SVM Continued and Intro to Bayesian Learning: 
*Max a Posteriori* and Max Likelihood Estimation

Thursday, 22 February 2007

William H. Hsu  
Department of Computing and Information Sciences, KSU  
http://www.kddresearch.org

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 (Naive) Bayes  
  - Learning over text
Selection and Building Blocks

- **Restricted Case: Selection Only**
  - \( f(t) \) = average fitness of population at time \( t \)
  - \( m(s, t) \) = number of instances of schema \( s \) in population at time \( t \)
  - \( \bar{u}(s, t) \) = average fitness of instances of schema \( s \) at time \( t \)

- **Quantities of Interest**
  - Probability of selecting \( h \) in one selection step
    \[ P(h) = \frac{f(h)}{\sum f(h)} \]
  - Probability of selecting an instance of \( s \) in one selection step
    \[ P(h \in s) = \sum_{h: f(h)} \frac{f(h)}{f(t)} = \frac{\bar{u}(s, t)}{f(t)} \cdot m(s, t) \]
  - Expected number of instances of \( s \) after \( n \) selections
    \[ E[m(s, t + 1)] \cdot \frac{\bar{u}(s, t)}{f(t)} \cdot m(s, t) \]

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., \( P(\text{Outlook} = \text{Sunny}) = 0.8 \)
  - 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 \( P(\text{Tomorrow's-Weather} = \text{Rain}) = 0.5 \)
  - College admissions: \( P(\text{Acceptance}) = p \)
    - 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 ($\Omega$): Range of a Random Variable $X$
- Probability Measure $Pr(\cdot)$
  - $\Omega$ denotes a range of “events”; $X: \Omega$
  - Probability $Pr$, or $P$, is a measure over $\Omega$
  - 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. $P(\Omega) = \sum_{x \in \Omega} P(X = x) = 1$
  - 3. $\forall X_1, X_2, \ldots, x_i \neq j \Rightarrow X_i \land X_j = \emptyset$.
    - $P(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} P(X_i)$
- Joint Probability: $P(X_1 \land X_2) =$ Probability of the Joint Event $X_1 \land X_2$
- Independence: $P(X_1 \land X_2) = P(X_1) \cdot P(X_2)$

Bayes’s Theorem

- Theorem
  - $P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{P(h \land D)}{P(D)}$
- $P(h)$ = 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 $D$)
- $P(h \mid D)$ = Probability of $h$ Given $D$
  - $\mid$ denotes conditioning - hence $P(h \mid D)$ 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)
- $P(h \land D)$ = Joint Probability of $h$ and $D$
  - Measures probability of observing $D$ and of $h$ being correct
Choosing Hypotheses

- **Bayes’s Theorem**
  \[
P(h | D) = \frac{P(D | h) p(h)}{P(D)} \]  
  where \( h \) is a hypothesis and \( D \) is the data.

- **MAP Hypothesis**
  - Generally want most probable hypothesis given the training data.
  - Define: \( \arg\max_{h \in O} f(x) \) = the value of \( x \) in the sample space \( O \) with the highest \( f(x) \).
  - Maximum a posteriori hypothesis, \( h_{MAP} \):
    \[
    h_{MAP} = \arg\max_{h \in O} P(h | D)
    = \arg\max_{h \in O} \frac{P(D | h) p(h)}{P(D)}
    = \arg\max_{h \in O} P(D | h) p(h)
    \]

- **ML Hypothesis**
  - Assume that \( p(h_i) = p(h_j) \) for all pairs \( i, j \) (uniform priors, i.e., \( P_H \sim \text{Uniform} \)).
  - Can further simplify and choose the maximum likelihood hypothesis, \( h_{ML} \):
    \[
    h_{ML} = \arg\max_{h \in O} P(D | h)
    \]

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
  - \( \alpha = P(\text{false negative for } H_0 = \text{Cancer}) = 0.02 \) (NB: for 1-point sample)
  - \( \beta = P(\text{false positive for } H_0 = \text{Cancer}) = 0.03 \) (NB: for 1-point sample)
  - \[
  P(\text{Cancer}) = 0.008 \quad P(+) | \text{Cancer} = 0.98 \quad P(+) | \neg \text{Cancer} = 0.03
  P(\neg \text{Cancer}) = 0.992 \quad P(-) | \text{Cancer} = 0.02 \quad P(-) | \neg \text{Cancer} = 0.97
  
  \[
  P(+) | H_0 \) P(H_0) = 0.0078, P(+) | H_1 \) P(H_1) = 0.0298 \Rightarrow h_{MAP} = H_1 = \neg \text{Cancer}
  \]
### Basic Formulas for Probabilities

- **Product Rule** (Alternative Statement of Bayes’s Theorem)
  \[ P(A \mid B) = \frac{P(A \land B)}{P(B)} \]
  - Proof: requires axiomatic set theory, as does Bayes’s Theorem

- **Sum Rule**
  \[ P(A \lor B) = P(A) + P(B) - P(A \land B) \]
  - Sketch of proof (immediate from axiomatic set theory)
    - Draw a Venn diagram of two sets denoting events A and B
    - Let \( A \lor B \) denote the event corresponding to \( A \lor B \)

- **Theorem of Total Probability**
  - Suppose events \( A_1, A_2, \ldots, A_n \) are mutually exclusive and exhaustive
    - **Mutually exclusive:** \( i \neq j \Rightarrow A_i \land A_j = \emptyset \)
    - **Exhaustive:** \( \sum P(A_j) = 1 \)
  - Then
    \[ P(B) = \sum_{i} P(B \mid A_i) \cdot P(A_i) \]
  - 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**
  - Emphasis on a particular \( H \) (the space of hidden Markov models)
Bayesian Learning Example: Unbiased Coin [1]

• Coin Flip
  – Sample space: \( \Omega = \{ \text{Head}, \text{Tail} \} \)
  – Scenario: given coin is either fair or has a 60% bias in favor of \text{Head}
    • \( h_1 \equiv \text{fair coin}: P(\text{Head}) = 0.5 \)
    • \( h_2 \equiv 60\% \text{ bias towards Head}: P(\text{Head}) = 0.6 \)
  – Objective: to decide between default (null) and alternative hypotheses

• \text{A Priori} (aka \text{Prior}) Distribution on \( H \)
  – \( P(h_1) = 0.75, P(h_2) = 0.25 \)
  – Reflects learning agent’s \text{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 \text{Head}
  – Q: What does the agent believe now?
  – A: Compute \( P(d) = P(d | h_1) P(h_1) + P(d | h_2) P(h_2) \)

\( P(\text{Head}) = 0.5 \times 0.75 + 0.6 \times 0.25 = 0.375 + 0.15 = 0.525 \)

• Bayesian Learning
  – Now apply Bayes’s Theorem
    • \( P(h_1 | d) = P(d | h_1) P(h_1) / P(d) = 0.375 / 0.525 = 0.714 \)
    • \( P(h_2 | d) = P(d | h_2) P(h_2) / P(d) = 0.15 / 0.525 = 0.286 \)
    • \text{Belief has been revised downwards for} \( h_1 \), \text{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

Bayesian Learning Example: Unbiased Coin [2]

• Bayesian Inference: Compute \( P(d) = P(d | h_1) P(h_1) + P(d | h_2) P(h_2) \)
  – \( P(\text{Head}) = 0.5 \times 0.75 + 0.6 \times 0.25 = 0.375 + 0.15 = 0.525 \)
  – This is the probability of the observation \( d = \text{Head} \)

• Bayesian Learning
  – Now apply Bayes’s Theorem
    • \( P(h_1 | d) = P(d | h_1) P(h_1) / P(d) = 0.375 / 0.525 = 0.714 \)
    • \( P(h_2 | d) = P(d | h_2) P(h_2) / P(d) = 0.15 / 0.525 = 0.286 \)
    • \text{Belief has been revised downwards for} \( h_1 \), \text{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
  – \( P(D) = (0.5)^{50} \times (0.5)^{50} \times 0.75 + (0.6)^{70} \times (0.4)^{30} \times 0.25 \)
  – Now \( P(h_1 | d) << P(h_2 | d) \)
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 $V_{S,H,D}$
- **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 \(<x_1, x_2, ..., x_m>\)
  - Let \(D\) denote the set of classifications: \(D = <c(x_1), c(x_2), ..., c(x_m)>\)

- Choose \(P(D | h)\)
  - \(P(D | h) = 1\) if \(h\) consistent with \(D\) (i.e., \(\forall x, h(x) = c(x)\))
  - \(P(D | h) = 0\) otherwise

- Choose \(P(h) \sim \text{Uniform}\)
  - Uniform distribution: \(P(h) = \frac{1}{|H|}\)
  - Uniform priors correspond to “no background knowledge” about \(h\)
  - Recall: maximum entropy

- MAP Hypothesis

\[
P(h|D) = \begin{cases} 
\frac{1}{|VS_D|} & \text{if } h \text{ is consistent with } D \\
0 & \text{otherwise}
\end{cases}
\]

Evolution of Posterior Probabilities

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

  \[
P(h) \rightarrow P(h|D_1) \rightarrow P(h|D_1, D_2)
\]

- Evidential Inference
  - Introduce data (evidence) \(D_i\): 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) \(D_j\): further belief revision
**Characterizing Learning Algorithms by Equivalent MAP Learners**

- **Inductive System**
  - Training Examples $D$
  - Hypothesis Space $H$
  - Candidate Elimination Algorithm
  - Output hypotheses

- **Equivalent Bayesian Inference System**
  - Training Examples $D$
  - Hypothesis Space $H$
  - Prior knowledge made explicit
  - $P(h) \sim \text{Uniform}$
  - $P(D \mid h) = \delta(h(\cdot), c(\cdot))$
  - Brute Force MAP Learner
  - Output hypotheses

**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...)
Terminology

- Introduction to Bayesian Learning
  - Probability foundations
    - Definitions: subjectivist, frequentist, logicist
    - (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

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