CPH 636 – Spring 2009 – Dr. Charnigo

Lecture 1

Textbook definitions. The author of your textbook, George Fernandez, describes data mining as a "powerful information technology tool with great potential for extracting previously unknown and potentially useful information from large databases" (page 1). Hastie et al describe data mining as an endeavor in which an analyst must "extract important features and trends" from large quantities of data (Preface). [[In our course materials, I will sometimes refer to "Hastie et al", "Ripley", or "Stroup and Teutsch". The book titles are *The Elements of Statistical Learning* (2001, Springer-Verlag), *Pattern Recognition and Neural Networks* (1996, Cambridge University Press), and *Statistics in Public Health* (1998, Oxford). The first two books are very good but assume a strong background in mathematical statistics. The third book does not assume such a background.]]

Differences between data mining and "ordinary" statistics. So, apart from the sizes of the data sets, what distinguishes data mining techniques from the methods that you learned about in your other statistics courses?

1. <u>Modeling and prediction receive priority over hypothesis tests</u>. In data mining, there is an emphasis on discovering patterns and developing models with a view to making predictions. In your other statistics courses, the emphasis was usually on testing specific hypotheses or estimating measures of effect. For example, is there a nonzero difference in mean response between a treatment group and a control group, or is the odds ratio comparing exposed people to non-exposed people different from unity? Granted, you may have touched on issues of model se-

lection and prediction (e.g., calculating a "prediction interval" in linear regression, employing the AIC to judge between two competing logistic regression models, or using "backward elimination" to select explanatory variables in proportional hazards regression), but these were not focal points in your other statistics courses.

2. A bigger sample is not always better in data mining. This statement seems counterintuitive. Indeed, thinking back to your introductory statistics course and its methods for making inferences about a population mean, you will recall that the power to reject $H_0 : \mu = \mu_0$ in favor of $H_1 : \mu \neq \mu_0$ at significance level 0.05 is

$$\Phi\left[-1.96 + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right],$$

where Φ is the standard normal cumulative distribution function ("Z table"), n is the sample size, $\mu_1 (\neq \mu_0)$ is the true value of the population mean, and σ is the true value of the population standard deviation. Since μ_1 and σ are determined by nature, as it were, the only way that you can change the power is by changing the sample size. With a larger sample size, you have more power. In fact, this is a rather general phenomenon in statistical hypothesis testing. [[Students with a strong background in mathematical statistics will recall the term "asymptotic consistency", a description applied to a hypothesis testing procedure for which the power at any fixed point within the alternative hypothesis approaches 1 as the sample size approaches infinity.]] Thus, if your main objective is hypothesis testing and you have 5000 observations available, you should use all 5000 of them to carry out the hypothesis test. *Yet, I will argue in Lecture 2 that, if you have* 5000 *observations available in a data mining problem, you should not use all* 5000 *of them to fit a statistical model!*

3. You will encounter new data analysis techniques. While some of the methods that you have encountered in your other statistics courses are useful in data mining (especially linear regression and logistic regression), in this course you will encounter a variety of techniques that I suspect most of you have not seen before. These will include discriminant analysis, trees, neural networks, nearest neighbor methods, principal components, and clustering.

Sketches of data mining techniques. Today I will provide very brief sketches of those data mining techniques that I suspect most of you have not seen before. Each of these techniques will be the subject of a full lecture later in the semester.

1. Discriminant analysis. The response variable defines two or more categories into which observations may be classified. You want to develop a rule for predicting into which category a new observation falls. Discriminant analysis constructs a rule based on functions of the explanatory variables called discriminants.

2. Trees. The response variable can be either continuous or categorical. A picture is worth a thousand words:

3. Neural networks. The response variable can be either continuous or categorical. Again, a picture is worth a thousand words:

4. Nearest neighbor methods. The response variable can be either continuous or categorical. You predict the response for a new observation based on the responses for previous observations that were "similar" on the explanatory variables.

5. Principal components. You have a large collection of variables, but none of them is a response variable *per se*. You want to condense this large collection into (perhaps) two or three new variables that capture the "essence" of the original collection.

6. Clustering. You have a large collection of variables, but none of them is a response variable *per se*. You want to group similar observations into clusters. This is different from discriminant analysis because there are no pre-existing categories determined by a response variable.

Applications are ubiquitous. Here are several examples of data mining applications in public health, medicine, and other areas.

1. (Ripley, page 14) Within a specific population, discover how variables like plasma glucose concentration, diastolic blood pressure, body mass index, and age relate to diabetic status.

2. (Hastie et al, page 1) Identify risk factors for prostate cancer using clinical and demographic variables.

3. (Hastie et al, page 1) Predict whether a heart attack patient will have another heart attack using demographic, diet, and clinical variables.

4. (Fernandez, page 4) Predict which customers will buy new health insurance policies and identify patterns that characterize fraudulent behavior.

5. (Hastie et al, page 5) In cell samples taken from cancer patients, look for patterns in gene expression to determine which cell samples are similar to each other.

6. (Hastie et al, page 1) Predict the price of a stock in the near future using economic and company-specific variables.

7. (Hastie et al, page 2) Discover patterns of words in e-mail messages for the purpose of automatically filtering spam e-mail from real e-mail.

8. (Fernandez, page 4) In marketing, identify products that are often purchased together.

Sources and examples of public health data. Stroup and Teutsch (Chapter 3) identify several "traditional" sources of public health data: vital statistics and census information, registries, public health surveillance systems, surveys of populations or providers, epidemic investigations, research studies, and evaluations of intervention programs. However, they are quick to point out that data maintained for other purposes sometimes warrant the attention of public health officials.

Here are a few specific examples.

1. Registry data. The Surveillance, Epidemiology, and End Result program of the National Cancer Institute provides data used to monitor trends in cancer and to develop priorities for cancer research.
2. Surveillance data. In the National Notifiable Disease Surveillance System, physicians and other providers report cases of specified health conditions.
3. Survey data. The Behavioral Risk Factor Surveillance System of the Centers for Disease Control and Prevention provides data for prevention activities.
4. Alternative data. The Hazardous Materials Information System of the Department of Transportation meets a public health need, as the reporting of spills in interstate commerce relates to the monitoring of environmental hazards.

Supervised learning versus unsupervised learning. Most applications of data mining techniques, as well as the data mining techniques themselves, fall on one or the other side of a paradigm that distinguishes scenarios with well-defined response variables from those without. Quoting Hastie et al (Preface), "In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures."

Discussion question. Consider the applications enumerated on page 5 of this lecture. Which applications exemplify supervised learning? unsupervised learning?

Challenges in data mining. Fernandez identifies some challenges in data mining, which I enumerate below. Interestingly, the sizes of the data sets and the availability of computational resources are not the main challenges (page 2).

1. While your emphasis may not be on testing specific hypotheses, you must nevertheless have a clear goal when data mining; otherwise, your efforts may be misdirected and, hence, not fruitful (page 5).

2. When data are entered by different people, and especially when data are drawn from several sources, there can be inconsistencies in format or coding (page 6). For example, in the absence of a clear uniform protocol for data entry, some people may use "0 = No, 1 = Yes" for a dichotomous variable, while others may use "0 = Yes, 1 = No"; some

people may use "999" for a missing value, while others may leave a blank; some people may use the metric system, while others may use the English system. If unrecognized and unaddressed, such inconsistencies can render your data mining ineffective.

3. Even if there are no inconsistencies in format or coding, there may still be gross mistakes or other anomalies in the data set. If unrecognized and unaddressed, such mistakes and anomalies can render your data mining ineffective (page 6).

Requirements for data content and organization. Ideally you want individual-level information rather than summary information (page 15), as summary information is of limited value for making predictions and is of almost no value for perceiving patterns. Reading the data into SAS will be facilitated if the data are organized into a matrix whose rows correspond to individual observations and whose columns correspond to variables (page 16).

SAS and the EXCELSAS macro. Fernandez can't give enough praise to SAS Institute and its products (page 10), but he makes a good point that the Enterprise Miner software may be prohibitively expensive for some people and organizations (page 11). For this reason, he has written a number of macros for data mining (page 12). We will use many of them this semester, beginning today with the EXCELSAS macro. I have written some detailed notes in {SASMacroInstr.txt} to help you get started with EXCELSAS.