Lecture 21

Uncertain Reasoning Discussion (2 of 4): Learning Bayesian Network Structure

Friday, March 10, 2000

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

Readings:
"Learning Bayesian Network Structure from Massive Datasets", Friedman, Nachman, and Pe'er
(Reference) Section 6.11, Mitchell

Lecture Outline

- Suggested Reading: Section 6.11, Mitchell
- Overview of Bayesian Learning (Continued)
- Bayes's Theorem (Continued)
  - 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
- Later
  - Applications of probability in KDD
    - Learning over text
    - Learning over hypermedia documents
    - General HCII (Yuhui Liu: March 13, 2000)
  - Causality (Yue Jiao: March 17, 2000)

Probability: Basic Definitions and Axioms

- Sample Space (Ω): Range of a Random Variable X
- Probability Measure $Pr(\bullet)$
  - Ω denotes a range of outcomes; $X: \Omega$
  - Probability P: *measure* over 2^{Ω} (power set of sample space, *aka* event space)
  - 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. $\forall x \in \Omega \,.\, 0 \le P(X = x) \le 1$
  - 2. $P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1$
  - 3. $\forall X_1, X_2, \ldots \ni i \ne j \Rightarrow X_i \wedge X_j = \emptyset$.

$$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$$

- Joint Probability: $P(X_1 \wedge X_2) \equiv$ Probability of the Joint Event $X_1 \wedge X_2$
- Independence: $P(X_1 \wedge X_2) = P(X_1) \bullet P(X_2)$

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: $\arg\max_{x \in \Omega} [f(x)] \equiv$ the value of x in the sample space Ω with the highest $f(x)$
  - Maximum *a posteriori* hypothesis, h_{MAP}

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

- 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}

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

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)

$P(Cancer) = 0.008$	$P(+ \mid Cancer) = 0.98$	$P(+ \mid \neg Cancer) = 0.03$
$P(\neg Cancer) = 0.992$	$P(- \mid Cancer) = 0.02$	$P(- \mid \neg Cancer) = 0.97$

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

Basic Formulas for Probabilities

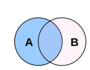
- Product Rule (Alternative Statement of Bayes's Theorem)

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

  - 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: $i \ne j \Rightarrow A_i \wedge A_j = \emptyset$
    - 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: $\Omega = \{Head, Tail\}$
  - 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 $P(d) = P(d \mid h_1)\, P(h_1) + P(d \mid h_2)\, P(h_2)$

Bayesian Learning Example: Unbiased Coin [2]

- Bayesian Inference: Compute $P(d) = P(d \mid h_1)\, P(h_1) + P(d \mid h_2)\, P(h_2)$
  - $P(Head) = 0.5 \cdot 0.75 + 0.6 \cdot 0.25 = 0.375 + 0.15 = 0.525$
  - 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
  - $P(D) = (0.5)^{50} \cdot (0.5)^{50} \cdot 0.75 + (0.6)^{70} \cdot (0.4)^{30} \cdot 0.25$
  - Now $P(h_1 \mid d) << P(h_2 \mid d)$

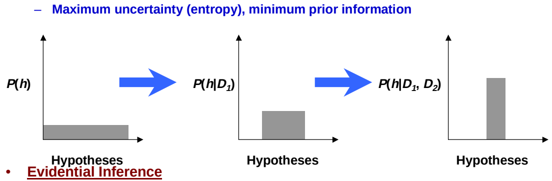
Bayesian Concept Learning and Version Spaces

- 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 h consistent with D (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: maximum entropy
- 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



- 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

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 (see CIS 798 Lecture 10)

Midterm Review: Topics Covered

- Review: Inductive Learning Framework
  - Search in hypothesis space *H*
  - Inductive bias: *preference for some hypotheses over others*
  - Search in space of *hypothesis languages*: bias optimization
- Analytical Learning
  - Learning architecture components: hypothesis languages, domain theory
  - Learning algorithms: EBL, hybrid (analytical and inductive) learning
- Artificial Neural Networks (ANN)
  - Architectures (hypothesis languages): MLP, Boltzmann machine, GLIM hierarchy
  - Algorithms: backpropagation (gradient), MDL, EM
  - Tradeoffs and improvements: momentum, wake-sleep, modularity / HME
- Bayesian Networks
  - Learning architecture: BBN (graphical model of probability)
  - Learning algorithms: CPT (e.g., gradient); structure (polytree, K2)
  - Tradeoffs and improvements: polytrees vs. multiply-connected BBNs, etc.

Midterm Review: Applications and Concepts

- Methods for Multistrategy (Integrated Inductive and Analytical) Learning
  - Analytical learning to drive inductive learning: EBNN, Phantom Induction, advice-taking agents
  - Interleaved analytical and inductive learning: Chown and Dietterich
- Artificial Neural Networks in KDD
  - Tradeoffs and improvements
    - Reinforcement learning models: temporal differences, ANN methods
    - Wake-sleep
    - Modularity (mixture models and hierarchical mixtures of experts)
    - Combining classifiers
  - Applications to KDD: learning for pattern (e.g., image) recognition, planning
- Bayesian Networks in KDD
  - Advantages of probability, causal networks (BBNs)
  - Applications to KDD: learning to reason

Terminology

- Introduction to Bayesian Learning
  - Probability foundations
  - Definitions: subjectivist, frequentist, logicist, objectivist
  - (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 Class: Presentation on Learning Bayesian (Belief) Network Structure
  - For more on Bayesian learning: MDL, BOC, Gibbs, Simple (Naïve) Bayes
  - Soon: user modeling using BBNs, causality