These questions are similar in form and content to those that could appear on our second 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.

Chapter 6 similarity, neighbours, clusters

Multiple Choice

In the following, choose the single best answer:

1) (True/False) Evaluation is more difficult for unsupervised data mining than supervised data mining

2) (True/False) kNN techniques are efficient in the "use" phase of predictive modeling.

3) (True/False) A 2-nearest neighbor model is more likely to overfit than a 20-nearest neighbor model (cf. Chapter 5).

4) Similarity measures are most essential for
   a) Naïve Bayes
   b) Tree Induction
   c) Hierarchical Clustering
   d) Logistic Regression

5) Which is not true of k-Nearest Neighbor (k-NN)?
   a) It can incorporate domain knowledge
   b) It builds a simple induction model
   c) It is robust to noisy data
   d) It is easy to explain how it works

Short Answer

9) Distance is a key notion underlying many data mining algorithms, such as k-nearest neighbor (k-NN). What problem is there with comparing

consumers using regular Euclidean distance, for example when they are described by age (in years), income (in dollars), and number of credit cards? How can this problem be fixed? See "Some Important Technical Details…" section

10) Similarity is a key notion underlying many data mining techniques. Assume that you are employed by Pandora to make music recommendations. Give Pandora an artist or song, and it will find similar music in terms of melody, harmony, lyrics, orchestration, vocal character and so on. If you plan to use k-NN algorithm to finish the music recommendation task, carefully describe how you would proceed. *Note: this question is not just about explaining how k-NN works. It is also asking about what information you'd use in your task, how you'd use it and so on.*

11) Similarity is a key notion underlying many data mining techniques. If you use Euclidean distance to find similar examples, how can you deal with categorical attributes? The k-nearest-neighbor technique estimates the target variable based on the k most similar examples. How exactly would you estimate the target variable for a regression problem? Explain the pros and cons of using different values for k, for example k=1 and k=N, where N is the total number of training examples. How would you choose k?
In answering "how would you choose k", you should present a *full* approach. You should answer this question as if you were asked it during an interview.

12) A key notion underlying k-means clustering and k-nearest neighbor methods is the same--what is it? Computing similarity between data instances.

Chapter 7 what is a good model

Multiple Choice

1) (True/False) The error rate of a classifier is equal to the number of incorrect decisions made over the total number of decisions made.

2) A binary classifier achieves 95% accuracy on a test set consisting of 95% positive and 5% negative instances. If we use the same classifier on a test set comprised of 50% positive and 50% negative instances, we expect to get:
   a) higher accuracy
   b) lower accuracy
   c) the same accuracy
   d) cannot be determined

Short answer

1) Two of your data scientists A and B are working on a project for preliminary screening of a population of people for the early detection of Provost's Quizinoma. Although very rare, this disease is deadly for the person bearing it if not identified in time, so your task is quite important. After preliminary screening, a $750 blood test can determine the presence of the disease with almost perfect accuracy. You decided to motivate your analysts by structuring their work as a competition: both data scientists A and B have to work independently on the problem and then present their results separately. After the competition period is over, on the test data, data scientist A reports 99.9% percent correctly classified instances from her model, while data scientist B reports only 86.3% percent correctly classified instances from his model. Describe carefully how you would determine which model is preferable? Illustrate with some hypothetical example numbers. [Think about the costs and benefits, not just the accuracies!]

2) In a classification application we are asked to predict whether kids are going to be infected with the flu virus during 2017 or not, and if yes vaccinate them against it. The vaccine costs $10. If a child is vaccinated, there is only a 10% chance that she will be infected. If a kid gets infected, the cost of treatment is about $1000. Write down the cost-benefit matrix for the problem.

	P	N
P	110	10
N	1000	0

The 110 for the "True Positive" comes from the 10% of $1000. Pay close attention to what the rows / columns mean in this case.

Matching

1)

c accuracy	a. TP/(TP+FP)
b recall	b. TP/(TP+FN)
a precision	c. 1 - (FP+FN)/(P+N)

Chapter 8 visualizing model performance

Multiple Choice

In the following, choose the single best answer:

1) The area under the ROC curve **is not**?
   a) equal to the Mann-Whitney-Wilcoxon statistic
   b) a measure of the quality of a model's probability estimates
   c) likely to be at least 0.5
   d) larger when false positive errors cost more

2) (True/False) Adding a budget constraint to the problem formulation might change the choice of the best ranking classifier.

3) (True/False) Each point in an ROC is a separate confusion matrix.

4) (True/False) A profit curve can assume negative values.

5) (True/False) If we create an ROC curve by sweeping a threshold from the test instances with the highest predicted classes to the lowest, then moving along this ROC curve from left to right, we will never move downward.

6) The points on a model's ROC curve
   a) represent the performance of different thresholds
   b) represent different rankings of examples
   c) represent the cost of different classifications

Short Answer

9 What exactly does the area under the ROC curve represent? Be as precise as possible.

The area under the ROC curve represents the probability that a randomly selected positive example will be ranked above a randomly selected negative example. This is the same as the Mann-Whitney-Wilcoxon statistic. (pages 219, 225 of the textbook).

10 Give a short example of using the same model used in different contexts with different thresholds to make different decisions.

An example would be a model that predicts the GPA a student will

achieve. When used used to evaluate prospective student applicants, we might want to set some threshold A to extend offers to the ones above that threshold. When used on current students, we might want to set a different threshold B to discover the students performing worst and offer them help (e.g. free tutoring).

11 Give two different reasons why using ROC curves can be more effective for assessing model quality than the percent of classifications that are correct (a.k.a. "vanilla" accuracy).

12 Last month your boss sent a mailing to 20,000 of your existing customers with a special offer on a Hoosfoos credeen. The response was exciting: 1% of them responded, which brought in $200,000 in revenue. She has now delegated to you the task of continuing the program, and has given you a budget of $10,000, which will allow you to target another 20,000 customers (out of your customer base of 100,000). You don't want to just target them randomly, as your boss did. You build a tree model and a logistic regression. Describe how to evaluate them as follows. Describe (a) the confusion matrix and (b) how you will fill it out for one of the models. Describe (c) the cost/benefit matrix for this problem, including the costs and benefits for this case. (d) Show the evaluation function you will use to compare your systems. (e) How do (a) and (c) come into play in this evaluation function?

Chapter 9

Short Answer

In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

1) Explain why Naive Bayes is naive.
   See the book..

2) Explain the meaning of each of the different terms in Bayes Rule. Describe one way that this rule is used for data mining.

Explanation of the different terms in chapter 9 of the book.

Chapter 10

Multiple Choice

1) (True/False) What is considered a stopword depends on the context of the textual data.

2) One key part of the data mining process is creating attributes to describe examples. In order to represent documents (such as emails) as examples, we create term (e.g., word) based attributes to describe the documents. Which of the following is not a common approach?
   a) whether or not the term appears in the document (binary attribute)
   b) term frequency (number of times term appears in document)
   c) term frequency/total number of terms in document
   d) term frequency times the term's frequency in the document corpus

Short Answer

3) The word 'good' does not always imply positive sentiment in a review. Give an example. Describe a way that we can circumvent this problem.
   An example is 'The movie was not good'. Using 2-grams instead we can catch 'not good' as a different feature than 'really good'. However this is approach is far from perfect - for example: 'I can not understand why people say the movie is not good'.