

CSCI 5622 Machine Learning

ML Evaluating Hypotheses

DATE	READ	DUE
Today, Sept 9	6	Prelim Proposal
Mon, Sept 14	See Web	---
Wed, Sept 16	4	Full Proposal

www.RodneyNielsen.com/teaching/CSCI5622-F09/

Instructor: Rodney Nielsen

Assistant Professor Adjunct, CU Dept. of Computer Science

Research Assistant Professor, DU, Dept. of Electrical & Computer Engr.

Research Scientist, Boulder Language Technologies

ML Hypothesis Evaluation

- **Limited sample versus population**
- **Hypothesis comparison**
- **Training and testing with limited data**
- **Distribution assumptions**

ML Confidence Intervals

- Confidence interval for binomial distribution

$$\mu \pm z_{1-\alpha} \sigma$$

$$error_s(h) \pm z_{1-\alpha} \sqrt{\frac{error_s(h)(1 - error_s(h))}{n}}$$

ML Limited Data Testing Procedure

- Partition D_0 into k disjoint subsets T_1, T_2, \dots, T_k of equal size ≥ 30

- For i from 1 to k , do

$$S_i \leftarrow \{D_0 - T_i\}$$

$$h_A \leftarrow L_A(S_i)$$

$$h_B \leftarrow L_B(S_i)$$

$$\delta_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$$

- Return

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^k \delta_i$$

ML Confidence Interval

- Paired tests, in this case the paired t -test, lead to tighter confidence interval
- $k-1$ is the degrees of freedom

$$\bar{\delta} \pm t_{1-\alpha, k-1} S_{\bar{\delta}}$$

$$S_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}$$

- Use McNemar if possible; see Deitterich, '96

ML **Bayesian Learning**

- **Probabilistic approach to inference**
- **Assumes instances are governed by distribution conditioned on the label**
- **Basis for learning algorithms that directly manipulate probabilities**
- **Framework for analyzing other ML algorithms**
- **Among best for some problems**

ML Bayesian Features

- **Adjust probability that a given hypothesis is correct vs. eliminating hypotheses**
- **Can incorporate prior knowledge (probabilities)**
- **Make probabilistic estimates**
- **Can classify according to multiple weighted h**
- **When computationally intractable, can still provide standard**

ML **Practical Difficulties**

- **Require initial knowledge of many probabilities**
 - If unknown, estimated using background knowledge, previous data, or assumptions about the underlying distribution
- **Significant computational cost to compute the Bayes optimal hypothesis**
 - Depends on number of hypotheses (often large)

Basic Probability Rules

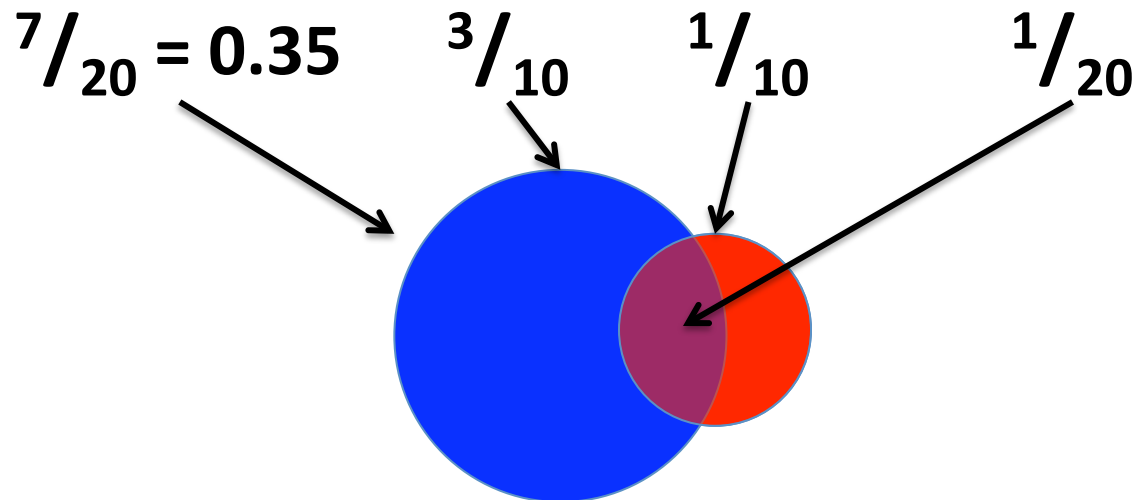
- **Sum rule**
 - $P(\text{Glasses}) = 0.3$
 - $P(\text{Sweater}) = 0.1$
 - $P(\text{Glasses} \vee \text{Sweater}) = ?$
 - $P = 0.0$ • $0.0 \leq P \leq 1.0$
 - $P = 0.03$ • $0.0 \leq P \leq 0.4$
 - $P = 0.1$ • $0.0 \leq P \leq 0.3$
 - $P = 0.2$ • $0.0 \leq P \leq 0.1$
 - $P = 0.3$ • $0.0 \leq P \leq 0.03$
 - $P = 0.4$ • $0.1 \leq P \leq 0.4$
 - $P = 1.0$ • $0.1 \leq P \leq 0.3$
 - $x \leq P \leq y$ • $0.3 \leq P \leq 0.4$

$$P(G \vee S) = ?$$

Basic Probability Rules

- **Sum rule**
 - $P(\text{Glasses}) = 0.3$
 - $P(\text{Sweater}) = 0.1$
 - $P(\text{Glasses} \vee \text{Sweater}) = ?$

$$P(G \vee S) = P(G) + P(S) - P(G \wedge S)$$



Basic Probability Rules

- **Product rule**
 - $P(\text{Glasses}) = 0.3$
 - $P(\text{Sweater}) = 0.1$
 - $P(\text{Glasses} \wedge \text{Sweater}) = ?$
 - $P = 0.0$
 - $P = 0.03$
 - $P = 0.1$
 - $P = 0.2$
 - $P = 0.3$
 - $P = 0.4$
 - $P = 1.0$
 - $x \leq P \leq y$

- $0.0 \leq P \leq 1.0$
- $0.0 \leq P \leq 0.4$
- $0.0 \leq P \leq 0.3$
- $0.0 \leq P \leq 0.1$
- $0.0 \leq P \leq 0.03$
- $0.1 \leq P \leq 0.4$
- $0.1 \leq P \leq 0.3$
- $0.3 \leq P \leq 0.4$

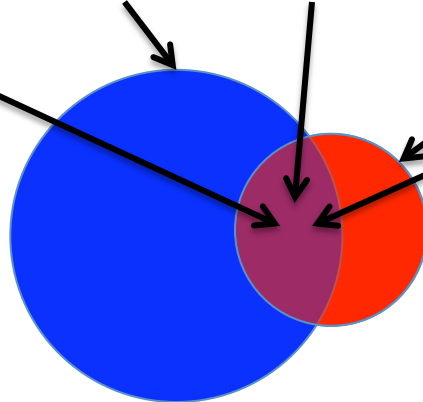
$$P(G \wedge S) = ?$$

Basic Probability Rules

- **Product rule**
 - $P(\text{Glasses}) = 0.3$
 - $P(\text{Sweater}) = 0.1$
 - $P(\text{Glasses} \wedge \text{Sweater}) = ?$

$$P(G \wedge S) = P(G)P(S|G) = P(S)P(G|S)$$

$$\frac{1}{20} = 0.05 \quad \frac{3}{10} \quad \frac{1}{6} \quad \frac{1}{10} \quad \frac{1}{2}$$



Bayes Theorem

- **Product rule**

$$P(A \wedge B) = P(B)P(A|B) = P(A)P(B|A)$$

- **Bayes Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Basic Probability Rules

- **Bayes Theorem**

- $P(\text{Glasses}) = 0.3$

- $P(\text{Sweater}) = 0.1$

- $P(\text{Glasses} \mid \text{Sweater}) = ?$

- | | |
|---------------------|--------------------------|
| • $P = 0.0$ | • $0.0 \leq P \leq 1.0$ |
| • $P = 0.03$ | • $0.0 \leq P \leq 0.4$ |
| • $P = 0.1$ | • $0.0 \leq P \leq 0.3$ |
| • $P = 0.2$ | • $0.0 \leq P \leq 0.1$ |
| • $P = 0.3$ | • $0.0 \leq P \leq 0.03$ |
| • $P = 0.4$ | • $0.1 \leq P \leq 0.4$ |
| • $P = 1.0$ | • $0.1 \leq P \leq 0.3$ |
| • $x \leq P \leq y$ | • $0.3 \leq P \leq 0.4$ |

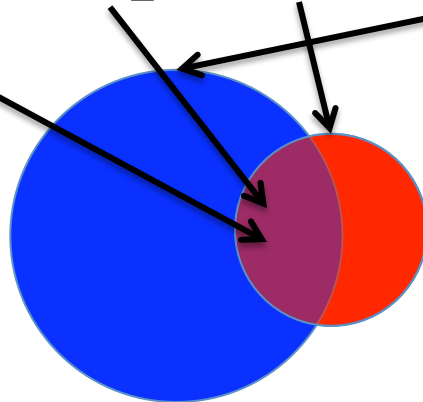
$$P(G|S) = ?$$

Basic Probability Rules

- **Bayes Theorem**
 - $P(\text{Glasses}) = 0.3$
 - $P(\text{Sweater}) = 0.1$
 - $P(\text{Glasses} \mid \text{Sweater}) = ?$

$$P(S|G) = P(G|S)P(S)/P(G)$$

$$\frac{1}{6} = 0.167 \quad \frac{1}{2} \quad \frac{1}{10} \quad \frac{3}{10}$$



Basic Probability Rules

- **Bayes Theorem**

- $P(\text{Glasses}) = 0.3$

- $P(\text{Sweater}) = 0.1$

- $P(\text{Sweater} \mid \text{Glasses}) = ?$

- | | |
|---------------------|--------------------------|
| • $P = 0.0$ | • $0.0 \leq P \leq 1.0$ |
| • $P = 0.03$ | • $0.0 \leq P \leq 0.4$ |
| • $P = 0.1$ | • $0.0 \leq P \leq 0.3$ |
| • $P = 0.2$ | • $0.0 \leq P \leq 0.1$ |
| • $P = 0.3$ | • $0.0 \leq P \leq 0.33$ |
| • $P = 0.4$ | • $0.1 \leq P \leq 0.4$ |
| • $P = 1.0$ | • $0.1 \leq P \leq 0.3$ |
| • $x \leq P \leq y$ | • $0.3 \leq P \leq 0.4$ |

$$P(S \mid G) = ?$$

Basic Probability Rules

- Theorem of total probability

- $P(\text{Glasses} \mid \text{Sweater}) = 0.2$

- $P(\text{Glasses} \mid \text{Shirt}) = 0.1$

- $P(\text{Glasses} \mid \text{WetSuit}) = 0.0$

- $P(\text{Glasses} \mid \text{Other}) = 0.1$

- $P = 0.0$
 - $0.0 \leq P \leq 1.0$

- $P = 0.1$
 - $0.0 \leq P \leq 0.4$

- $P = 0.2$
 - $0.0 \leq P \leq 0.2$

- $P = 0.3$
 - $0.0 \leq P \leq 0.1$

- $P = 0.4$
 - $0.1 \leq P \leq 0.4$

- $P = 1.0$
 - $0.1 \leq P \leq 0.2$

- $x \leq P \leq y$
 - $0.2 \leq P \leq 0.4$

$$P(G) = ?$$

Basic Probability Rules

- **Theorem of total probability**

- $P(\text{Glasses} \mid \text{Sweater}) = 0.4$; $P(\text{Sweater}) = 0.1$

- $P(\text{Glasses} \mid \text{Shirt}) = 0.2$; $P(\text{Shirt}) = 0.6$

- $P(\text{Glasses} \mid \text{WetSuit}) = 0.0$; $P(\text{WetSuit}) = 0.1$

- $P(\text{Glasses} \mid \text{Other}) = 0.1$; $P(\text{Other}) = 0.2$

$$P(x \wedge y) = 0; x \neq y; x, y \in \{\text{Sweater}, \text{Shirt}, \text{WetSuit}, \text{Other}\}$$

$$\sum_{x \in \{\text{Sweater}, \text{Shirt}, \text{WetSuit}, \text{Other}\}} P(x) = 1.0$$

$$P(\text{Glasses}) = \sum_{x \in \{\text{Sweater}, \text{Shirt}, \text{WetSuit}, \text{Other}\}} P(\text{Glasses} \mid x) P(x)$$

$$P(\text{Glasses}) = 0.4 \cdot 0.1 + 0.2 \cdot 0.6 + 0.0 \cdot 0.1 + 0.1 \cdot 0.2 = 0.18$$

Bayes Theorem

- Find the most likely hypothesis given:
 - The prior probability of h , the probability of the data given h , and the observed data

$$P(h|\mathbf{D}) = \frac{P(\mathbf{D}|h)P(h)}{P(\mathbf{D})}$$

- $P(h)$ is the prior probability of h
- $P(D)$ is the prior probability of D
- $P(D|h)$: the prob. of D given h (likelihood)
- $P(h|D)$: posterior prob. of h given the data

Bayes Learning

- **Maximum a posteriori hypothesis**

$$\begin{aligned}h_{MAP} &\equiv \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h)\end{aligned}$$

- **Maximum likelihood hypothesis**

$$h_{ML} \equiv \arg \max_{h \in H} P(D|h)$$

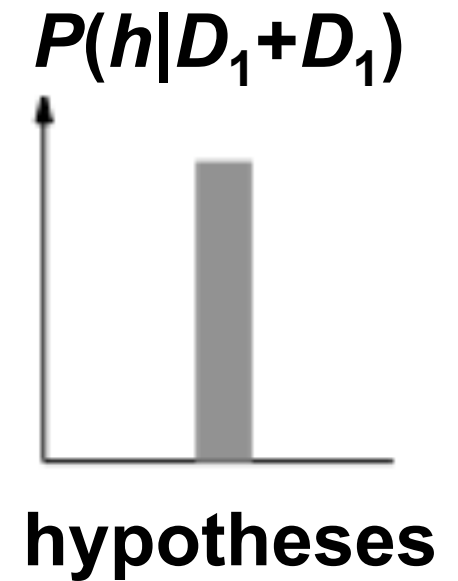
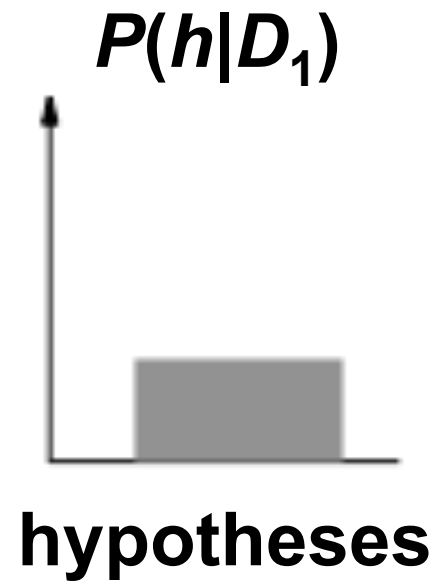
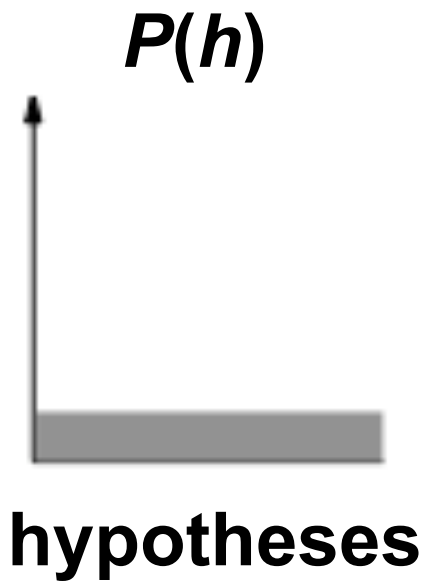
Bayes Learning

- **Brute-force MAP concept learning**

$$h_{MAP} \equiv \arg \max_{h \in H} P(D|h)P(h)$$

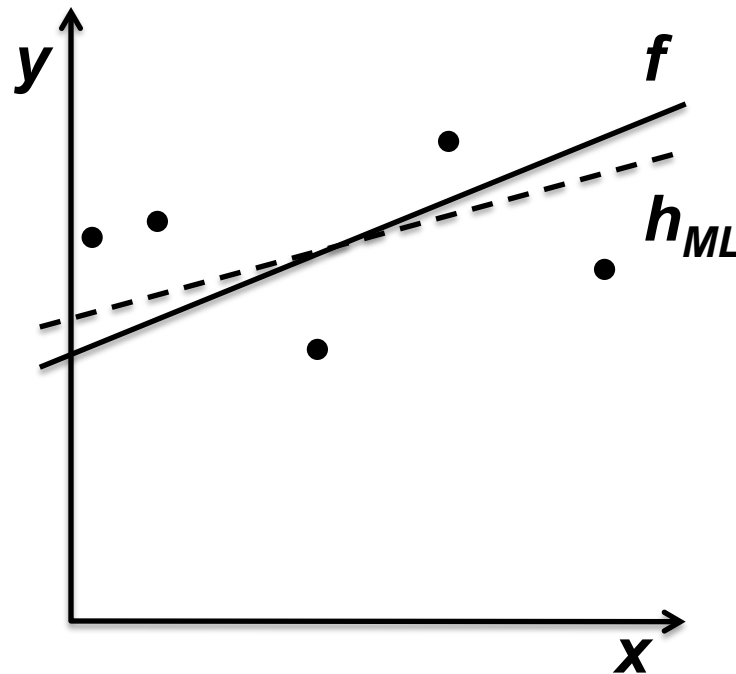
- Choose $P(D|h)$ and $P(h)$ according to prior knowledge about the learning problem
- If no prior knowledge, could assume $P(h) = 1/|H|$
- If assume no noise and target concept is in H , then $P(D|h) = 1$ if consistent and 0 otherwise \rightarrow
 $P(h|D) = 1/|V_{S_{H,D}}|$ if consistent and 0 otherwise

ML Evolution of Posteriors $P(h | D)$



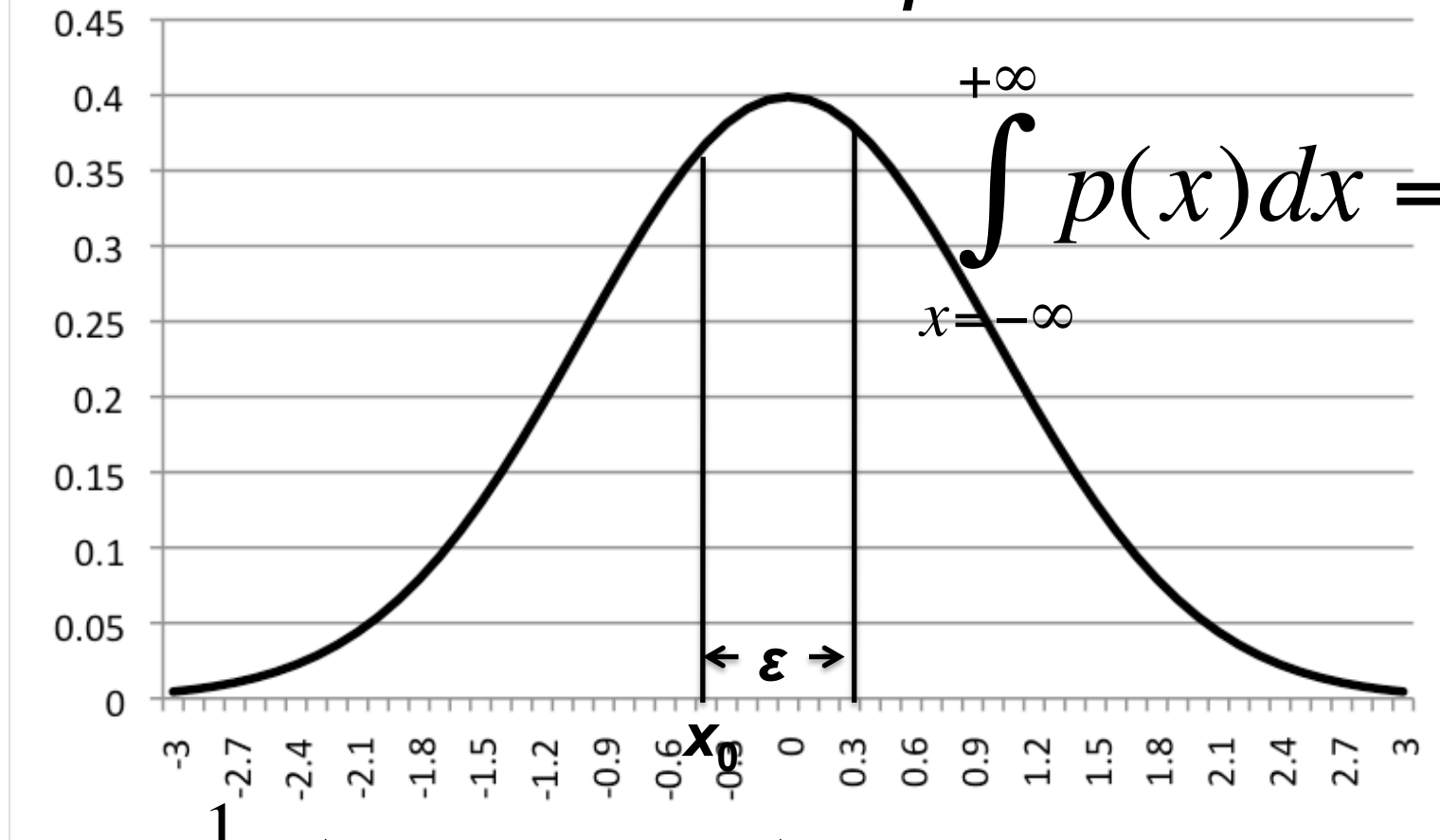
ML h_{ML} & Least Error² Hypothesis

- Under specific assumptions, the Least Squared Error hypothesis is a Maximum Likelihood hypothesis



ML Probability Density Functions

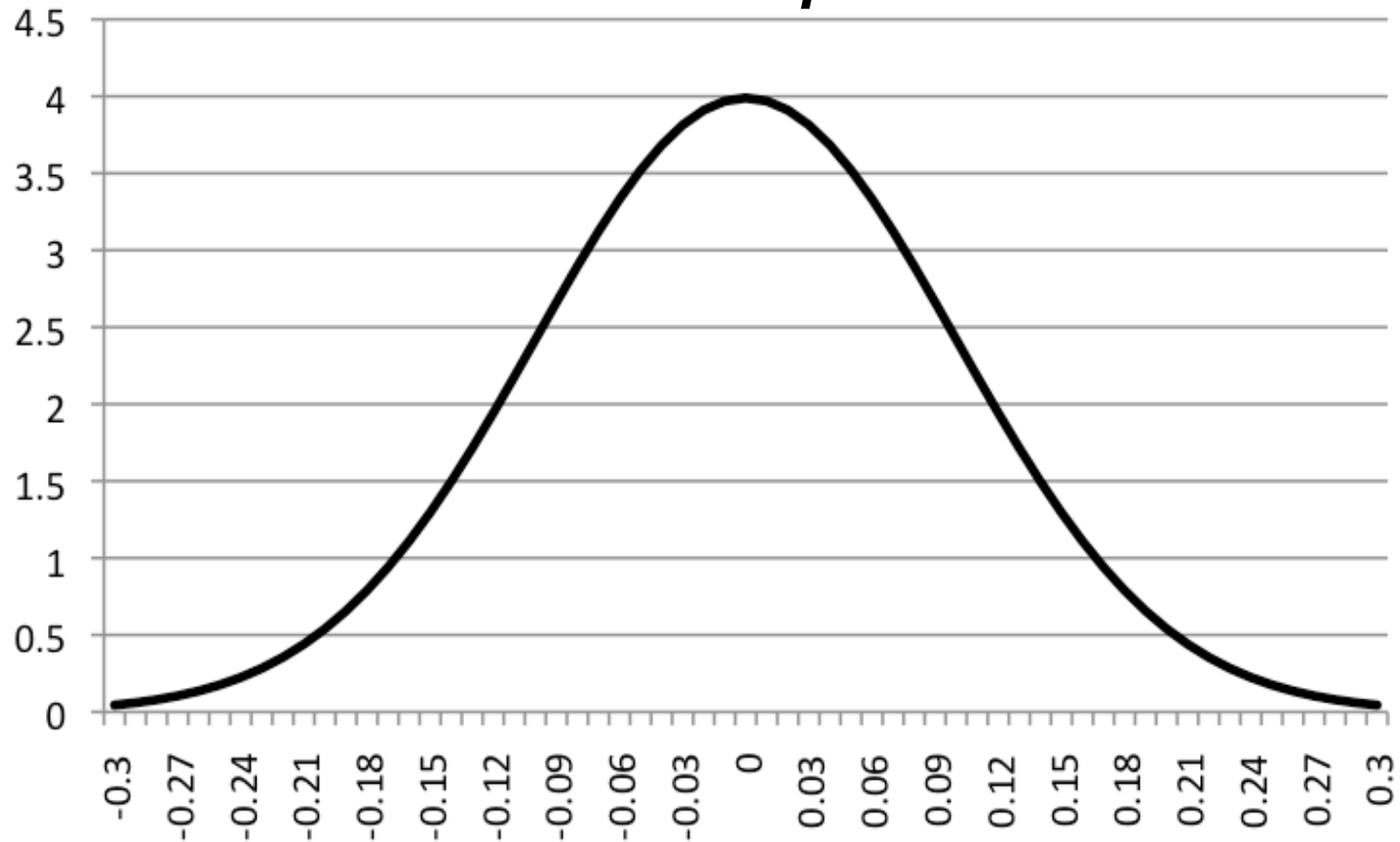
Normal distribution with $\mu = 0$ and $\sigma = 1$



$$p(x_0) = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} P(x_0 \leq x < x_0 + \epsilon)$$

ML Probability Density Function

Normal distribution with $\mu = 0$ and $\sigma = 0.1$



ML h_{ML} & Least Error² Hypothesis

$$h_{ML} = \arg \max_{h \in H} p(D|h)$$

$$= \arg \max_{h \in H} \prod_{i=1}^n p(d_i|h)$$

$$= \arg \max_{h \in H} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2}$$

$$h_{ML} = \arg \max_{h \in H} \ln \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2}$$

$$= \arg \max_{h \in H} \sum_{i=1}^n \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} (d_i - h(x_i))^2$$

$$h_{ML} = \arg \min_{h \in H} \sum_{i=1}^n (d_i - h(x_i))^2$$

Assume noise follows a Normal distribution

ML h_{ML} for Predicting Probabilities

- We may return to this when we discuss neural networks

M Min Description Length, Occam & Bayes

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

$$= \arg \max_{h \in H} \log_2 P(D|h) + \log_2 P(h)$$

$$= \arg \min_{h \in H} -\log_2 P(D|h) - \log_2 P(h)$$

$$= \arg \min_{h \in H} L_{C_{D|h}}(D|h) + L_{C_h}(h)$$

$$h_{MDL} = \arg \min_{h \in H} L_{C_1}(D|h) + L_{C_2}(h)$$

ML Bayes Optimal Classifier

- h_{MAP} : The most probable hypothesis given the training data
- Often more important: The most probably classification of a new instance given the training data
 - Ex: Three friends give you different opinions about a ML algorithm. You trust one about 33% more than the others, but the others agree with each other. What do you decide?
 $0.4:O_1$ vs. $2*0.3:O_2$

Bayes Optimal Classifier

- **The most probably classification of a new instance given the training data**

$$P(y_j|D) = \sum_{h_i \in H} P(y_j|h_i)P(h_i|D)$$

$$\hat{y} = \operatorname{argmax}_{y_j \in C} \sum_{h_i \in H} P(y_j|h_i)P(h_i|D)$$

- **No other classification using the same hypothesis space and prior knowledge can outperform this method on average**

ML **Naïve Bayes Classifier**

- **Practical, easy to implement**
- **Sometimes results as good as other classifiers, usually not**

$$\hat{y}_{MAP} = \arg \max_{y_j \in C} P(y_j | \mathbf{x}); \quad \mathbf{x} = \langle x_1, x_2 \cdots x_d \rangle$$

$$\hat{y}_{MAP} = \arg \max_{y_j \in C} P(x_1, x_2 \cdots x_d | y_j) P(y_j) / P(\mathbf{x})$$

- **Usually impossible to estimate $P(x_1, x_2 \cdots x_d | y_j)$**

Naïve Bayes Classifier

- Assume independence of attributes given the class

$$P(x_1, x_2 \cdots x_d | y_j) = \prod_i P(x_i | y_j)$$

$$\hat{y}_{NB} = \arg \max_{y_j \in C} P(y_j) \prod_i P(x_i | y_j)$$

- Now we need only estimate the $P(y_j)$ and the $P(x_i | y_j)$ rather than all possible $P(x_1, x_2 \cdots x_d, | y_j)$

ML Estimating Probabilities

- Fewer estimates for Naïve Bayes, but some are still problematic
- Use the *m*-estimate

$$\frac{n_c + mp}{n + m}$$

- Often set $p=1/k$ and $m=k$

$$\frac{n_c + 1}{n + k}$$

ML

Are you a skier

$$P(\text{Skier}) = 12$$

$$P(\text{Nonskier}) = 12$$

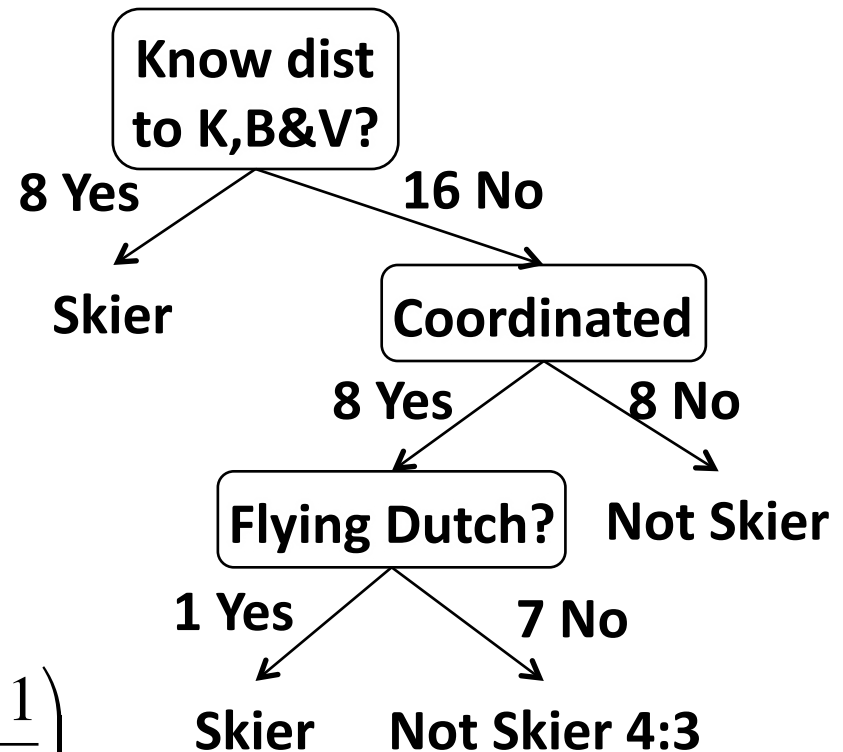
$$P(x_1 = \text{yes} | \text{S}) = \frac{8}{12}; P(x_1 | \text{N}) = \frac{0}{12}$$

$$P(x_2 = \text{yes} | \text{S}) = \frac{9}{12}; P(x_2 | \text{N}) = \frac{4}{12}$$

$$P(x_3 = \text{yes} | \text{S}) = \frac{1}{12}; P(x_3 | \text{N}) = \frac{0}{12}$$

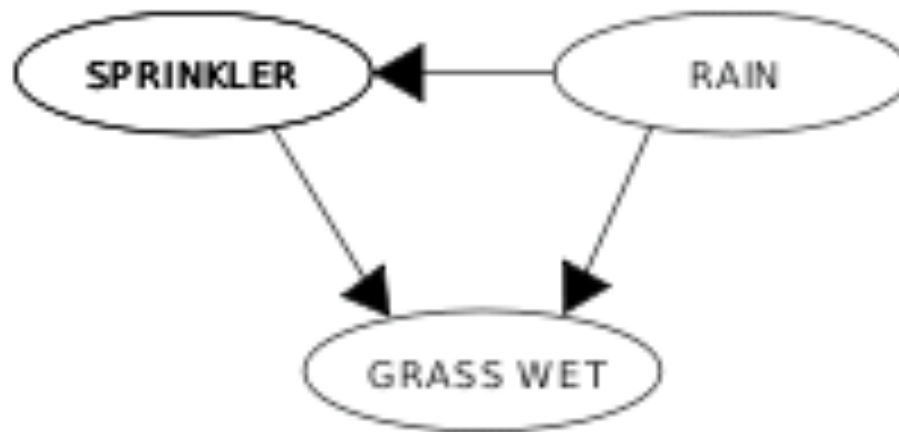
$$\hat{y}_{\text{Skier}, \text{NB}}(\mathbf{x} = \text{yes}, \text{no}, \text{no}) = \frac{1}{2} \left(\frac{8}{12} \cdot \frac{3}{12} \cdot \frac{11}{12} \right)$$

$$\hat{y}_{\text{Nonskier}, \text{NB}}(\mathbf{x} = \text{yes}, \text{no}, \text{no}) = \frac{1}{2} \left(\frac{0}{12} \cdot \frac{8}{12} \cdot \frac{12}{12} \right)$$



ML Bayesian Belief Networks

		SPRINKLER	
		T	F
RAIN	F	0.4	0.6
	T	0.01	0.99



		RAIN	
		T	F
		0.2	0.8

		GRASS WET	
		T	F
SPRINKLER	RAIN	0.0	1.0
F	F	0.8	0.2
F	T	0.9	0.1
T	F	0.99	0.01
T	T		