Lecture 9

Bayes's Theorem, MAP, and Maximum Likelihood Hypotheses

Thursday, September 23, 1999

William H. Hsu
Department of Computing and Information Sciences, KSU
http://www.cis.ksu.edu/~bhsu

Readings:
Sections 6.1-6.5, Mitchell

Lecture Outline

- Read Sections 6.1-6.5, Mitchell
- 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
- Next Week: Sections 6.6-6.13, Mitchell; Roth; Pearl and Verma
  - More Bayesian learning: MDL, BOC, Gibbs, Simple (Naïve) Bayes
  - Learning over text

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*(*Outlook* = *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*(*Tomorrow's-Weather* = *Rain*) = 0.5
  - College admissions: *Pr*(*Acceptance*) ≡ *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 (Ω): Range of a Random Variable *X*
- Probability Measure *Pr*(•)
  - Ω denotes a range of "events"; X: Ω
  - Probability *Pr*, or *P*, is a *measure* over Ω
  - 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. $\forall X_1, X_2, \ldots \ni i \neq j \Rightarrow X_i \wedge X_j = \varnothing \,.$
    $$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$$
- Joint Probability: $P(X_1 \wedge X_2) \equiv$ Probability of the Joint Event $X_1 \wedge X_2$
- Independence: $P(X_1 \wedge X_2) = P(X_1) \bullet P(X_2)$

Bayes's Theorem

- **Theorem**

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

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

Choosing Hypotheses

- **Bayes's Theorem**

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

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

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

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

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

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
  - $\alpha \equiv P$(false negative for $H_0 \equiv Cancer$) = 0.02 (*NB*: for 1-point sample)
  - $\beta \equiv P$(false positive for $H_0 \equiv Cancer$) = 0.03 (*NB*: for 1-point sample)

$$\begin{array}{lll} P(Cancer) = 0.008 & P(+ \mid Cancer) = 0.98 & P(+ \mid \neg Cancer) = 0.03 \\ P(\neg Cancer) = 0.992 & P(- \mid Cancer) = 0.02 & P(- \mid \neg Cancer) = 0.97 \end{array}$$

  - $P(+ \mid H_0)\, P(H_0) = 0.0078$, $P(+ \mid H_A)\, P(H_A) = 0.0298 \Rightarrow h_{MAP} = H_A \equiv \neg Cancer$

Basic Formulas for Probabilities

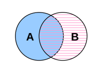
- **Product Rule (Alternative Statement of Bayes's Theorem)**

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

  - 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 $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive
    - Mutually exclusive: $i \neq j \Rightarrow A_i \wedge A_j = \emptyset$
    - Exhaustive: $\sum P(A_i) = 1$
  - Then $P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)$
  - Proof: follows from product rule and 3rd 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: $P(D \mid h)$, the "probability that the model generated the data"
- **Backward Problem: Pattern Recognition / Prediction Step**
  - Given: model h, observations D
  - Maximize: $P(h(X) = x \mid h, D)$ 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 $P(h \mid D)$ is maximized (i.e., MAP hypothesis)
- **More Info**
  - http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/HiddenMarkovModels.html
  - Emphasis on a particular H (the space of hidden Markov models)

Bayesian Learning Example: Unbiased Coin [1]

- **Coin Flip**
  - Sample space: Ω = {*Head*, *Tail*}
  - Scenario: given coin is either fair or has a 60% bias in favor of *Head*
    - h_1 ≡ fair coin: $P(Head) = 0.5$
    - h_2 ≡ 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 ≡ a single coin toss, comes up *Head*
  - Q: What does the agent believe now?
  - A: Compute $P(d) = P(d \mid h_1)\, P(h_1) + P(d \mid h_2)\, P(h_2)$

Bayesian Learning Example: Unbiased Coin [2]

- Bayesian Inference: Compute $P(d) = P(d \mid h_1)\, P(h_1) + P(d \mid h_2)\, P(h_2)$
  - $P(Head) = 0.5 \bullet 0.75 + 0.6 \bullet 0.25 = 0.375 + 0.15 = 0.525$
  - This is the probability of the observation $d = Head$
- Bayesian Learning
  - Now apply Bayes's Theorem
    - $P(h_1 \mid d) = P(d \mid h_1)\, P(h_1) / P(d) = 0.375 / 0.525 = 0.714$
    - $P(h_2 \mid d) = P(d \mid h_2)\, P(h_2) / P(d) = 0.15 / 0.525 = 0.286$
    - *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
  - $P(D) = (0.5)^{50} \bullet (0.5)^{50} \bullet 0.75 + (0.6)^{70} \bullet (0.4)^{30} \bullet 0.25$
  - Now $P(h_1 \mid d) << P(h_2 \mid d)$

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:

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

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

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

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?

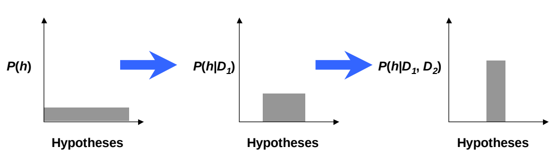
Bayesian Concept Learning and Version Spaces

- Assumptions
  - Fixed set of instances $<x_1, x_2, \ldots, x_m>$
  - Let D denote the set of classifications: $D = <c(x_1), c(x_2), \ldots, c(x_m)>$
- Choose $P(D \mid h)$
  - $P(D \mid h) = 1$ if <u>h consistent with D</u> (i.e., $\forall x_i \,.\, h(x_i) = c(x_i)$)
  - $P(D \mid h) = 0$ otherwise
- Choose $P(h)$ ~ Uniform
  - Uniform distribution: $P(h) = \frac{1}{|H|}$
  - Uniform priors correspond to "no background knowledge" about h
  - Recall: <u>maximum entropy</u>
- MAP Hypothesis

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

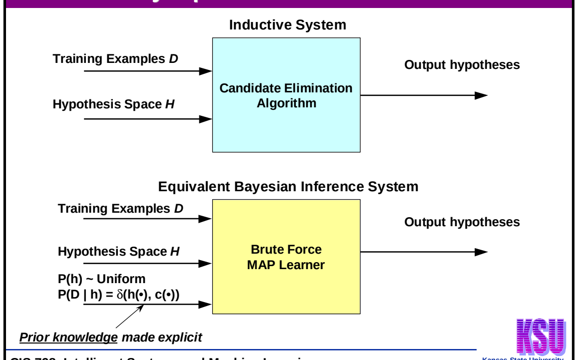
Evolution of Posterior Probabilities

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

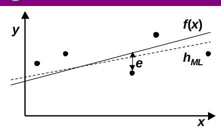


- <u>Evidential Inference</u>
  - Introduce data (<u>evidence</u>) D_1: <u>belief revision</u> occurs
    - Learning agent revises conditional probability of inconsistent hypotheses to 0
    - Posterior probabilities for remaining $h \in VS_{H,D}$ revised upward
  - Add more data (evidence) D_2: further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

Characterizing Learning Algorithms by Equivalent MAP Learners



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



- **Problem Definition**
  - Target function: any real-valued function *f*
  - Training examples $<x_i, y_i>$ where y_i is <u>noisy training value</u>
    - $y_i = f(x_i) + e_i$
    - e_i is random variable (<u>noise</u>) i.i.d. ~ Normal (0, σ), *aka* <u>Gaussian noise</u>
  - Objective: approximate *f* as closely as possible
- **Solution**
  - Maximum likelihood hypothesis h_{ML}
  - Minimizes <u>s</u>um of <u>s</u>quared <u>e</u>rrors (<u>SSE</u>)

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

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

- **Derivation of Least Squares Solution**
  - Assume noise is Gaussian (prior knowledge)
  - Max likelihood solution: $h_{ML} = \arg\max_{h \in H} p(D \mid h)$

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

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

- **Problem: Computing Exponents, Comparing Reals - Expensive!**
- **Solution: Maximize <u>Log Prob</u>**

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

$$= \arg\max_{h \in H} \sum_{i=1}^{m} \left[-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \right]$$

$$= \arg\max_{h \in H} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2$$

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

Learning to Predict Probabilities

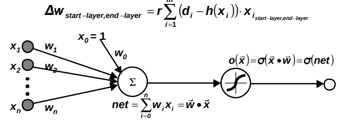
- **Application: Predicting Survival Probability from Patient Data**
- **Problem Definition**
  - Given training examples $<x_i, d_i>$, where $d_i \in \mathcal{H} \equiv \{0, 1\}$
  - Want to train neural network to output a probability given x_i (not a 0 or 1)
- **<u>M</u>aximum <u>L</u>ikelihood <u>E</u>stimator (<u>MLE</u>)**
  - In this case can show:

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

  - Weight update rule for a sigmoid unit

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

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



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 <u>mode</u> versus the <u>integral</u>
    - One finds the maximum, the other the area
- **Refined Objective**
  - Want to determine the <u>most probable classification</u>
  - Need to *combine* the prediction of all hypotheses
  - Predictions must be *weighted by their conditional probabilities*
  - Result: <u>B</u>ayes <u>O</u>ptimal <u>C</u>lassifier (next time…)

Terminology

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

Summary Points

- **Introduction to Bayesian Learning**
  - Framework: using probabilistic criteria to search *H*
  - Probability foundations
    - Definitions: subjectivist, <u>objectivist</u>; Bayesian, frequentist, logicist
    - Kolmogorov axioms
- **Bayes's Theorem**
  - Definition of conditional (posterior) probability
  - Product rule
- **<u>M</u>aximum <u>*A P*</u>*osteriori* (<u>MAP</u>) and <u>M</u>aximum <u>L</u>ikelihood (<u>ML</u>) 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