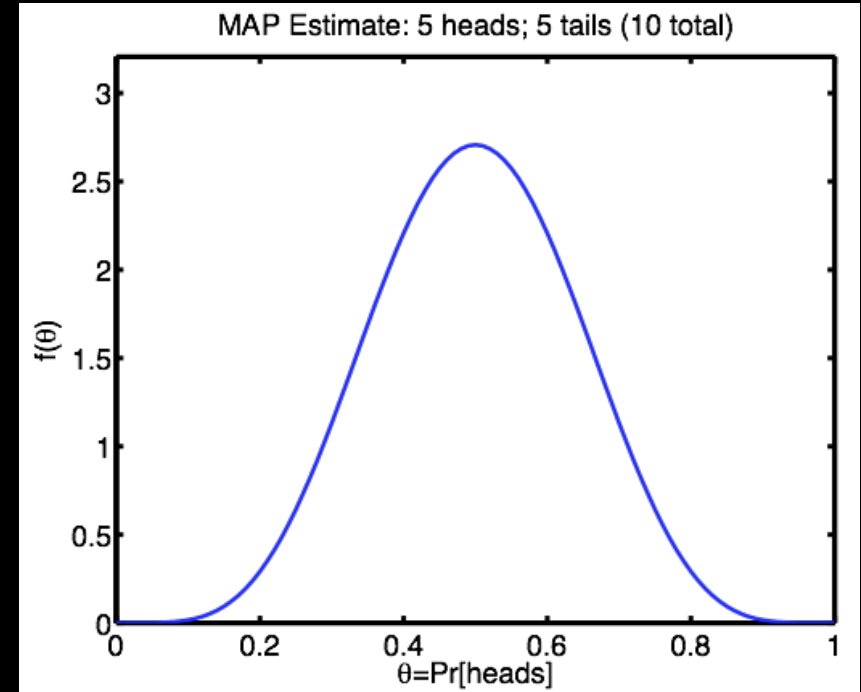
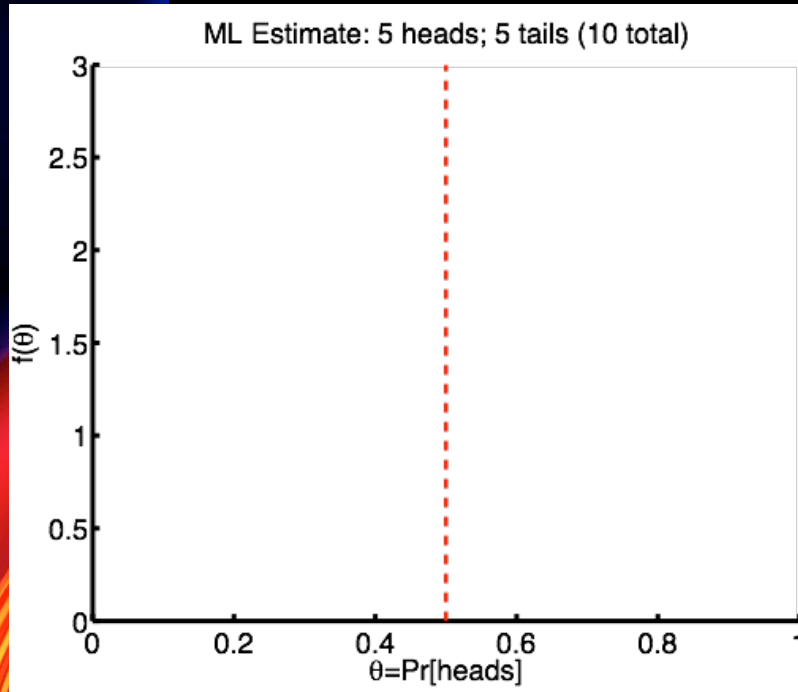
1 flip total

Example: Coin flipping



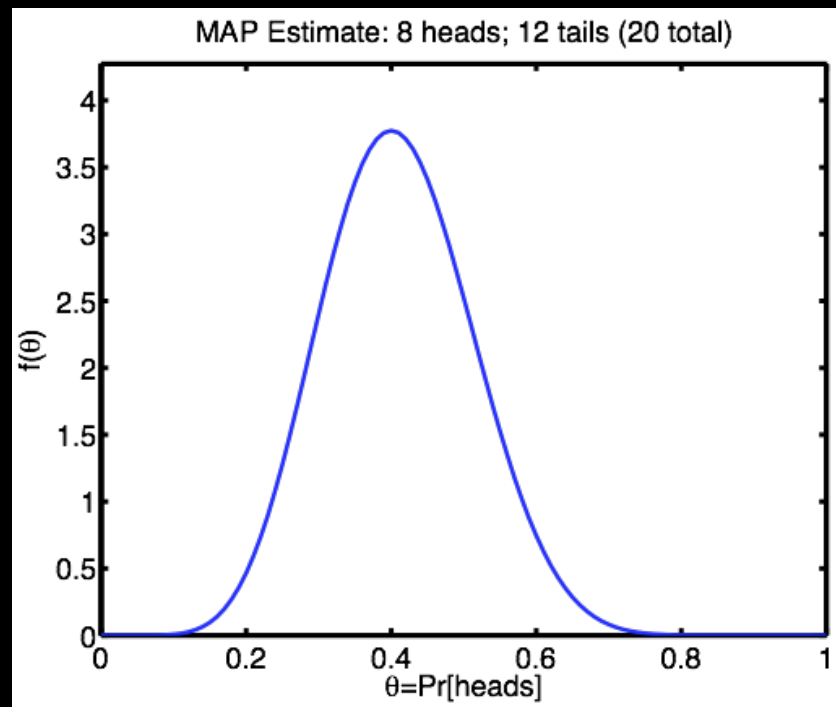
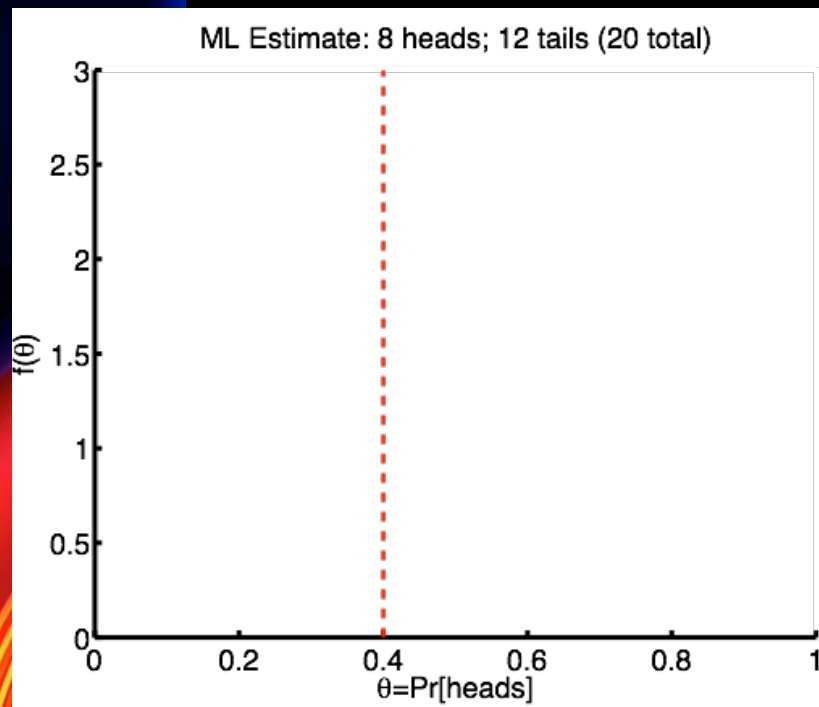
5 flips total

Example: Coin flipping



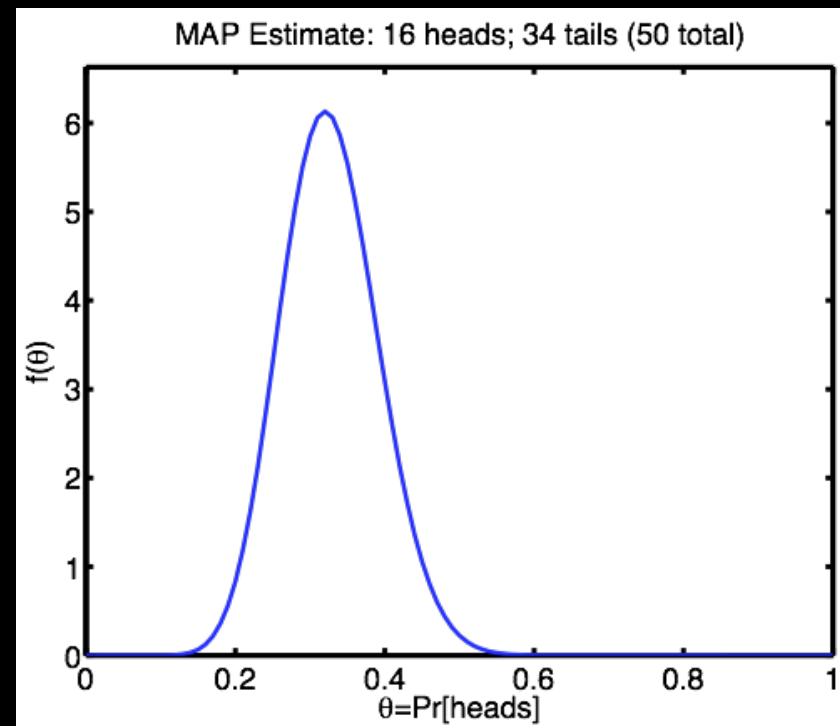
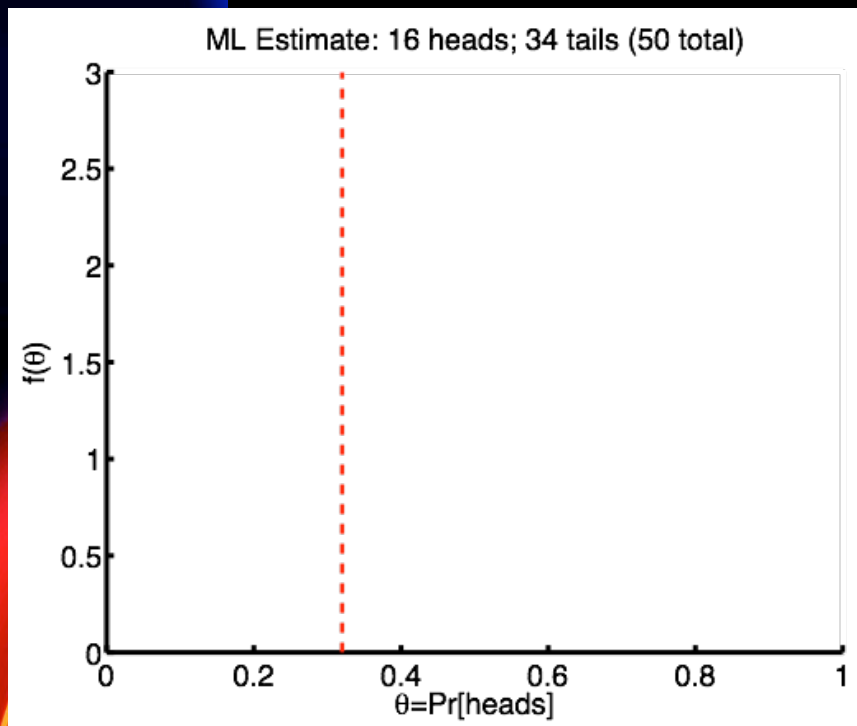
10 flips total

Example: Coin flipping



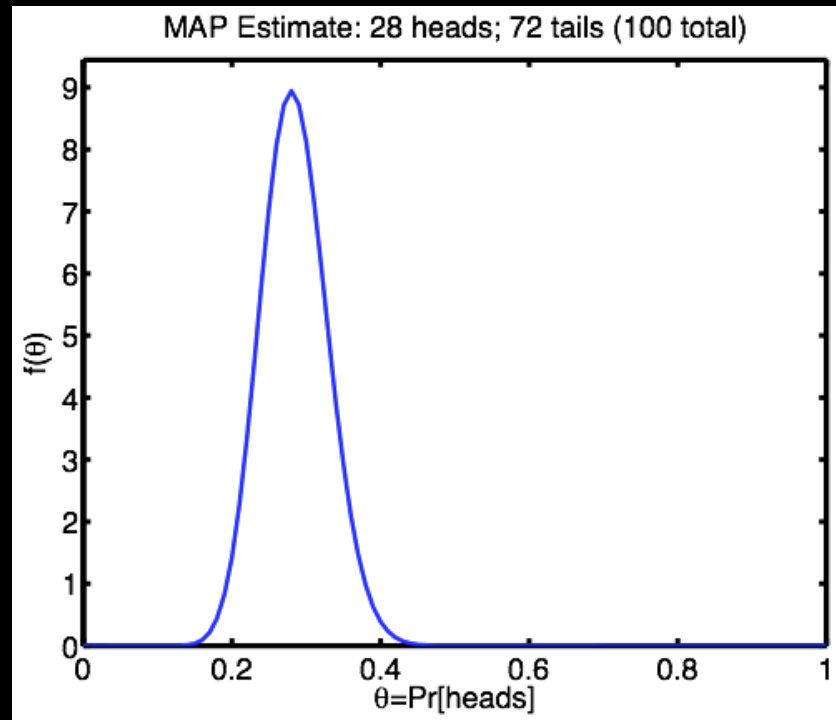
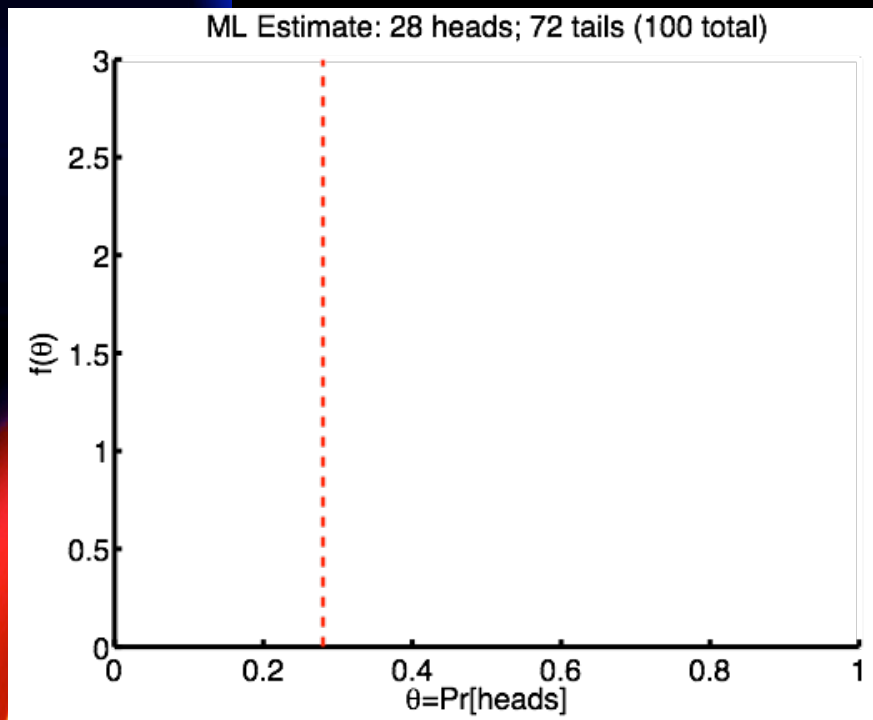
20 flips total

Example: Coin flipping



50 flips total

Example: Coin flipping



100 flips total

Parametric Classification

- Remember Naïve Bayes?
- Maximum likelihood estimator (MLE)

$$C^{\text{predict}} = \underset{c}{\operatorname{argmax}} P(X_1 = u_1 \cdots X_m = u_m \mid C = c)$$

- Maximum a-posteriori (MAP) classifier

$$C^{\text{predict}} = \underset{c}{\operatorname{argmax}} P(C = c \mid X_1 = u_1 \cdots X_m = u_m)$$

Parametric Classification

Discriminant (take the max over i):

$$g_i(x) = p(x | C_i)P(C_i)$$

or equivalently

$$g_i(x) = \log p(x | C_i) + \log P(C_i)$$

If we assume $p(x|C_i)$ are Gaussian:

$$p(x | C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right]$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x - \mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

- Given the sample $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$

1-D data $x \in \mathfrak{R}$

$$r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_j, j \neq i \end{cases}$$

Class indicators

- ML estimates of priors are

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

Observed
frequency

Observed
mean

Observed
variance

- Discriminant becomes

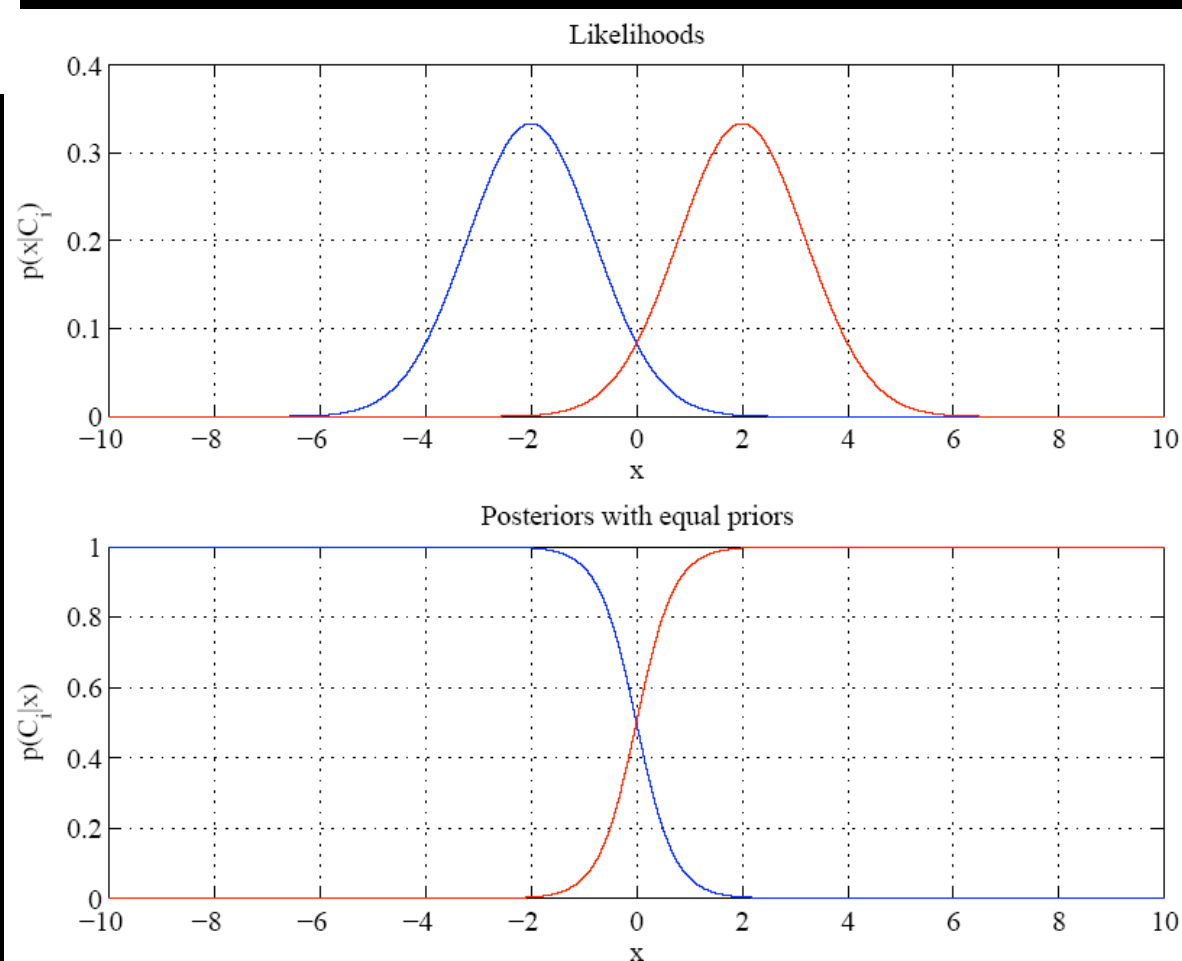
$$g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

Example: 2 classes (same variance)

Assume both classes
have the same prior

Likelihood =
Gaussian with
mean, std dev

Posterior =
discriminant $g(x)$
normalized by $P(x)$

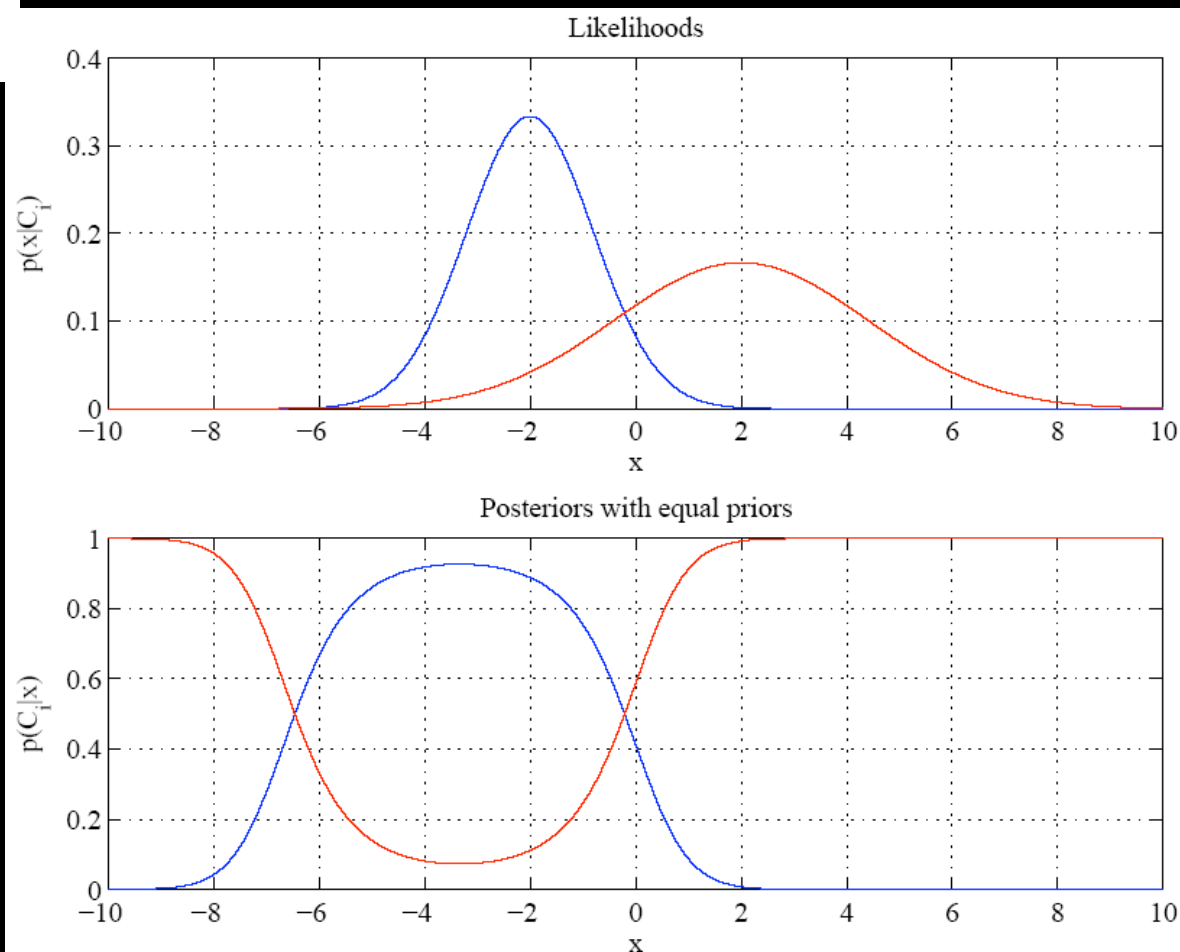


Example: 2 classes (diff variance)

Assume both classes
have the same prior

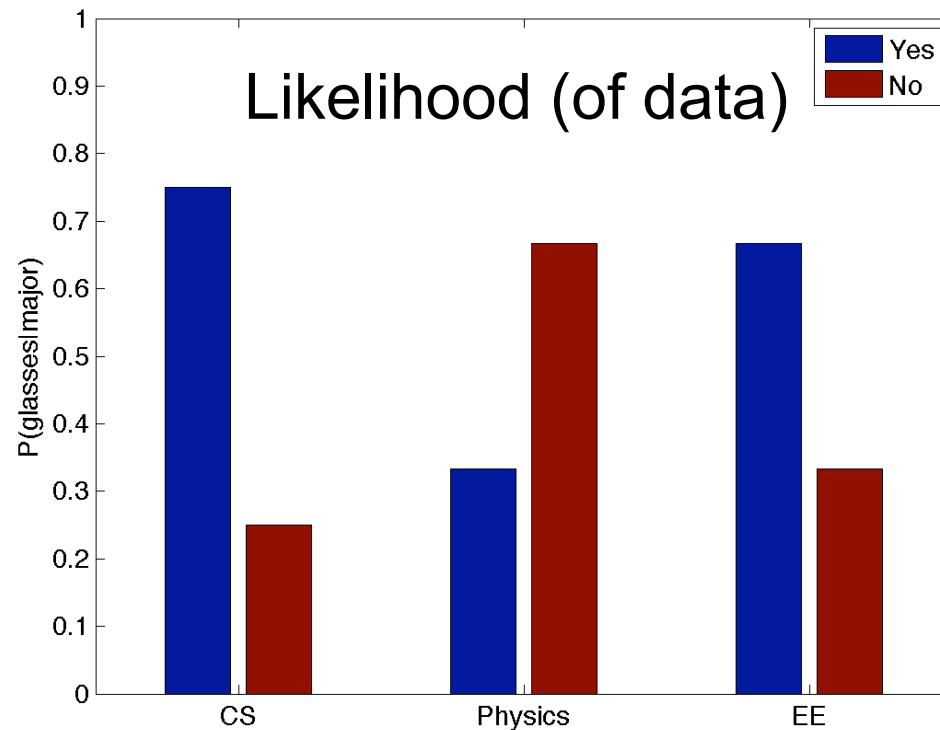
Likelihood =
Gaussian with
mean, std dev

Posterior =
discriminant $g(x)$
normalized by $P(x)$



Example: Predicting student's major

- From HW 3: Use “glasses” feature: yes or no
 - Discrete distribution

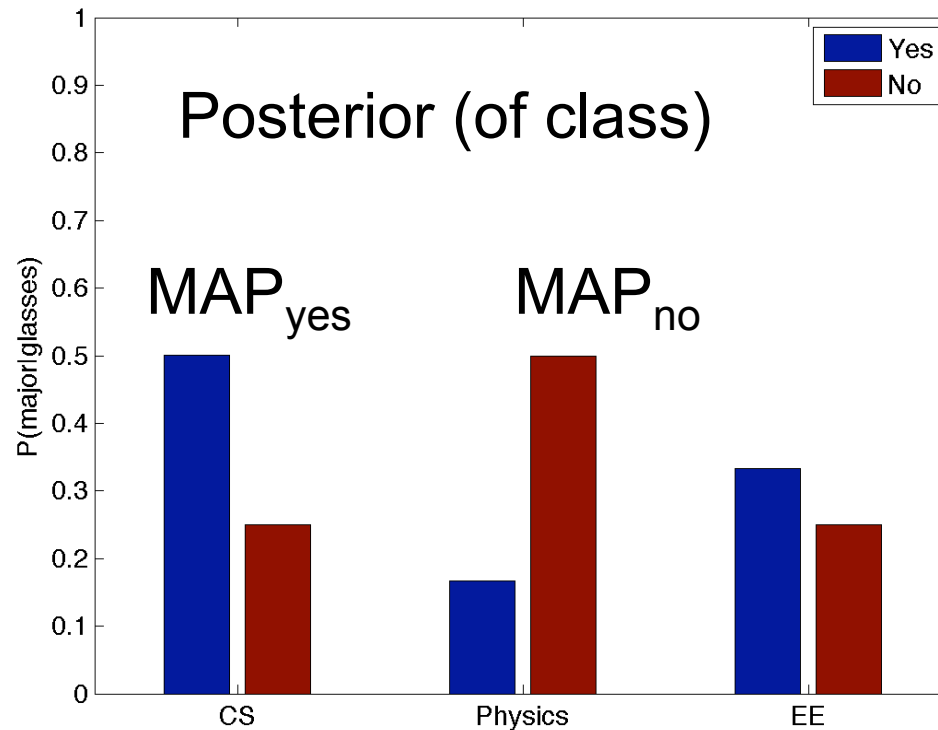


Example: Predicting student's major

Bayes: $P(\text{major}|\text{glasses}) = P(\text{glasses}|\text{major}) P(\text{major}) / P(\text{glasses})$

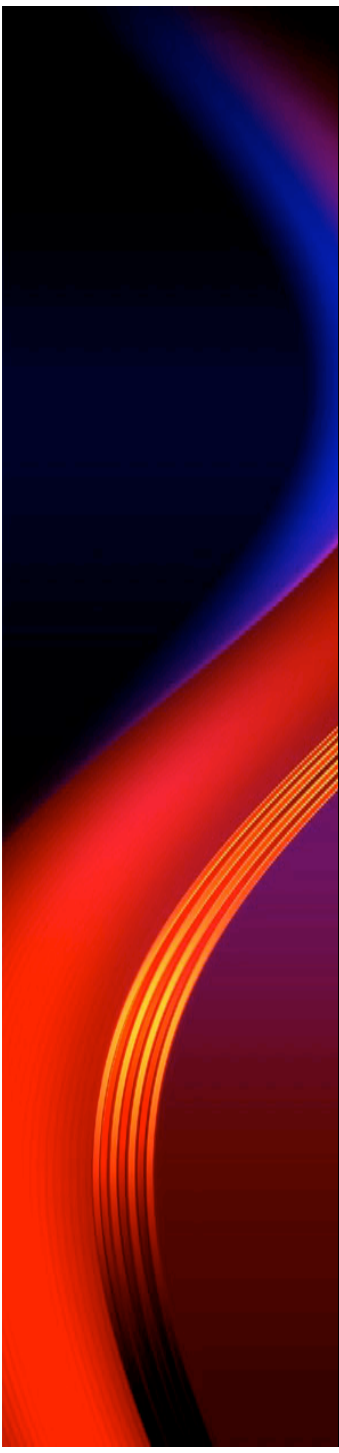
Priors: $P(\text{CS}) = 0.4$, $P(\text{Physics}) = 0.3$, $P(\text{EE}) = 0.3$

Probability of data: $P(\text{yes}) = 0.6$, $P(\text{no}) = 0.4$



Summary: Key Points for Today

- Parametric methods
 - Data comes from distribution
 - Bernoulli, Gaussian, and their parameters
 - How good is a parameter estimate? (bias, variance)
- Bayes estimation
 - ML: use the data
 - MAP: use the prior and the data
 - Bayes estimator: integrated estimate (weighted)
- Parametric classification
 - Maximize the posterior probability



Next Time

- Clustering!
(read Ch. 7.1-7.4, 7.8... optional)
- No reading questions