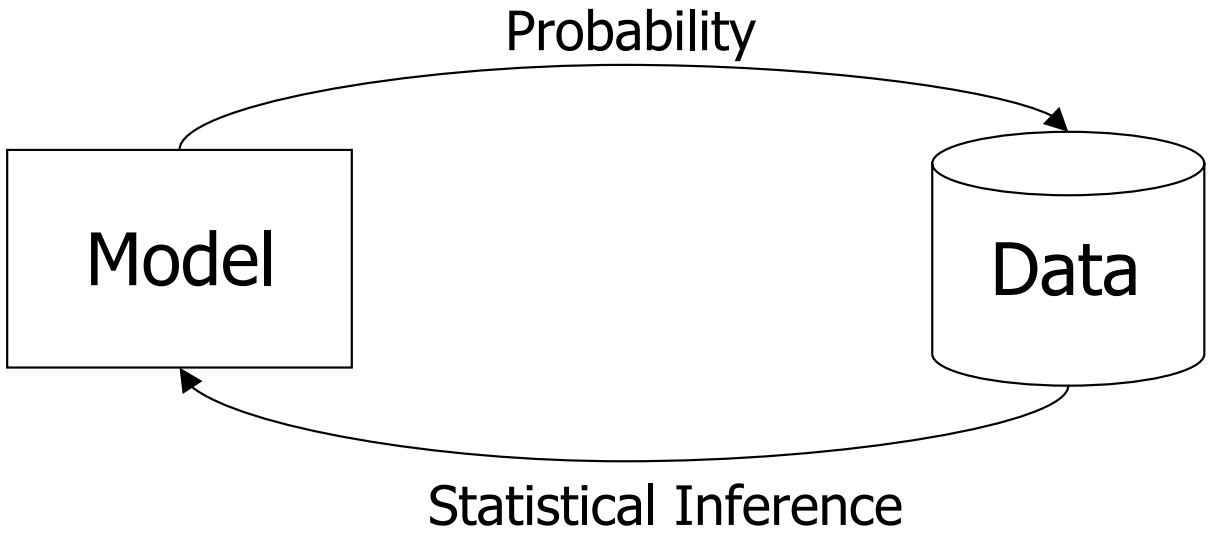
CMSC828G **Principles of Data Mining** Lecture #8

- Today's Reading:
  - HMS, chapter 4, 5
- Today's Lecture:
  - Parameter Estimation cont.
  - Hypothesis Testing
  - Data mining: the component view
- Upcoming Due Dates:
  - H1 due now

Statistical Inference



- **Statistical inference**: inferring properties of an unknown distribution from data generated by that distribution.
- **Estimate** parameters of the model from data
- Use **likelihood** function:

Statistical Inference, cont.

- assumes that the sample has been drawn from the population at random
- The model specifies the distribution for the population; the probability that a particular value for the variable will appear in the sample
- If we have a model M for the data, we can compute the probability that a random sampling process will lead to the data D = {x(1),…,x(x)}:

$$p(D \mid M)$$

- if we assume the probability of each data point is independent, or 'drawn at random' then

$$p(D \mid \theta, M) = \prod_{i=1}^{n} p(x(i) \mid \theta, M)$$

Statistical Inference, cont.

$$p(D \mid \theta, M) = \prod_{i=1}^{n} p(x(i) \mid \theta, M)$$

- based on this value, we can decide how realistic the assumed model is.
- if the assumed model is unlikely to have generated the data, we might reject the model; this is the principle behind hypothesis testing
- we can also estimate population values for the parameters…

Parameter estimation

- Maximum Likelihood Estimation cont.
- Bayesian Estimation

Likelihood Function

- Let D = {x(1),…,x(n)}
- independently sampled, from the same distribution p(x|θ)
  'independent and identically distributed', iid
- The **likelihood function**, L(θ| x(1),…,x(n)) captures the probability of the data as a function of θ

$$L(\theta, D) = L(\theta \mid x(1),\ldots, x(n))$$
$$= p(x(1),\ldots, x(n) \mid \theta)$$
$$= \prod_{i=1}^{n} p(x(i) \mid \theta)$$

Likelihood

- L(X, θ) for a parameter θ and data X is the **relative** probability of getting the data for the different possible parameters θ.
- We must remember that it is **not** a probability distribution.
- For a probability distribution, θ is fixed and you can get all possible values of the random variable X.
- For the likelihood funtion, X=x is fixed and you consider how the probability of getting this x changes when you change θ.

Maximum Likelihood Estimation (MLE)

- Most widely used method of parameter estimation
- Using the likelihood function, L(θ| x(1),…,x(n))

$$
\begin{aligned}
L(\theta, D) &= L(\theta \mid x(1),\ldots, x(n)) \\
&= p(x(1),\ldots, x(n) \mid \theta) \\
&= \prod_{i=1}^{n} p(x(i) \mid \theta)
\end{aligned}
$$

- choose value that maximizes the likelihood function

$$\hat{\theta}_{MLE}$$

Sufficient Statistics

- (informal definition)
- s(D) is a sufficient statistic for θ if the likelihood L(θ) depends on D only through s(D), i.e., L(θ,D)=L(θ,s(D))
- e.g., for our binomial example, r, the number of customers who have purchased milk, is a sufficient statistic (assuming n, the number of customers, is known)
- Why is this important?

for large datasets, instead of working with the entire data set, we can compute and store just sufficient statistics

Bayesian Estimation

- frequentist approach: parameters of a population are fixed but unknown, and data is a random sample from that population
- Bayesian approach: the data is known, and the parameters are the random variables. θ has a distribution of possible values and the observed data provide evidence for different values

Bayesian Estimation, cont.

- Before observing the data, we have a distribution over possible values for θ which is called the **prior distribution**

$$p(\theta)$$

- By analyzing the data, we can update our beliefs to take into account the observed data. This leads to a distribution over possible values for θ given D which is called the **posterior distribution**

$$p(\theta \mid D)$$

- We use Bayes' theorem to obtain the posterior:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

Bayesian Estimation, cont.

- The posterior gives a distribution over parameter values. If we would like a single value (a **point estimate**, as in the MLE case), we can use the same princple, choose the value of θ which maximizes the posterior.
- This is called the **maximum a posteriori (MAP)** method

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid D)$$

MLE is a special case of MAP with 'flat' prior

Bayesian Estimation, cont.

- For a given data set D and a particular model, P(D) is constant, so we often write:

$$\underbrace{p(\theta \mid D)}_{\text{posterior}} \propto \underbrace{p(D \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

Example

- Consider our example of customers milk buying habits, where we are trying to estimate θ, the probability that a customer will buy milk
- The most commonly chosen model for the prior of a binomial rv is a **Beta distribution**.
- The beta distribution has parameters $\alpha > 0$, $\beta > 0$, and is:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta^{\beta-1})$$

$$E[\theta] = \frac{\alpha}{\alpha+\beta}, \quad \text{var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- relative sizes of α and β control the location and spread of the distribution for θ.
- The variance is inversely proportional to $\alpha + \beta$; size of the sum controls the narrowness of the prior. If $\alpha + \beta$ is large, the prior will be sharply peaked.

Example

- Beta prior:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta^{\beta-1})$$

- Likelihood for Binomial model:

$$L(\theta \mid D) = \theta^{r}(1-\theta^{n-r})$$

- Posterior:

$$p(\theta \mid D) = p(D \mid \theta)p(\theta)$$

$$= \theta^{r}(1-\theta^{n-r})\theta^{\alpha-1}(1-\theta^{\beta-1})$$

$$= \theta^{r+\alpha-1}(1-\theta^{n-r+\beta-1})$$

- Beta is like a Binomial with α - 1 prior successes and β - 1 prior failures;
- can think of Beta as having equivalent sample size $\alpha + \beta - 2$
- Posterior is Binomial Beta distribution with parameters r + α and n - r + β

BETA is CONJUGATE PRIOR for BINOMIAL

Bayesian Approach to Prediction

- A fully Bayesian approach is characterized by maintaining a distribution over models
- In order to make a prediction about a new data point x(n+1) not in our training data set D, we average over all possible values of θ, weighted by the posterior probability p(θ|D):

$$p(x(n+1) \mid D) = \int p(x(n+1), \theta \mid D) d\theta$$

$$= \int p(x(n+1) \mid \theta) p(\theta \mid D) d\theta$$

Of course, computationally, this is much more challenging than the maximum likelihood approach… MCMC

Classical Hypothesis Testing

- Two hypothesis H_0 (the null hypothesis) and H_1 (the alternative hypothesis)
- Outcome of a hypothesis test is 'reject H_0' or 'do not reject H_0'
- Example:
  - H_0: there is no difference in taste between coke and pespi against H_1: there is a difference.
- The hypotheses are often statements about population parameters like expected value and variance:
  - for example H_0 might be that the expected value of the height of ten year old boys in the Scottish population is not different from that of ten year old girls
- A hypothesis might also be a statement about the distributional form of a characteristic of interest,
  - the height of ten year old boys is normally distributed within the Scottish population

Classical Hypothesis Testing, cont.

- One general form: $H_0 : \theta = \theta_0$ and $H_1: \theta \neq \theta_0$
- Using the observed data, we calculate some test statistic t_0
- The statistic will vary from sample to sample, so it is a random variable
- If we assume H_0 is true, we can determine the expected distribution for the statistic t
- We can find range of extreme values that probability less than 0.05 of occurring
- If t_0 lies in this range, we reject H_0 at the 5% significance level
- one-sided and two-sided tests

Errors

		Decision	
		Reject H0	**Don't reject H0**
Truth	**H0**	Type I Error	Right Decision
	H1	Right Decision	Type II Error

- type I error: null hypothesis is rejected when it is in fact true;
- e.g., in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug
- a type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.
- A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error.
- **P(type I error)** = significance level = α

Z-test

Example

Testing for Independence

- Two nominal variables $X = x_1, \ldots, x_r$ and $Y = y_1, \ldots, y_s$
- $H_0 = I(X,Y)$
- Let $n(x_i)/n$ be then proportion of observations with $X = x_i$, similarly for $n(y_j)/n$
- If X and Y are independent then we expect $n(x_i\ y_j) = n(x_i)\ n(y_j)$
- Number the possible combinations 1, …, t
- Let E_k denote the expected number of occurrences of combination k; let O_k denoted the observed occurrences

$$X^2 = \sum_{k=1}^{t} \frac{(E_k - O_k)}{E_k}$$

If H_0 is true, X^2 follows a chi-squared distribution

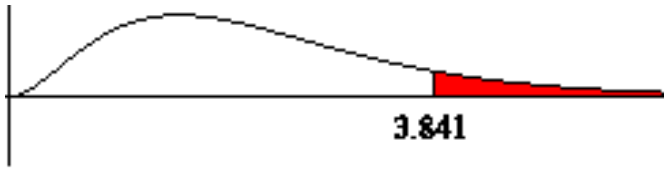
Example

- A = age of mother and B = birth weight

	≤ 2500g	> 2500g	
≤ 20	10	40	50
> 20	15	135	150
	25	175	200

Rejection Rule:

Reject if $\mathbf{X^2 > 3.841}$, $df = (2\text{-}1)(2\text{-}1)=1$



$$E_{00} = \frac{50 \cdot 25}{200}, E_{01} = \frac{50 \cdot 175}{200}, E_{10} = \frac{150 \cdot 25}{200}, E_{11} = \frac{150 \cdot 175}{200}$$

$$X^2 = \frac{(10-6.25)^2}{6.25} + \frac{(40-43.75)^2}{43.75} + \frac{(15-18.75)^2}{18.75} + \frac{(135-131.25)^2}{131.25} = 3.428$$

Bayesian Hypothesis Testing

- Compare the posterior probabilities of H0 and H1:

$$p(H_i \mid D) \propto p(D \mid H_i)p(H_i)$$

- Ratio factors into prior odds and likelihood ratio, called the **Bayes Factor:**

$$\frac{p(H_0 \mid D)}{p(H_1 \mid D)} \propto \frac{p(D \mid H_0)p(H_0)}{p(D \mid H_1)p(H_1)}$$

Overview of Data Mining Algorithms

- A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns.

Component View

- Data mining algorithm components:
  - task
  - structure or model
  - score function
  - search or optimization method
  - data management technique

Task

- visualization, classification, clustering, regression, anomaly detection

Structure

- functional form of the model or pattern we are fitting to the data
- e.g., linear regression model, hierarchical clustering model, neural network, Bayesian network, association rule

Score Function

- used to judge the quality of our fitted models or patterns based on observed data
- e.g., misclassification error, squared error, loglikelihood, etc.
- score function is what we try to maximize when we fit parameters
- May be based on *goodness of fit* alone (how well the model can describe the data) or can try to capture *generalization* performance (how well the model will describe data we have not yet seen)

Search Method

- The optimization method we use to search over parameters and structures, I.e., computational procedures and algorithms used to find the maximum (or minimum) of the score function
- if model structure is fixed, this is just search in parameter space
- if we consider a set of different structures, or a family of models, then must search over both structures and associated parameter spaces.
- **optimization is at the heart of any data mining algorithm**

Data Management Technique

- methods used for storing, indexing and retrieving data
- many statistical and machine learning algorithms do not specify any data management technique, essentially assuming that the data set fits in main memory so random access to any data point is free
- However for massive data sets that exceed capacity of main memory, accessing data is orders of magnitude slower, and the physical location of the data and manner in which it is accessed can be critical to algorithm efficiency

3 Algorithms

	CART	Backpropagation	A Priori
Task	classification and regression	regression	rule pattern discovery
Structure	decision tree	neural network	association rules
Score Function	cross-validated loss function	squared error	support/accuracy
Search Method	greedy search over structure	gradient descent on parameters	breadth-first with pruning
Data Management Technique	unspecified	unspecified	linear scans

Next Time

- Reading:
  - HMS, chapter 5, 6 cont.

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