Statistics 202
Fall 2012
Data Mining
Practice Final Exam

Prof. J. Taylor

This quiz is closed book but you are allowed two double-sided pages of notes. Show your work for full credit. This exam has 14 pages. You are allowed to use a calculator.

I understand and accept the Stanford University Honor Code.
Name: ___________________________________
Signature: ___________________________________

1	
2	
3	
4	
5	
6	
Total	

Q. 1) (True or False)

(a) _______ Neural networks are often used for clustering.

(b) _______ A rule-based classifier is determined by a set of mutually exclusive rules.

(c) _______ Generally, the test error for a classifier is higher than its training error.

(d) _______ The silhouette statistic is used to measure the quality of a classifier.

(e) _______ An ROC curve can summarize the performance of any classifier.

(f) _______ Principal Component Analysis is scale invariant.

(g) _______ Single linkage hierarchical clustering usually produces more compact (or tight) clusters than complete linkage hierarchical clustering.

(h) _______ Grubbs' test is an outlier detection method designed to control the probability of false detection of an outlier.

(i) _______ Fitting a mixture model is a supervised learning problem.

(j) _______ All classifiers use the same measure of impurity to estimate the relevant parameters.

Q. 2) Below is a table of similarities between five points.

1.00	0.92	0.35	0.22	0.21
0.92	1.00	0.61	0.44	0.16
0.35	0.61	1.00	0.37	0.10
0.22	0.44	0.37	1.00	0.33
0.21	0.16	0.10	0.33	1.00

(a) Cluster these five points using complete linkage. Draw a dendrogram to depict the clustering you obtain.

(b) Some clustering algorithms only work on distance matrices. A standard way to convert similarity matrices S_{ij} to distances D_{ij} is as follows:

$$D_{ij} = 1 - S_{ij}.$$

What if you had used the conversion

$$D_{ij} = \sqrt{S_{ii} - 2S_{ij} + S_{jj}}?$$

Will this result in the same clustering?

(c) Repeat (b) when using group average linkage instead of complete linkage.

Work space for Q.2

Q. 3) In a classification problem with classes A and B, 2 continuous features and one binary feature, you decide to use a naive Bayes classifier. There are 80 observations in class A and 20 observations in class B.

The sample means and standard deviations for the continuous features are given below.

Group	Feature	$\widehat{\mu}$	$\widehat{\sigma^2}$
A	1	2	3
A	2	1	1.5
B	1	-1	2
B	2	0	2.5

For the binary feature, you observe

Group	0	1
A	70	10
B	10	10

(a) What is the rule you use to decide whether an observation is in class A or class B? Be as explicit as possible.

(b) Suppose that you actually did not observe the class labels but still believed that this naive Bayes model is appropriate. How might you estimate the parameters of the model?

(c) Outline an algorithm for estimating these parameters.

Work space for Q.3

Q. 4) Suppose that your friend, a postdoc in biology, asks for your help in understanding an algorithm to detect some *novel* cancer cell types based on some point data in p dimensions that might represent different features of each of a number of cells.

Your friend tells you:

> Novel cell types are characterized by being isolated from the generally large group of "normal" cells in the data. However, smaller (but not tiny) clusters of cells away from the large "normal" cluster are probably not *novel* cancer cell types.

Based on an some already published results, your friend also tells you:

> The probability I mistakenly label a truly non-novel cancer cell *novel* is about 1%, while the probability I label a truly novel cancer cell *novel* is 90%. I also know that, in my samples, 95% of the cells are not *novel* cancer cells.

(a) Suppose an experiment generates 10000 cells. Write down an "expected" confusion matrix for this algorithm.

(b) If *positive* is *novel*, what is the TPR (true positive rate) of this particular algorithm?

(c) If you were given the data, i.e. a data matrix $\boldsymbol{X}$ along with labels $\boldsymbol{Y}$ indicating whether an observation is novel or not, how might you find a model that would allow you to find *novel* cancer cells?

Work space for Q.4

Q. 5) Consider the three classification methods: support vector machines, random forests and boosting. Describe each method in a few sentences. What are the major differences between them?

Work space for Q.5

Q. 6) Consider clustering a dataset $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. Suppose that for some subset $S \subset \{1, \ldots, n\}$ of the observations, we observe labels that divide the X's L different classes.

(a) Describe the K-means algorithm to cluster $\boldsymbol{X}$ into K clusters.

(b) Can you propose a modification of the K-means algorithm that uses these labels in such a way that all observations from a given class remain in the same cluster and allows you to cluster the data into $K \geq L$ clusters.

(c) Would you expect this to always yield a better clustering than if the class labels are ignored?

Work space for Q.6

Extra space

Extra space