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

Hypothesis Evaluation

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

Confidence Intervals

- **Confidence interval for binomial distribution**

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

$$error_{\mathbf{S}}(h) \pm z_{1-\alpha}\sqrt{\frac{error_{\mathbf{S}}(h)\left(1-error_{\mathbf{S}}(h)\right)}{n}}$$

ML Limited Data Testing Procedure

- **Partition D_0 into k disjoint subsets $T_1, T_2, \ldots, T_k$ of equal size >= 30**
- **For i from 1 to k, do**

$$\mathbf{S}_i \leftarrow \{D_0 - T_i\}$$

$$h_A \leftarrow L_A(\mathbf{S}_i)$$

$$h_B \leftarrow L_B(\mathbf{S}_i)$$

$$\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$$

- **Return**

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

Confidence Interval

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

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

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

- **Use McNemar if possible; see Deitterich, '96**

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

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**

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

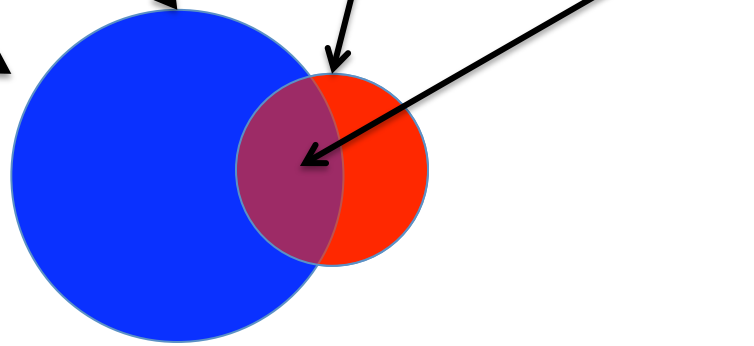
$P(G \vee S) = ?$

Basic Probability Rules

- **Sum rule**
  - ***P*(Glasses) = 0.3**
  - ***P*(Sweater) = 0.1**
  - ***P*(Glasses ∨ Sweater) = ?**

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

$^{7}/_{20} = 0.35$ $^{3}/_{10}$ $^{1}/_{10}$ $^{1}/_{20}$



Basic Probability Rules

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

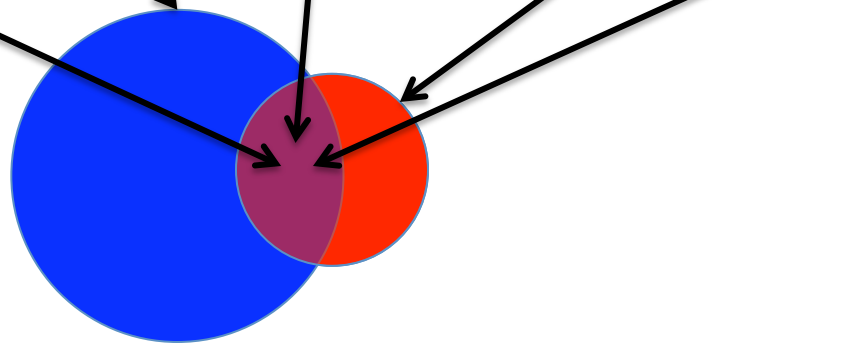
$P(G \wedge S) = ?$

Basic Probability Rules

- **Product rule**
  - ***P*(Glasses) = 0.3**
  - ***P*(Sweater) = 0.1**
  - ***P*(Glasses ∧ Sweater) = ?**

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

$^1/_{20} = 0.05$ $^3/_{10}$ $^1/_6$ $^1/_{10}$ $^1/_2$



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**
  - $P(\text{Glasses}) = 0.3$
  - $P(\text{Sweater}) = 0.1$
  - $P(\text{Glasses} \mid \text{Sweater}) = ?$
    - $P = 0.0$
    - $P = 0.03$
    - $P = 0.1$
    - $P = 0.2$
    - $P = 0.3$
    - $P = 0.4$
    - $P = 1.0$
    - $x \le P \le y$
    - $0.0 \le P \le 1.0$
    - $0.0 \le P \le 0.4$
    - $0.0 \le P \le 0.3$
    - $0.0 \le P \le 0.1$
    - $0.0 \le P \le 0.03$
    - $0.1 \le P \le 0.4$
    - $0.1 \le P \le 0.3$
    - $0.3 \le P \le 0.4$

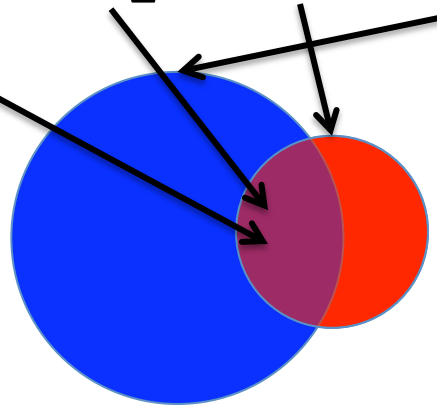
$P(G|S) = ?$

Basic Probability Rules

- **Bayes Theorem**
  - ***P*(Glasses) = 0.3**
  - ***P*(Sweater) = 0.1**
  - ***P*(Glasses | Sweater) = ?**

$$P(\mathrm{S}|\mathrm{G}) = P(\mathrm{G}|\mathrm{S})P(\mathrm{S})/P(\mathrm{G})$$

$^1/_6 = 0.167$ $^1/_2$ $^1/_{10}$ $^3/_{10}$



ML

Basic Probability Rules

- **Bayes Theorem**
  - $P(\text{Glasses}) = 0.3$
  - $P(\text{Sweater}) = 0.1$
  - $P(\text{Sweater} \mid \text{Glasses}) = ?$
    - $P = 0.0$ — $0.0 \leq P \leq 1.0$
    - $P = 0.03$ — $0.0 \leq P \leq 0.4$
    - $P = 0.1$ — $0.0 \leq P \leq 0.3$
    - $P = 0.2$ — $0.0 \leq P \leq 0.1$
    - $P = 0.3$ — $0.0 \leq P \leq 0.33$
    - $P = 0.4$ — $0.1 \leq P \leq 0.4$
    - $P = 1.0$ — $0.1 \leq P \leq 0.3$
    - $x \leq P \leq y$ — $0.3 \leq P \leq 0.4$

$P(S|G) = ?$

ML

Basic Probability Rules

- **Theorem of total probability**
  - $P(\text{Glasses} \mid \text{Sweater}) = 0.2$
  - $P(\text{Glasses} \mid \text{Shirt}) = 0.1$
  - $P(\text{Glasses} \mid \text{WetSuit}) = 0.0$
  - $P(\text{Glasses} \mid \text{Other}) = 0.1$

• $P = 0.0$	• $0.0 \le P \le 1.0$
• $P = 0.1$	• $0.0 \le P \le 0.4$
• $P = 0.2$	• $0.0 \le P \le 0.2$
• $P = 0.3$	• $0.0 \le P \le 0.1$
• $P = 0.4$	• $0.1 \le P \le 0.4$
• $P = 1.0$	• $0.1 \le P \le 0.2$
• $x \le P \le y$	• $0.2 \le P \le 0.4$

$P(\text{G}) = ?$

Basic Probability Rules

- **Theorem of total probability**
  - ***P*(Glasses | Sweater) = 0.4; *P*(Sweater) = 0.1**
  - ***P*(Glasses | Shirt) = 0.2; *P*(Shirt) = 0.6**
  - ***P*(Glasses | WetSuit) = 0.0; *P*(WetSuit) = 0.1**
  - ***P*(Glasses | Other) = 0.1; *P*(Other) = 0.2**

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

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

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

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

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**

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

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

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

- **Maximum likelihood hypothesis**

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

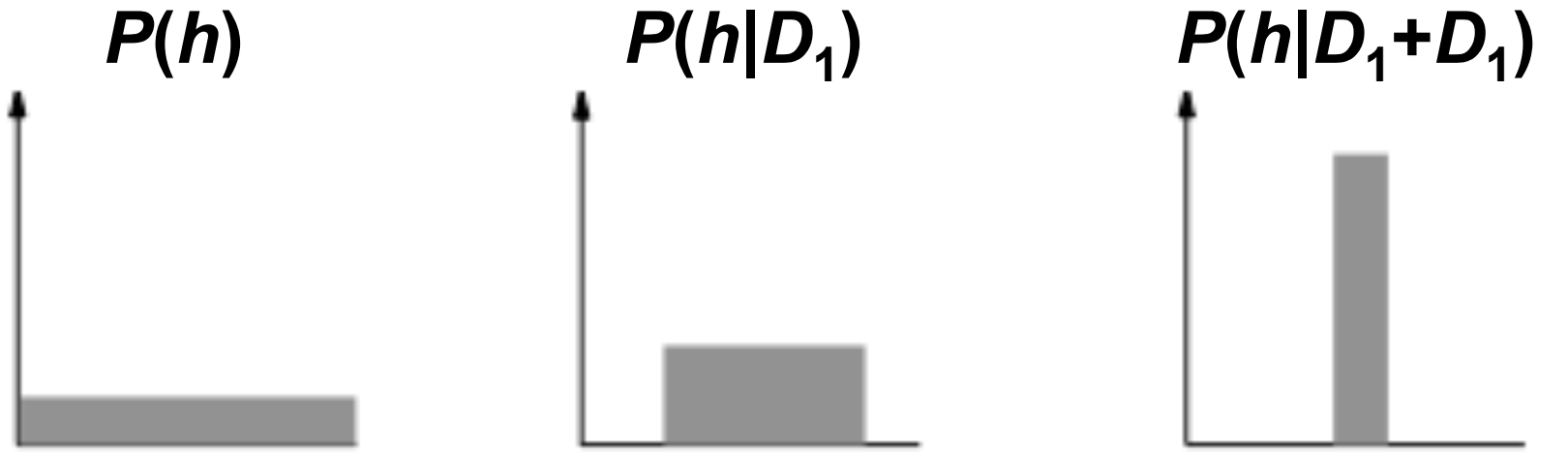
Bayes Learning

- **Brute-force MAP concept learning**

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

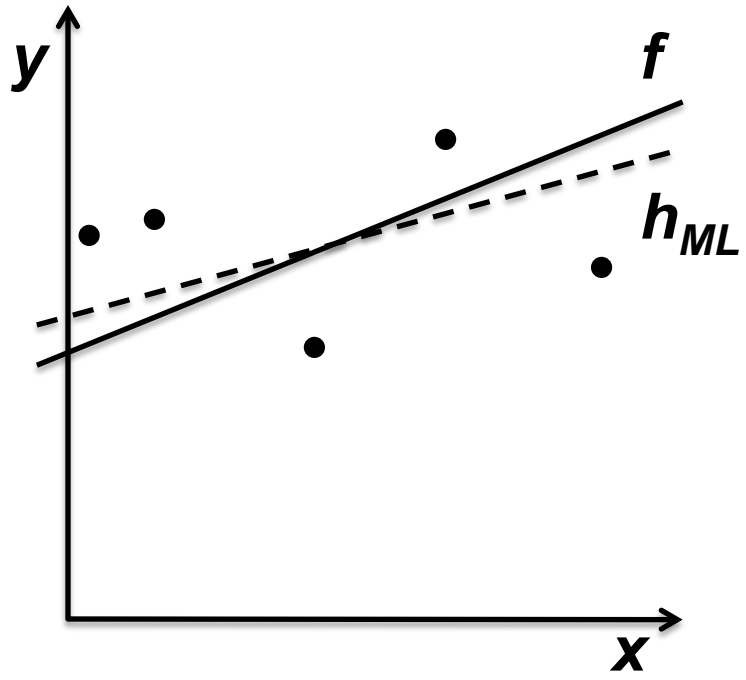
  - **Choose $P(D|h)$ and $P(h)$ according to prior knowledge about the learning problem**
  - **If no prior knowledge, could assume $P(h) = 1/|H|$**
  - **If assume no noise and target concept is in H, then $P(D|h) = 1$ if consistent and 0 otherwise → $P(h|D) = 1/|VS_{H,D}|$ if consistent and 0 otherwise**

Evolution of Posteriors $P(h|D)$

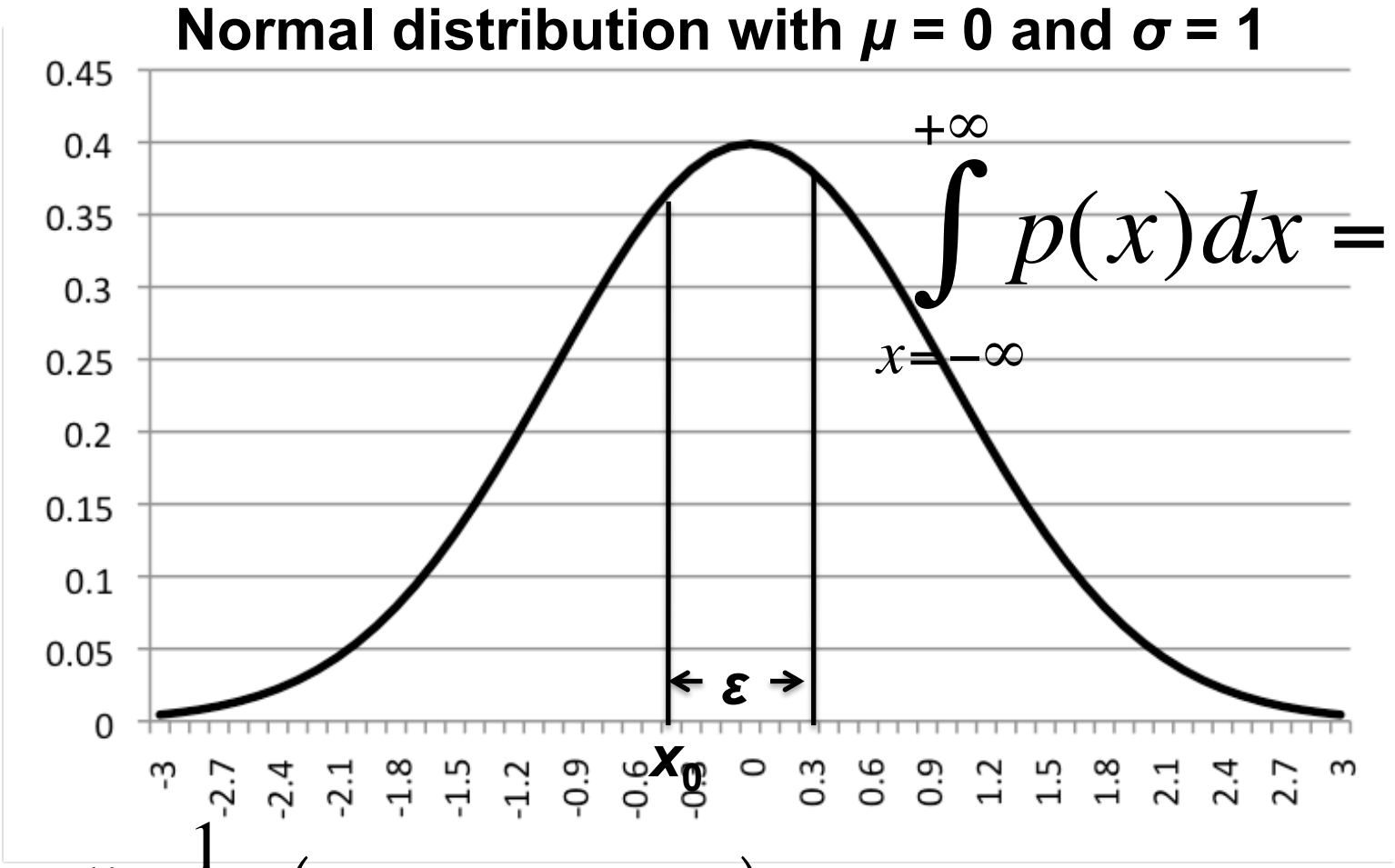


ML h_{ML} & Least Error^2 Hypothesis

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



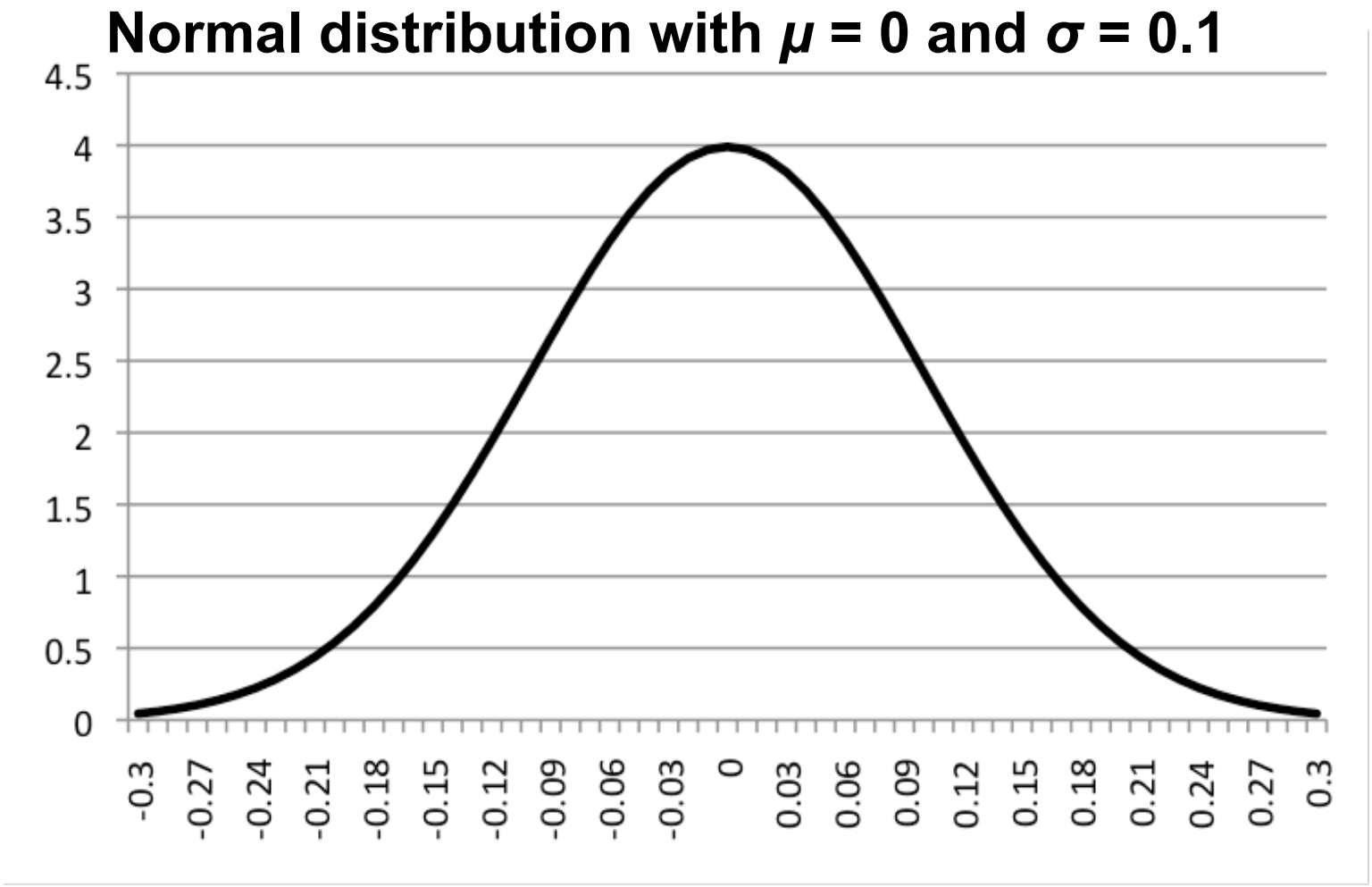
Probability Density Functions



$$\int_{x=-\infty}^{+\infty} p(x)dx = 1.0$$

$$p(x_0) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} P\left(x_0 \le x < x_0 + \varepsilon\right)$$

Probability Density Function



ML h_{ML} & Least Error² Hypothesis

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

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

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

Assume noise follows a Normal distribution

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

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

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

ML h_{ML} for Predicting Probabilities

- **We may return to this when we discuss neural networks**

ML Min Description Length, Occam & Bayes

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

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

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

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

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

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**

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

$$\hat{y} = \arg\max_{y_j \in C} \sum_{h_i \in H} P\left(y_j|h_i\right)P\left(h_i|D\right)$$

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

Naïve Bayes Classifier

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

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

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

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

Naïve Bayes Classifier

- **Assume independence of attributes given the class**

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

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

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

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***

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

Are you a skier

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

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

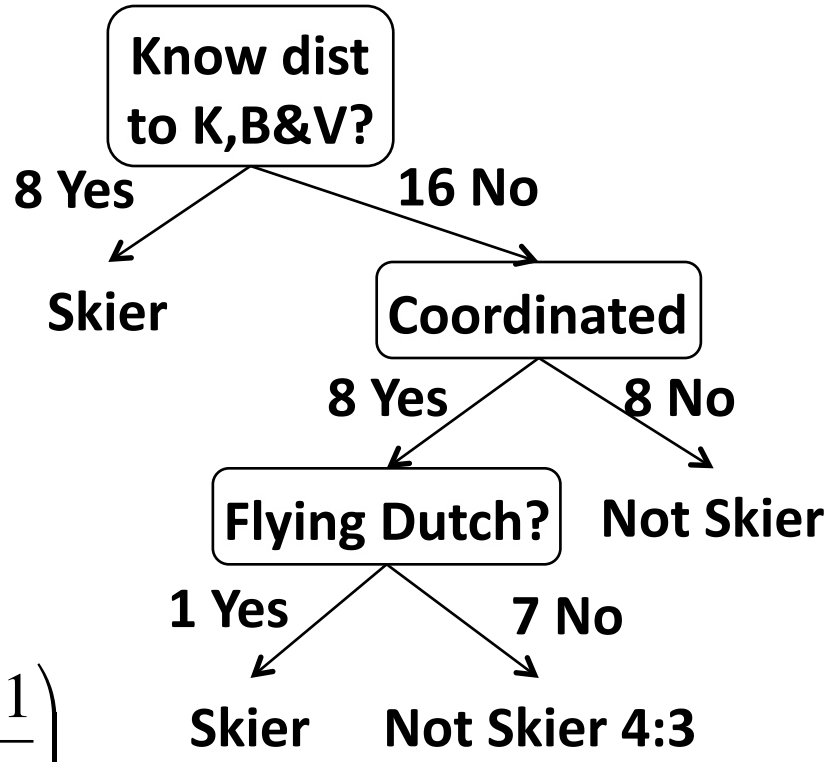
$$P(x_1 = yes|\text{S}) = \frac{8}{12}; P(x_1|\text{N}) = \frac{0}{12}$$

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

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

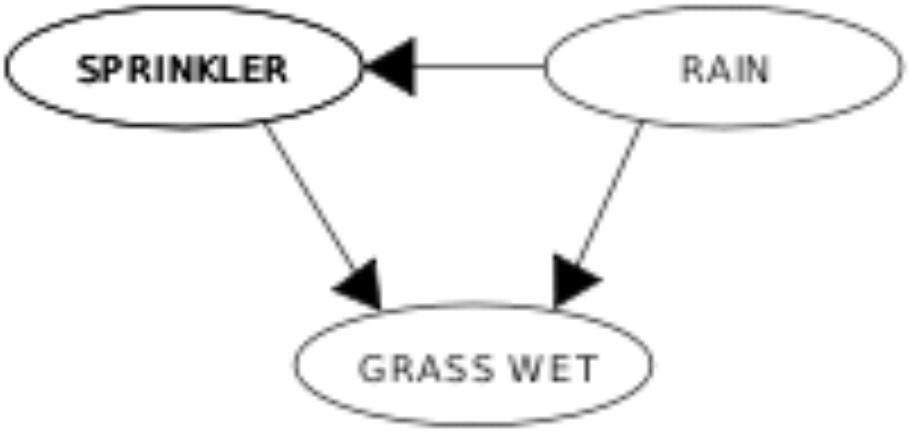
$$\hat{y}_{Skier,NB}(\mathbf{x} = \text{yes,no,no}) = \frac{1}{2}\left(\frac{8}{12} \cdot \frac{3}{12} \cdot \frac{11}{12}\right)$$

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



Bayesian Belief Networks

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



	RAIN	
T	F	
0.2	0.8	

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