DUKE UNIVERSITY
The Fuqua School of Business

MGRECON 491 Data Mining — Gallant
Final Exam — Term 1
Due October 17 , 2008

Prepare a PowerPoint presentation that addresses the following questions. The best way to organize your thoughts is to think of this as a presentation you will be giving to a client who needs to understand the skills involved in data mining in order to evaluate employee or vendor performance. The overall limit is 40 slides.

1. Define data mining. Be sure that your definition makes a distinction between the methods and goals of data mining and the methods and goals of traditional statistical inference. (At most two slides.)
2. Give an example of a application of data mining that has strategic value to a firm. This example can come from your own experience, your own independent reading, one of the examples from the pre-assignment (Thomas H. Davenport, "Competing on Analytics," *Harvard Business Review*, January 2006) or from the cases at the end of the text (Michael J. A. Berry and Gordon Linoff, *Mastering Data Mining*, Wiley, New York). Make sure that the example describes a specific data mining task, its strategic value, the tools used, and the conclusions reached. (At most five slides.)
3. Describe the main divisions of the subject. Give a concrete example of each. List the tools used within each. (At most two slides.)
4. Describe how each of the following tools work: Linear regression, decision trees, neural nets, logistic regression, principal components, discriminant analysis, hierarchical clustering, K-means, and association rules. (At most two slides per tool.)
5. Define a feature, a derived feature, a dummy variable, and an interaction. List the prediction tools that are automatic feature finders and those that are not. (At most two slides.)
6. Describe boosting and give one example. (At most two slides.)

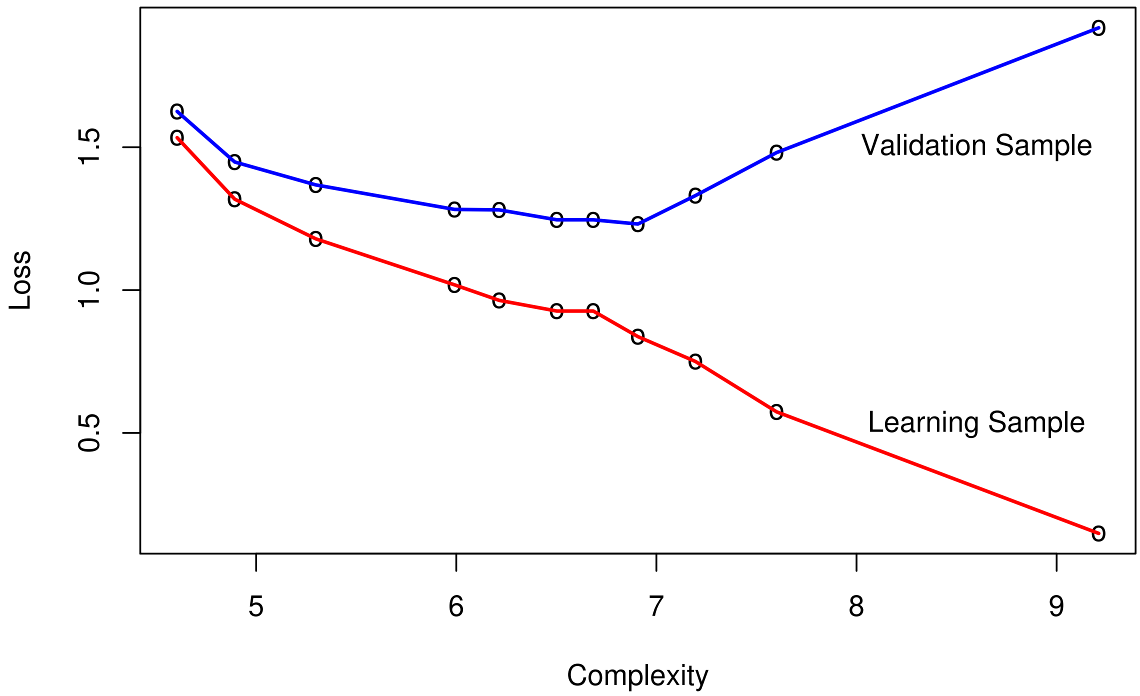

Figure 1 Learning and Validation Loss. Every tool will produce a picture like this as complexity increases: Loss is underestimated in the learning sample and correctly estimated in the validation sample. Complexity in regression increases when a variable is added; in nearest neighbors when the number of neighbors decreases, in a neural net when a hidden layer is added, in a decision tree when a leaf is added. Correct complexity is where validation loss is minimized.

7. Explain the notion of a loss function and loss. Distinguish between population loss and sample loss. Illustrate with a specific example such as mean squared error or classification error rate. (At most two slides.)

8. Describe how validation is used to select a model. Make sure the role of the training sample is explained and that you include a discussion of hold-out-sample validation, ten-part validation, and cross validation. (At most three slides.)

9. Explain why the loss in the training sample and the validation sample must behave as it does in Figure 1. In your discussion describe in as non-technical a way that you can the quantity s_{cf} that accounts for the difference between the two graphs in the case when the loss is mean squared error. (At most five slides.)

10. Describe the construction and use of a lift chart. (At most 2 slides.)