CMSC828G **Principles of Data Mining** Lecture #5

- Announcements:
  - Eiman's office hours changed to T 12:30-1:30, Th 9:30-10:30
  - Book errata: if you find any mistakes, send email to eiman@cs.umd.edu

- Today's Reading:
  - HMS, chapter 4

- Today's Lecture:
  - Dealing with Uncertainty
  - Probability Review
  - Parameter Estimation

- Upcoming Due Dates:
  - P1 writeup due 2/13, in class presentation 2/14
  - H1 due 2/21

Uncertainty

- ubiquitous

Methods for Dealing with Uncertainty

- fuzzy logic
- possibility theory
- rough sets
- certainty factors
- nonmonotonic logic
- PROBABILITY

A **Brief** History of Probability

- 1654, Blaise Pascal and Pierre de Fermat develop first mathematical theory of probability to analysis a game of dice.
- 1812, Pierre de Laplace applied probabilistic ideas to many scientific and practical problems and introduced a host of new ideas and mathematical techniques in his book, *Théorie Analytique des Probabilités*. until this point, probability theory was solely concerned with developing a mathematical analysis of games of chance.
- 1933, Andrei Kolmogorov states axioms of modern probability theory.

Probability

- Probability Theory
  - interpretation of statements
- Probability Calculus
  - manipulation of mathematical representation

Probability Theory

- different perspectives on meaning of probability
- many debates and controversies
- two important different views:
  Frequentist vs. Bayesian

Frequentist View

- probability is an objective concept
- long-term relative frequency of an event occurring under many repeated "similar" trials
- e.g., number of times we observe heads in repeated coin tosses
- dominant perspective throughout most of the last century

Bayesian Viewpoint

- probability is "degree-of-belief", or "degree-of-uncertainty".
- To the Bayesian, probability lies subjectively in the mind, and can--with validity--be different for people with different information
- e.g., the probability that Bush will be reelected in 2004.
- In contrast, to the frequentist, probability lies objectively in the external world.
- The Bayesian viewpoint has been gaining popularity in the past decade, largely due to the increase computational power that makes many of the calculations that were previously intractable, feasible.

Bayesian Statistics

- Central tenet: explicit characterization of all forms of uncertainty in the problem:
  - uncertainty about any parameters
  - uncertainty about model
  - uncertainty about forecast
- Subjective probability – internal state of individual
- Fortunately, if we adopt tenets of rational behavior, the set of axioms underlying subjective probability is the same as for frequentist viewpoint.
- Thus, while the interpretations may differ, the calculus is the same.

Things that Bayesians can do that Frequentists can't

- Calculate the probability of something that occurred in the past but whose outcome is not known. To the frequentist, this is meaningless. e.g., What is the probability that fossil A predates fossil B?
- Calculate the probability of an event resulting from a trial that can occur once only. Frequentists cannot.
- Computing the probability an HIV test giving a false negative when selecting someone from the general population. The question is Bayesian, because we are estimating the probability of someone already selected (past event) being HIV-positive. From a frequentist point of view, the person either is or is not HIV-positive (their status would be an unknown parameter), but we could not make probability statements about it.
- Here, we are beginning with a degree-of-belief probability statement about the person's HIV status in the absence of a blood test, and then updating our probability estimate by conditioning on the data--the result of the blood test.

Bayesian Inference

- Begin with "prior probability" estimates, based on whatever expert information is available, then you update those estimates by conditioning on observed data.

Inference Differences

- The fundamental difference between Bayesian inference and frequentist inference, is how they define probability. The appropriate interpretations of their inferential results are different as well. Frequentists, for example, may say, "this interval is the result of a procedure that had a 95% chance of creating an interval that would contain the mean". Bayesians may say "this interval has a 95% chance of containing the mean".
- Regarding tests of significance, frequentists examine the probability of data (statistics) given models (hypothesis). Bayesians examine the probability of different models given data.
- The red flag that pops up in people's minds when they first learn about Bayesian methods is due to the word "subjective". "Statistical analysis should be objective!" they cry. And many (but not all) Bayesians agree: statistical analysis should be as objective as possible. But there is no such thing as total objectivity, and subjectivity creeps into frequentist methods as well.
- the nature of Bayesian inference: "Extraordinary claims require extraordinary evidence." --

Random Variables

- Intuition: When the outcomes of an event that produces random results are numerical, the numbers obtained are called random variables. The sample space for the event is just a list containing all possible values of the random variable. This can be defined more precisely in terms of measure theory, however it will not be necessary for our data mining purposes.
- **Random variable**: The outcome of a random phenomenon.
- **Discrete random variable X** has a finite number of possible values, $x \in \{x_1, \ldots, x_m\}$.
- The **probability distribution** of X is written P(X=x).
- $P(X) = \{p(x_1), \ldots, p(x_m)\}$ is the **probability mass function**, p(x) denotes the probability that X takes on the particular value x.
- The **cumulative density** P(x) is the probability that X will take on a value less than x (if values are ordered):

$$P(x) = \sum_{x_i < x} p(x_i)$$

Random Variables, cont.

- A **continuous random variable** takes all values in an interval of numbers.
- The behavior of a continuous random variable X is described by a continuous curve f(x), or p(x), the **probability density function**
- If we want to make probability statements about X, we will have to consider the probability that X falls in an interval (a, b), that is:

$$P(a \le x \le b) = \int_a^b f(x)\,dx$$

Expected Value

- If X is discrete random variable with probability mass function p(x),

$$E(X) = \sum_{x} x\, p(x)$$

- If X is continuous random variable with probability density function f(x)

$$E(X) = \int_{x} x\, f(x) dx$$

- Variance of X:

$$Var(X) = E(X^2) - (E(X))^2$$

- Linearity of Expectation: E(X+Y) = E(X) + E(Y)

Common distributions

- A random variable X has a **Bernoulli distribution** with parameter *p* if it can assume a value of 1 with a probability of *p* and the value of 0 with a probability of (1-*p*).

$$P(x) = p^{x}(1-p)^{1-x}$$

- The mean of a random variable X that has a **Bernoulli distribution** with parameter p is

$$E(X) = 1(p) + 0(1-p) = p$$

- The variance of X is

$$Var(X) = E(X^2) - (E(X))^2 = 1^2(p) + 0^2(1-p) - p^2 = p(1-p)$$

- e.g. A random variable whose value represents the outcome of a fair coin toss (1 for heads, 0 for tails, or vice-versa) is a **Bernoulli** variable with parameter p = 0.5.

Binomial distribution

- A random variable X describing the number of successful outcomes in n independent Bernoulli trials, each with a probability p of success, has a binomial distribution:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- The mean of a random variable X that has a binomial distribution with parameters n and p is

$$E(X) = np$$

- The variance of X is

$$Var(X) = np(1-p)$$

- e.g., A random variable whose value represents the number of heads in a sequence of n tosses of a fair coin toss (1 for heads, 0 for tails, or vice-versa) is a binomial variable with parameters n and p = 0.5.

Multinomial distribution

- Generalization of binomial distribution to the case where there are k possible outcomes; outcome i has probability p_i of occurring. Let X_i denote the number of times in n trials the ith outcome occurs. Then $X_1,\ldots,X_k$ has a multinomial distribution:

$$P(x_1, \ldots, x_k) = \binom{n}{x_1 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

- The mean of a random variable X_i that has a binomial distribution with parameters n and p is

$$E(X_i) = np_i$$

- The variance of X_i is

$$Var(X_i) = np(1-p)$$

- e.g., A random variable whose value represents the number of outs, singles, doubles, triples, homeruns, etc. in a sequence of n times at bats.

Normal (Gaussian) Distribution

- Density:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}$$

- μ is mean and σ^2 is variance

- important distribution

- **central limit theorem**: the distribution of the mean of a sample of n observations becomes more and more like the normal distribution as n increases, regardless of the form of the population distribution from which data is drawn

Multivariate Random Variables

- A multivariate random variable **X** is set $X_1,\ldots,X_p$ random variables
- The density function f(X) is called the **joint density function**
- $f(x_1,\ldots,x_p)$ shorthand for $f(X_1 = x_1,\ldots, X_p = x_p)$
- The density of any subset of the complete set of variables is a **marginal density** of the joint
- e.g. suppose we have three rvs, X_1, X_2 and X_3, with joint $f(X_1, X_2, X_3)$. The marginal density of X_1 is

$$f(x_1) = \sum_{x_2, x_3} f(x_1, x_2, x_3)$$

- if the rvs are discrete. If the rvs are continuous,

$$f(x_1) = \iint_{x_2, x_3} f(x_1, x_2, x_3)$$

Multivariate Random Variables, cont.

- The density of any subset of the complete set of variables conditioned on particular values of the other variables is a **conditional density** of the joint
- e.g. suppose we have three rvs, X_1, X_2 and $X_{3,}$ with joint $f(X_1, X_2, X_3)$. The conditional density of X_1 igiven $X_2 = x_2$ and $X_3 = x_3$ is

$$f(x_1 \mid x_2, x_3) = \frac{f(x_1, x_2, x_3)}{f(x_2, x_3)}$$

Example

- customer purchases: Bread, Bagels and Butter (R,A,U)

Bread	Bagels	Butter	p(r,a,u)
0	0	0	0.24
0	0	1	0.06
0	1	0	0.12
0	1	1	0.08
1	0	0	0.12
1	0	1	0.18
1	1	0	0.04
1	1	1	0.16

Independence

- Two variables X and Y are **independent** if

$$P(x,y) = P(x)P(y)$$

- or, equivalently,

$$P(x \mid y) = P(x)$$

- in other words, Y carries no information about X

Example #1

Butter	p(u)
0	0.52
1	0.48

Bread	Bagels	Butter	p(r,a,u)
0	0	0	0.24
0	0	1	0.06
0	1	0	0.12
0	1	1	0.08
1	0	0	0.12
1	0	1	0.18
1	1	0	0.04
1	1	1	0.16

Bagels	p(a)
0	0.6
1	0.4

Bread	p(r)
0	0.5
1	0.5

Bagels	Butter	p(a,u)
0	0	0.36
0	1	0.24
1	0	0.16
1	1	0.24

$P(a,u)=P(a)P(u)$?

Bread	Bagels	p(r,a)
0	0	0.3
0	1	0.2
1	0	0.3
1	1	0.2

$P(r,a)=P(r)P(a)$?

Conditional Independence

- Let **X**, **Y** and **Z** be sets of random variables. **X** and **Y** are conditional independent given **Z**, denoted I(**X**,**Y**|**Z**), iff, for all values of **X**, **Y** and **Z**

$$P(x, y \mid z) = P(x \mid z)P(y \mid z)$$

Example #2

Hotdogs	Mustard	Ketchup	p(h,m,k)
0	0	0	0.576
0	0	1	0.144
0	1	0	0.064
0	1	1	0.016
1	0	0	0.004
1	0	1	0.036
1	1	0	0.016
1	1	1	0.144

Mustard	p(m)
0	0.76
1	0.24

Ketchup	p(k)
0	0.66
1	0.34

Mustard	Ketchup	p(m,k)
0	0	0.58
0	1	0.18
1	0	0.08
1	1	0.16

P(m,k)=P(m)P(k)?

Example #2

H	M	K	p(h,m,k)
0	0	0	0.576
0	0	1	0.144
0	1	0	0.064
0	1	1	0.016
1	0	0	0.004
1	0	1	0.036
1	1	0	0.016
1	1	1	0.144

Mustard	Hotdogs	p(m|h)
0	0	0.9
0	1	0.2
1	0	0.1
1	1	0.8

Ketchup	Hotdogs	p(m|h)
0	0	0.8
0	1	0.1
1	0	0.2
1	1	0.9

Mustard	Ketchup	Hotdogs	p(m,k|h)
0	0	0	0.72
0	1	0	0.18
1	0	0	0.08
1	1	0	0.02
0	0	1	0.02
0	1	1	0.18
1	0	1	0.08
1	1	1	0.72

$$P(m,k|h)=P(m|h)P(k|h)?$$

Example #1

Bread	Bagels	Butter	p(r,a,u)
0	0	0	0.24
0	0	1	0.06
0	1	0	0.12
0	1	1	0.08
1	0	0	0.12
1	0	1	0.18
1	1	0	0.04
1	1	1	0.16

Bread	Butter	p(r|u)
0	0	0.69…
0	1	0.29…
1	0	0.30…
1	1	0.70…

Bagels	Butter	p(a|u)
0	0	0.69…
0	1	0.5
1	0	0.30…
1	1	0.5

Bread	Bagels	Butter	p(r,a|u)
0	0	0	0.46…
0	1	0	0.23…
1	0	0	0.23…
1	1	0	0.08…
0	0	1	0.12…
0	1	1	0.17…
1	0	1	0,38…
1	1	1	0.33…

$$P(r,a|u)=P(r|u)P(a|u)?$$

Summary

- Example 1: $I(X,Y|\emptyset)$ and not $I(X,Y|Z)$
- Example 2: $I(X,Y|Z)$ and not $I(X,Y|\emptyset)$

- conclusion: independence does not imply conditional independence!

Sequential Data

- sequence of values, x_1, ..., x_n
- a common assumption made is that next value in the sequence is independent of all of the past values given the current value:

$$P(X_j \mid X_{j-1} \ldots X_1) = P(X_j \mid X_{j-1})$$

- **first-order Markov assumption**
- Allows factorization of joint into simpler products:

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{j=2}^{n} p(X_j \mid X_{j-1})$$

Next Time

- Reading:
  - HMS, chapter 4
- Topic:
  - Student data presentations

References

- *Principles of Data Mining*, Hand, Mannila, Smyth. MIT Press, 2001.
- http://www.cc.gatech.edu/classes/cs6751_97_winter/Topics/stat-meas/probHist.html
- http://mathforum.org/epigone/apstat-l/folkixblon