Final Exam for STAT 8090, Fall 2019, Dr. Y. Zhao

Note: (1) Show all your work; otherwise no credit will be given.

(2) Due date is 12/10/19 (Tuesday) at 5:00 PM.

(3) Please note that no exam papers will be accepted after the due time, and it will be counted as zero.

(4) You have to finish the final exam by yourself and cannot discuss with anyone else, neither check answers with anyone else.

(5) The final project report should be in the form as a formal report. It should include (1)-(7) of the project.

(6) Please show your codes in details.

Classification of vowel sounds

Aim: This project aims to study the classification of eleven vowel sounds with multivariate analysis methods.

Datasets: Both the training data and the test data contain 11 classes and 10 predictors. The classes correspond to 11 vowel sounds, each contained in 11 different words. Here are the words, preceded by the symbols that represent them:

Vowel	Word	Vowel	Word	Vowel	Word	Vowel	Word
i:	heed	O	hod	I	hid	C:	hoard
E	head	U	hood	A	had	u:	who'd
a:	hard	3:	heard	Y.	hud		

In the training data set ("vowel-train.txt"), each of eight speakers spoke each word six times. There are thus 528 training observations. In the test data set ("vowel-test.txt"), each of seven speakers spoke each word six times, and there are thus 462 test observations. The ten predictors $(x.1, x.2, \ldots, x.10)$ are derived from the digitized speech in a rather complicated way, but standard in the speech recognition world. The variable y is the class index for each observation.

1 [15 points] Conduct a principal component analysis for the training data set. Illustrate the percent that each eigenvalue contributes to the total sample variance. How many principal components will be enough to explain at least 90% of the total smaple variance in the training data?

2. [15 points] Suppose that you decide to use K principal components in part (1) to summarize the original training data. Get the scores of these K components for each observation, and then use these scores to conduct the linear discriminant analysis (LDA). What is the misclassification error rate based on the training data set? Apply the obtained linear discriminant rule to the test data set, what is the error rate?

3. [10 points] Repeat the work in part (2) and do the quadratic discriminant analysis (QDA). Does QDA give lower testing error rate?

4. [15 points] For the original training data set, conduct linear discriminant analysis and quadratic discriminant analysis, and get the error rates. Apply these rules to the test data and get the error rates. Summarize these error rates with those from parts (2) and (3) in a table and comment on the results.

5. [15 points] What classes do you think are most difficult to distinguish from others? If these classes are removed, how does the test error rates for LDA and QDA change? Compare with results from part (4).

6. [. : 20 points] To simplify it, select observations from classes 1, 3, 6 and 10 and conduct clustering analysis on these observations. Try the hierarchical clustering methods, K-means method and model-based clustering method. Comment on the performance of these methods.

7. [10 points] What is the final conclusion based on the previous methods?