

## Lecture 06 of 42

# Bayes's Theorem, MAP, and Maximum Likelihood Hypotheses

Monday, 04 February 2008

William H. Hsu

Department of Computing and Information Sciences, KSU

<http://www.kddresearch.org>

Readings:

Sections 6.1-6.5, Mitchell



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences

## Lecture Outline

- Read Sections 6.4, Han & Kamber (2006)
- Overview of Bayesian Learning
  - Framework: using probabilistic criteria to generate hypotheses of all kinds
  - Probability: foundations
- Bayes's Theorem
  - Definition of conditional (posterior) probability
  - Ramifications of Bayes's Theorem
    - Answering probabilistic queries
    - MAP hypotheses
- Generating Maximum A Posteriori (MAP) Hypotheses
- Generating Maximum Likelihood Hypotheses
- Rest of this Week: 6.4 – 6.5, Han & Kamber, Chapter 6 Mitchell
  - More Bayesian learning: MDL, BOC, Gibbs, Simple (Naïve) Bayes
  - Learning over text



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences

## Bayesian Learning

- **Framework: Interpretations of Probability [Cheeseman, 1985]**
  - Bayesian subjectivist view
    - A measure of an agent's belief in a proposition
    - Proposition denoted by random variable (sample space: range)
    - e.g.,  $Pr(\text{Outlook} = \text{Sunny}) = 0.8$
  - Frequentist view: probability is the *frequency of observations* of an event
  - Logicist view: probability is inferential evidence in favor of a proposition
- **Typical Applications**
  - HCI: learning natural language; intelligent displays; decision support
  - Approaches: prediction; sensor and data fusion (e.g., bioinformatics)
- **Prediction: Examples**
  - Measure *relevant parameters*: temperature, barometric pressure, wind speed
  - Make statement of the form  $Pr(\text{Tomorrow's-Weather} = \text{Rain}) = 0.5$
  - College admissions:  $Pr(\text{Acceptance}) \equiv p$ 
    - Plain beliefs: unconditional acceptance ( $p = 1$ ) or categorical rejection ( $p = 0$ )
    - Conditional beliefs: depends on reviewer (use probabilistic model)



## Two Roles for Bayesian Methods

- **Practical Learning Algorithms**
  - Naïve Bayes (*aka simple Bayes*)
  - Bayesian belief network (BBN) structure learning and parameter estimation
  - Combining prior knowledge (prior probabilities) with observed data
    - A way to incorporate background knowledge (BK), *aka domain knowledge*
    - Requires prior probabilities (e.g., annotated rules)
- **Useful Conceptual Framework**
  - Provides “gold standard” for evaluating other learning algorithms
    - Bayes Optimal Classifier (BOC)
    - Stochastic Bayesian learning: Markov chain Monte Carlo (MCMC)
  - Additional insight into Occam's Razor (MDL)



## Probabilistic Concepts versus Probabilistic Learning

- **Two Distinct Notions: Probabilistic Concepts, Probabilistic Learning**
- **Probabilistic Concepts**
  - Learned concept is a *function*,  $c: X \rightarrow [0, 1]$
  - $c(x)$ , the target value, denotes the probability that the label 1 (i.e., *True*) is assigned to  $x$
  - Previous learning theory is applicable (with some extensions)
- **Probabilistic (i.e., Bayesian) Learning**
  - Use of a probabilistic criterion in selecting a hypothesis  $h$ 
    - e.g., “most likely”  $h$  given observed data  $D$ : MAP hypothesis
    - e.g.,  $h$  for which  $D$  is “most likely”: max likelihood (ML) hypothesis
    - May or may not be stochastic (i.e., search process might still be deterministic)
  - NB:  $h$  can be deterministic (e.g., a Boolean function) or probabilistic



## Probability: Basic Definitions and Axioms

- **Sample Space ( $\Omega$ ): Range of a Random Variable  $X$**
- **Probability Measure  $Pr(\bullet)$** 
  - $\Omega$  denotes a range of “events”;  $X: \Omega$
  - **Probability  $Pr$ , or  $P$ , is a *measure over***
  - In a general sense,  $Pr(X = x \in \Omega)$  is a measure of belief in  $X = x$ 
    - $P(X = x) = 0$  or  $P(X = x) = 1$ : plain (aka categorical) beliefs (can’t be revised)
    - All other beliefs are subject to revision
- **Kolmogorov Axioms**
  - 1.  $\forall x \in \Omega . 0 \leq P(X = x) \leq 1$
  - 2.  $P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1$
  - 3.  $\forall X_1, X_2, \dots \ni i \neq j \Rightarrow X_i \wedge X_j = \emptyset .$
$$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$$
- **Joint Probability:  $P(X_1 \wedge X_2) \equiv$  Probability of the Joint Event  $X_1 \wedge X_2$**
- **Independence:  $P(X_1 \wedge X_2) = P(X_1) \cdot P(X_2)$**



## Bayes's Theorem

- Theorem

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

- $P(h)$   $\equiv$  Prior Probability of Hypothesis  $h$ 
  - Measures **initial beliefs** (BK) **before** any information is obtained (hence **prior**)
- $P(D)$   $\equiv$  Prior Probability of Training Data  $D$ 
  - Measures probability of obtaining sample  $D$  (i.e., expresses  $D$ )
- $P(h | D)$   $\equiv$  Probability of  $h$  Given  $D$ 
  - $|$  denotes **conditioning** - hence  $P(h | D)$  is a **conditional** (*aka posterior*) probability
- $P(D | h)$   $\equiv$  Probability of  $D$  Given  $h$ 
  - Measures probability of observing  $D$  given that  $h$  is correct (“**generative**” model)
- $P(h \wedge D)$   $\equiv$  **Joint** Probability of  $h$  and  $D$ 
  - Measures probability of observing  $D$  and of  $h$  being correct



## Choosing Hypotheses

- Bayes's Theorem

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

- MAP Hypothesis

- Generally want most probable hypothesis given the training data
- Define:  $\arg \max_{x \in \Omega} [f(x)]$   $\equiv$  the value of  $x$  in the sample space  $\Omega$  with the highest  $f(x)$
- **Maximum a posteriori** hypothesis,  $h_{MAP}$

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

- ML Hypothesis

- Assume that  $p(h_i) = p(h_j)$  for all pairs  $i, j$  (**uniform priors**, i.e.,  $P_H \sim$  Uniform)
- Can further simplify and choose the **maximum likelihood** hypothesis,  $h_{ML}$

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$



## Bayes's Theorem: Query Answering (QA)

- **Answering User Queries**
  - Suppose we want to perform intelligent inferences over a database *DB*
    - Scenario 1: *DB* contains records (instances), some “labeled” with answers
    - Scenario 2: *DB* contains probabilities (annotations) over propositions
  - QA: an application of probabilistic inference
- **QA Using Prior and Conditional Probabilities: Example**
  - Query: *Does patient have cancer or not?*
  - Suppose: patient takes a lab test and result comes back positive
    - Correct + result in only 98% of the cases in which disease is actually present
    - Correct - result in only 97% of the cases in which disease is not present
    - Only 0.008 of the entire population has this cancer
  - $\alpha \equiv P(\text{false negative for } H_0 \equiv \text{Cancer}) = 0.02$  (NB: for 1-point sample)
  - $\beta \equiv P(\text{false positive for } H_0 \equiv \text{Cancer}) = 0.03$  (NB: for 1-point sample)
 

$P(\text{Cancer}) = 0.008$	$P(+   \text{Cancer}) = 0.98$	$P(+   \neg \text{Cancer}) = 0.03$
$P(\neg \text{Cancer}) = 0.992$	$P(-   \text{Cancer}) = 0.02$	$P(-   \neg \text{Cancer}) = 0.97$
  - $P(+ | H_0) P(H_0) = 0.0078$ ,  $P(+ | H_A) P(H_A) = 0.0298 \Rightarrow h_{MAP} = H_A \equiv \neg \text{Cancer}$



## Basic Formulas for Probabilities

- **Product Rule (Alternative Statement of Bayes's Theorem)**

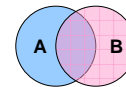
$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$

- Proof: requires axiomatic set theory, as does Bayes's Theorem

- **Sum Rule**

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- Sketch of proof (immediate from axiomatic set theory)
  - Draw a Venn diagram of two sets denoting events *A* and *B*
  - Let  $A \cup B$  denote the event corresponding to  $A \vee B$ ...



- **Theorem of Total Probability**

- Suppose events  $A_1, A_2, \dots, A_n$  are mutually exclusive and exhaustive
  - Mutually exclusive:  $i \neq j \Rightarrow A_i \wedge A_j = \emptyset$
  - Exhaustive:  $\sum P(A_i) = 1$
- Then  $P(B) = \sum_{i=1}^n P(B | A_i) \cdot P(A_i)$
- Proof: follows from product rule and 3<sup>rd</sup> Kolmogorov axiom



## MAP and ML Hypotheses: A Pattern Recognition Framework

- **Pattern Recognition Framework**
  - Automated speech recognition (ASR), automated image recognition
  - Diagnosis
- **Forward Problem: One Step in ML Estimation**
  - Given: model  $h$ , observations (data)  $D$
  - Estimate:  $P(D | h)$ , the “probability that the model generated the data”
- **Backward Problem: Pattern Recognition / Prediction Step**
  - Given: model  $h$ , observations  $D$
  - Maximize:  $P(h(X) = x | h, D)$  for a new  $X$  (i.e., find best  $x$ )
- **Forward-Backward (Learning) Problem**
  - Given: model space  $H$ , data  $D$
  - Find:  $h \in H$  such that  $P(h | D)$  is maximized (i.e., MAP hypothesis)
- **More Info**
  - [http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/\\_HiddenMarkovModels.html](http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/_HiddenMarkovModels.html)
  - Emphasis on a particular  $H$  (the space of hidden Markov models)



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences

## Bayesian Learning Example: Unbiased Coin [1]

- **Coin Flip**
  - Sample space:  $\Omega = \{Head, Tail\}$
  - Scenario: given coin is either fair or has a 60% bias in favor of Head
    - $h_1 \equiv$  fair coin:  $P(Head) = 0.5$
    - $h_2 \equiv$  60% bias towards Head:  $P(Head) = 0.6$
  - Objective: to decide between default (null) and alternative hypotheses
- **A Priori (aka Prior) Distribution on  $H$** 
  - $P(h_1) = 0.75$ ,  $P(h_2) = 0.25$
  - Reflects learning agent's *prior beliefs* regarding  $H$
  - Learning is revision of agent's beliefs
- **Collection of Evidence**
  - First piece of evidence:  $d \equiv$  a single coin toss, comes up Head
  - Q: What does the agent believe now?
  - A: Compute  $P(d) = P(d | h_1) P(h_1) + P(d | h_2) P(h_2)$



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences

## Bayesian Learning Example: Unbiased Coin [2]

- **Bayesian Inference: Compute  $P(d) = P(d | h_1) P(h_1) + P(d | h_2) P(h_2)$** 
  - $P(\text{Head}) = 0.5 \cdot 0.75 + 0.6 \cdot 0.25 = 0.375 + 0.15 = 0.525$
  - This is the probability of the observation  $d = \text{Head}$
- **Bayesian Learning**
  - Now apply Bayes's Theorem
    - $P(h_1 | d) = P(d | h_1) P(h_1) / P(d) = 0.375 / 0.525 = 0.714$
    - $P(h_2 | d) = P(d | h_2) P(h_2) / P(d) = 0.15 / 0.525 = 0.286$
    - *Belief has been revised downwards for  $h_1$ , upwards for  $h_2$*
    - The agent still thinks that the fair coin is the more likely hypothesis
  - Suppose we were to use the ML approach (i.e., assume equal priors)
    - Belief is revised upwards from 0.5 for  $h_1$
    - Data then supports the bias coin better
- **More Evidence: Sequence  $D$  of 100 coins with 70 heads and 30 tails**
  - $P(D) = (0.5)^{50} \cdot (0.5)^{50} \cdot 0.75 + (0.6)^{70} \cdot (0.4)^{30} \cdot 0.25$
  - Now  $P(h_1 | d) \ll P(h_2 | d)$



Kansas State University  
Department of Computing and Information Sciences

CIS 732: Machine Learning and Pattern Recognition

## Brute Force MAP Hypothesis Learner

- **Intuitive Idea: Produce Most Likely  $h$  Given Observed  $D$**
- **Algorithm Find-MAP-Hypothesis ( $D$ )**
  - 1. FOR each hypothesis  $h \in H$   
Calculate the conditional (i.e., posterior) probability:

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

- 2. RETURN the hypothesis  $h_{MAP}$  with the highest conditional probability

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



Kansas State University  
Department of Computing and Information Sciences

CIS 732: Machine Learning and Pattern Recognition

## Relation to Concept Learning

- **Usual Concept Learning Task**
  - Instance space  $X$
  - Hypothesis space  $H$
  - Training examples  $D$
- **Consider *Find-S* Algorithm**
  - Given:  $D$
  - Return: most specific  $h$  in the version space  $VS_{H,D}$
- **MAP and Concept Learning**
  - Bayes's Rule: Application of Bayes's Theorem
  - What would Bayes's Rule produce as the MAP hypothesis?
- **Does *Find-S* Output A MAP Hypothesis?**



## Bayesian Concept Learning and Version Spaces

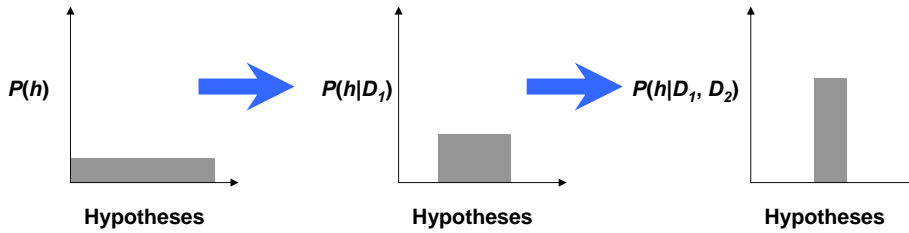
- **Assumptions**
  - Fixed set of instances  $\langle x_1, x_2, \dots, x_m \rangle$
  - Let  $D$  denote the set of classifications:  $D = \langle c(x_1), c(x_2), \dots, c(x_m) \rangle$
- **Choose  $P(D | h)$** 
  - $P(D | h) = 1$  if  $h$  consistent with  $D$  (i.e.,  $\forall x_i . h(x_i) = c(x_i)$ )
  - $P(D | h) = 0$  otherwise
- **Choose  $P(h) \sim$  Uniform**
  - Uniform distribution:  $P(h) = \frac{1}{|H|}$
  - Uniform priors correspond to “no background knowledge” about  $h$
  - Recall: maximum entropy
- **MAP Hypothesis**

$$P(h | D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$$



## Evolution of Posterior Probabilities

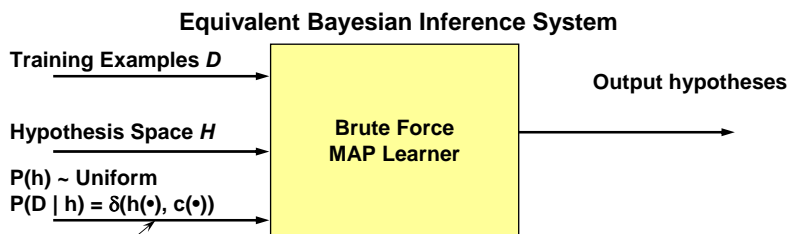
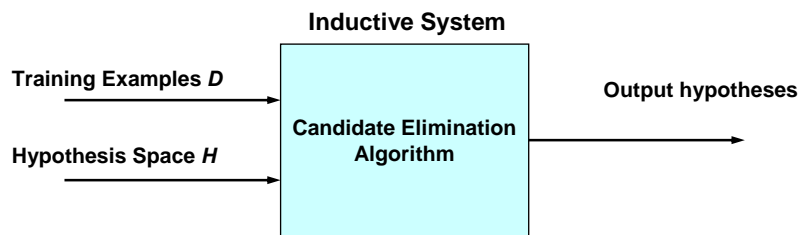
- **Start with Uniform Priors**
  - Equal probabilities assigned to each hypothesis
  - Maximum uncertainty (entropy), minimum prior information



- **Evidential Inference**
  - Introduce data (evidence)  $D_1$ : belief revision occurs
    - Learning agent revises conditional probability of inconsistent hypotheses to 0
    - Posterior probabilities for remaining  $h \in VS_{H,D}$  revised upward
  - Add more data (evidence)  $D_2$ : further belief revision



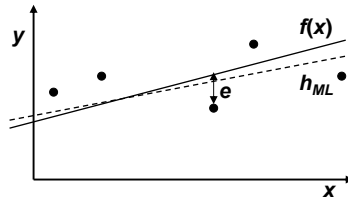
## Characterizing Learning Algorithms by Equivalent MAP Learners



*Prior knowledge made explicit*



## Maximum Likelihood: Learning A Real-Valued Function [1]



- **Problem Definition**
  - Target function: any real-valued function  $f$
  - Training examples  $\langle x_i, y_i \rangle$  where  $y_i$  is noisy training value
    - $y_i = f(x_i) + e_i$
    - $e_i$  is random variable (noise) i.i.d.  $\sim$  Normal  $(0, \sigma)$ , aka Gaussian noise
  - Objective: approximate  $f$  as closely as possible
- **Solution**
  - Maximum likelihood hypothesis  $h_{ML}$
  - Minimizes sum of squared errors (SSE)

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



## Maximum Likelihood: Learning A Real-Valued Function [2]

- **Derivation of Least Squares Solution**
  - Assume noise is Gaussian (prior knowledge)
  - Max likelihood solution:  $h_{ML} = \arg \max_{h \in H} p(D | h)$
- **Problem: Computing Exponents, Comparing Reals - Expensive!**
- **Solution: Maximize Log Prob**

$$\begin{aligned} h_{ML} &= \arg \max_{h \in H} \sum_{i=1}^m \left[ \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2} \left( \frac{d_i - h(x_i)}{\sigma} \right)^2 \right] \\ &= \arg \max_{h \in H} \sum_{i=1}^m \left[ -\frac{1}{2} \left( \frac{d_i - h(x_i)}{\sigma} \right)^2 \right] \\ &= \arg \max_{h \in H} \sum_{i=1}^m -(d_i - h(x_i))^2 \\ &= \arg \min_{h \in H} \sum_{i=1}^m (d_i - h(x_i))^2 \end{aligned}$$



## Learning to Predict Probabilities

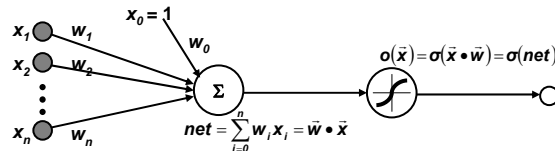
- **Application: Predicting Survival Probability from Patient Data**
- **Problem Definition**
  - Given training examples  $\langle x_i, d_i \rangle$ , where  $d_i \in \mathbb{H} \equiv \{0, 1\}$
  - Want to train neural network to output a probability given  $x_i$  (not a 0 or 1)
- **Maximum Likelihood Estimator (MLE)**
  - In this case can show:

$$h_{ML} = \arg \max_{h \in \mathbb{H}} \sum_{i=1}^m [d_i \ln h(x_i) + (1 - d_i) \ln (1 - h(x_i))]$$

- **Weight update rule for a sigmoid unit**

$$W_{start-layer, end-layer} = W_{start-layer, end-layer} + \Delta W_{start-layer, end-layer}$$

$$\Delta W_{start-layer, end-layer} = r \sum_{i=1}^m (d_i - h(x_i)) \cdot x_{i, start-layer, end-layer}$$



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences



## Most Probable Classification of New Instances

- **MAP and MLE: Limitations**
  - Problem so far: “find the most likely hypothesis given the data”
  - Sometimes we just want the best classification of a new instance  $x$ , given  $D$
- **A Solution Method**
  - Find best (MAP)  $h$ , use it to classify
  - *This may not be optimal, though!*
  - Analogy
    - Estimating a distribution using the mode versus the integral
    - One finds the maximum, the other the area
- **Refined Objective**
  - Want to determine the most probable classification
  - Need to *combine* the prediction of all hypotheses
  - Predictions must be *weighted by their conditional probabilities*
  - Result: Bayes Optimal Classifier (next time...)

CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences



## Terminology

- **Introduction to Bayesian Learning**
  - Probability foundations
    - Definitions: subjectivist, frequentist, logician
    - (3) Kolmogorov axioms
- **Bayes's Theorem**
  - Prior probability of an event
  - Joint probability of an event
  - Conditional (posterior) probability of an event
- **Maximum A Posteriori (MAP) and Maximum Likelihood (ML) Hypotheses**
  - MAP hypothesis: highest conditional probability given observations (data)
  - ML: highest likelihood of generating the observed data
  - ML estimation (MLE): estimating parameters to find ML hypothesis
- **Bayesian Inference: Computing Conditional Probabilities (CPs) in A Model**
- **Bayesian Learning: Searching Model (Hypothesis) Space using CPs**



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences

## Summary Points

- **Introduction to Bayesian Learning**
  - Framework: using probabilistic criteria to search  $H$
  - Probability foundations
    - Definitions: subjectivist, objectivist; Bayesian, frequentist, logicist
    - Kolmogorov axioms
- **Bayes's Theorem**
  - Definition of conditional (posterior) probability
  - Product rule
- **Maximum A Posteriori (MAP) and Maximum Likelihood (ML) Hypotheses**
  - Bayes's Rule and MAP
  - Uniform priors: allow use of MLE to generate MAP hypotheses
  - Relation to version spaces, candidate elimination
- **Next Week: 6.6-6.10, Mitchell; Chapter 14-15, Russell and Norvig; Roth**
  - More Bayesian learning: MDL, BOC, Gibbs, Simple (Naïve) Bayes
  - Learning over text



CIS 732: Machine Learning and Pattern Recognition

Kansas State University  
Department of Computing and Information Sciences