*These questions are similar in form and content to those that could appear on our first quiz. Note that I took these from prior quizzes, and in those classes we might have given different emphasis to different concepts, so do not scrutinize the exact questions too closely -- they are meant as a guide so you are not surprised by the sorts of questions that might appear and can prepare yourselves sufficiently.*

# Chapters 1 & 2 data analytic thinking, supervised vs unsupervised, the data mining process

**2.1 Multiple Choice**

In the following, choose the single best answer.

1) (True/False) We can build unsupervised data mining models when we lack labels for the target variable in the training data.

2) (True/False) For supervised data mining the value of the target variable is known when the model is used.

3) (True/False) Estimating the probability of a fraudulent transaction is an example of data mining.

4) (True/False) Finding the most profitable customer is an example of an unsupervised learning task

5) (True/False) Finding the characteristics that differentiate my most profitable customers from my less profitable customers is an example of an unsupervised learning task.

6) (True/False) Choosing which customers are most likely to leave is an example of the use of DM results.

7) (True/False) Discovering patterns of the defaults on auto loans is not an example of the model in use.

8) Which is <u>not</u> a reason why data mining technologies are attracting significant attention nowadays?
a) There is too much data for manual analysis
b) Data are difficult to transfer from databases
c) Data can be a resource for competitive advantage
d) Machine learning algorithms are easily available

9) Which of the following techniques is generally considered data mining?
   a) Data Warehousing
   b) On-line Analytical Processing
   c) Hypothesis Testing
   d) Linear Regression

10) Regression is distinguished from classification by:
   a) class probability estimation
   b) numerical attributes
   c) numerical target variable
   d) hypothesis testing

***Q) Label each case as describing either data mining (DM), or the use of the results of data mining (Use).***

***a) _____ Choose customers who are most likely to respond to an on-line ad.***
***b) _____ Discover rules that indicate when an account has been defrauded.***
***c) _____ Find patterns indicating what customer behavior is more likely to lead to response to an on-line ad.***
***d) _____ Estimate probability of default for a credit application.***
***e) _____ Predict whether a customer is pregnant***

**2.2 Short Answer**

In the following, give brief answers (at most 2 sentences per question).

11) What is a *leak* in predictive modeling? Are leaks really a problem? Give a brief example.

# Chapter 3 feature selection with entropy and information gain, tree induction

### 3.1 Multiple Choice

In the following, choose the single best answer:

1) (True/False) Induction reasons from general knowledge to specific facts.

2) (True/False) Estimating whether a review on Amazon is fake or not is an example of descriptive modelling.

3) Entropy
   a) is a measure of information gain
   b) is used to calculate information gain
   c) is a measure of correlation between numeric variables
   d) has a strong odor

4) (True/False) Classification trees cannot be extended to give us estimates of the probability of customer churn.

### 3.2 Short Answer

In the following, give brief answers (at most 2 sentences per question).

5) What does it mean for one attribute to give information about another attribute? Give an example of how one would find an attribute that gives information about another attribute.

# Chapter 4 linear discriminants, linear regression, logistic regression, SVMs

## 4.1 Multiple Choice

In the following, choose the single best answer:

1) (True/False) Support-Vector Machines (SVMs) approach classification problems by finding the widest possible bar that fits between points of two different classes.

2) Which of the following is not true about logistic regression:
   a) Logistic regression predicts probability of membership in a certain class.
   b) Logistic regression takes a categorical target variable in training data.
   c) A logistic regression represents the odds of class membership as a linear function of the attributes.
   d) Logistic regression requires numeric attributes and categorical attributes should be converted to numeric attributes.

3) Which of the following does not describe SVM (support vector machine)?
   a) SVMs are based on supervised learning
   b) SVM chooses the line to minimize the margin between two classes
   c) SVM can be applied when the data are not linearly separable

## 4.2 Short Answer

In the following, give brief answers (at most 2 sentences per question).

4) When we fit a model to data, we find the optimal model parameters. What does this mean?

### 4.3 Matching

In the following, choose the best matching for each set; each letter should be used once.

| | |
|---|---|
| __ Logistic regression | a. numerical target variable not bounded |
| __ Support Vector Machines | b. decision nodes |
| __ Linear Regression | c. log odds |
| __ Classification Trees | d. widest margin |

# Chapter 5 cv, overfitting

### 5.1 Multiple Choice

In the following, choose the single best answer:

1) (True/False) Cross-validation is used to estimate generalization performance

2) (True/False) Adding more complexity to a model will generally increase its performance on the training set.

3) (True/False) Pruning is a technique for reducing complexity.

4) (True/False) Complex models generally give better generalization performance than simple models

5) A fitting curve plots:
   a) True positive rate vs. false positive rate⍰
   b) True positive rate vs. false negative rate
   c) Generalization performance vs. size of training set
   d) Generalization performance vs. model complexity

6) Which is not a technique for reducing/avoiding overfitting in tree induction?
   a) choose largest improvement in information gain⍰
   b) stop growing tree based on the number of training examples at a leaf⍰
   c) select tree size based on validation data
   d) reduce tree size by cutting off branches and replacing them with leaves

7) Which is <u>not</u> a benefit of using cross-validation for model induction evaluation?
   a) It provides an estimate of generalization performance
   b) It provides statistics on estimated performance, so that we can understand how performance will vary across data sets
   c) It's quick to compute relative to other holdout methods
   d) It makes better use of limited data by using all of the data for both training and testing

8) Learning curves
   a) Are used to select an optimal parameter complexity
   b) Are equivalent to fitting curves
   c) Plot true positive rate vs false positive rate
   d) Can illustrate whether obtaining more data would be a good investment
   e) Are shown for a given amount of training data

9) Which is <u>not</u> a way of performing complexity control in logistic regression?
   a) pruning
   b) using domain knowledge to pick relevant attributes
   c) automated attribute selection using information gain
   d) including complexity penalization parameter in the objective function
   e) evaluating the model using cross validation

10) More complex models
   a) have better predictive performance
   b) tend to overfit more
   c) are easier to train than simpler models
   d) are very interpretable

### 5.2 Short Answer

1) Using a linear model that perfectly separates a set of data points with two labels is not always a good idea. Why is that? Give an example.

# Other Questions

By this point you should be able to formulate a supervised predictive modeling problem from a business problem. Revisit questions like:

**Q) MTC (MegaTelCo) has decided to use supervised learning to address its problem of churn in its wireless phone business. As a consultant to MTC, you realize that a main task in the business understanding/data understanding phases of the data mining process is to define the target variable. In one or two sentences, please suggest a definition for the target variable. Be as precise as possible—someone else will be implementing your suggestion.** ***(Remember: it should make sense from a business point of view, and it should be reasonable that MTC would have data available to know the value of the target variable for historical customers.)***