CSE4305 Assignment 1

Classification and Clustering for the Microarray Data Analysis

Due 11:55PM, June 14, 2020

Part a

In this assignment you will be implementing 2 programs to analyze microarray data: (1) K-nearest neighbor classification (a supervised machine learning algorithm) and (2) K-means analysis (an unsupervised machine learning algorithm).

The ribosome is a large complex of many proteins that facilitates the translation of mRNA into protein; they are the cellular machinery responsible for linking together the correct sequence of amino acids from a sequence of codons. The cell often regulates the amount of protein by controlling the transcription level of the protein's gene.

Note: In this assignment, ribosomal proteins mean those proteins present in the ribosome located in the cytoplasm, not those proteins found in the mitochondrial ribosome. It is important to make this distinction since we do not expect the levels of mitochondrial and cytoplasmic ribosomes to be coordinated.

Performance measures:

Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)

Accuracy = (TP+TN)/total

What you will need to submit:

A source program for an executable file called "knn" (obviously the extension will vary by language) performs k-nearest neighbor classification with k and p (in that order) values specified either as command line arguments or others. The parameter p of the program knn takes values in [0, 1]. If (p*100) percent or more of the K neighbors are ribosomal genes, classify the unknown gene as ribosomal, otherwise classify it as non-ribosomal. Output should be displayed on screen and saved in a file called "knn.out" in the format shown below.

```
k: 10
p: 0.75
sensitivity: 0.85
specificity: 0.92
accuracy: 0.68
```

A second source program for an executable file called "kmc" performs k-means clustering. You will use this to answer questions, and we will check code and comments and make sure it runs.

Data

The data consists of 79 microarray experiments where the expression levels of 2,467 genes in *S. Cervisiae* (baker's yeast) were measured across 79 different conditions or time points. Of the 2,467 genes, 121 were previously characterized as ribosomal genes -- the remainders are various other genes with known functions. The 79 experiments include measurements of expression taken after environmental changes were imposed on the yeast. For example, some of these conditions were starvation (causing the yeast to form spores), changing the sugar supply (causing the yeast to ferment rather than respire), and synchronizing the cells to force them to pass through the stages of cell division at the same time. The experiments are described in the following paper:

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. Cluster analysis and display of genome-wide expression patterns. PNAS. 95:14863-14868.

The expression patterns for the 121 ribosomal genes can be found in the tab-delimited data file ribo-data.txt. Each column in the file represents one experimental condition (there are 79 columns in all). The expression patterns for the 2,346 non-ribosomal genes can be found in the data file nonribo-data.txt. The columns in this file are organized the same way, with 79 columns corresponding to the same order of experiments as in ribosome file. The file experiments.txt lists to which experiment each column corresponds. The common gene name and a short functional annotation for both the ribosomal set and the non-ribosomal set can be found in the files ribo-names.txt and nonribo-names.txt respectively.

K-nearest neighbor classification with microarray data

Implement the K-nearest neighbor strategy discussed in class. The basic idea is the following: given an expression vector for an unclassified gene, find the K "closest" expression vectors from the classified genes and let them vote on the classification for the query gene. If some p percent or more of the K neighbors are ribosomal genes, classify the unknown gene as ribosomal, otherwise classify it as non-ribosomal. Your function should take in a positive training set, a negative training set, a test set, a positive integer for K, and a value between 0 and 1 for p. It should return a vector of predictions -- one prediction for each row in the test set. Use Euclidean distance as your distance metric.

Cross-Validation

To assess the accuracy of a classifier, we typically use an approach known as n-fold cross-validation. In this assignment, you are to do 6-fold cross-validation. For the 6-fold cross-validation,

1. Divide the training set into 6 groups (Divide the positive set into 6 groups and the negative set into 6 groups).
2. From the files ribo.dat and nonribo.dat, you have the positive and negative sets. Divide both the positive and negative sets into 6 groups. Remove group 1 from both positive and negative sets, and use that as your test set against the remaining 5 groups. Then do that with the second group, third, etc.
3. Repeat step 2 6 times.

To get the accuracy, sensitivity, specificity, you average the results you got from testing each of the 6 held-out groups (consisting of both positive and negative data points). You can calculate those values because you know the actual classification of each gene (you know which file you got it from).

K-means clustering with microarray data

Implement the K-means clustering method discussed in class. The basic idea of K-means clustering is the following:

- Choose an initial partition of the data into K clusters (the centers can be picked at random, or specified by the user).
- For each data point, assign it to a cluster such that the Euclidean distance from the data point to the center is minimal.
- For each cluster, recalculate the centers based on all the data points that belong to that cluster.
- Iterate until converge (no data points change cluster) or after a certain number of iterations (e.g. 50).

Your program should take a dataset, an integer for K, and a set of centers (or if not specified, pick the centers randomly), and return a vector indicating to which cluster each data point in the dataset belongs to.

Part b

Varying K in K Nearest Neighbors

With p=50%, report the cross-validation accuracy for the following values of K (accuracy =(TP+TN)/total):

1. K=1:

Answer:

2. K=5
Answer:

3. K=20:
Answer:

4. K=50:
Answer:

5. K=100:
Answer:

6. How is the performance affected by different values of K?
Answer:

7. Are there any non-ribosomal genes that are consistently misclassified as ribosomal? If yes, list the gene numbers.
Answer:

Part c

Varying p: Sensitivity versus Specificity Tradeoff

Sensitivity and specificity are usually estimated from performance on a test set as follows:
Sensitivity = (correctly classified ribosomal genes)/(all the ribosomal genes in the test set)
Specificity = (correctly classified non-ribosomal genes)/(all the non-ribosomal genes in the test set)

Using the value of K that gave you the best performance from **part b**, try the following different values of p and report the sensitivity and specificity your classifier obtains from 6-fold cross-validation:

If you can't see a change in sensitivity and specificity for the K that gives you the best accuracy, choose a different K that also has good accuracy.

8. P=5%:
Answer:
1. Sensitivity =
2. Specificity =

9. P=25%:

Answer:

1. Sensitivity =

2. Specificity =

10. p=50%:

Answer:

1. Sensitivity =

2. Specificity =

11. p=75%:

Answer:

1. Sensitivity =

2. Specificity =

12. p=90%:

Answer:

1. Sensitivity =

2. Specificity =

13. p=100%:

Answer:

1. Sensitivity =

2. Specificity =

14. What general trend does sensitivity follow with increasing p? (Choose all that apply)

a. decreasing

b. constant

c. increasing

15. What general trend does specificity follow with increasing p? (Choose all that apply)

a. decreasing

b. constant

c. increasing

16. When might you be more interested in having high sensitivity? (Choose all that apply)

a. if low false positive rate is desired
b. if low false negative rate is desired
c. if high true positive rate is desired
d. if high true negative rate is desired

17. When might you be more interested in having high specificity? (Choose all that apply)
a. when low false positive rate is desired
b. when low false negative rate is desired
c. when high true positive rate is desired
d. when high true negative rate is desired

K-means clustering with microarray data.

18. K=2. Pick the first data point in both ribo.txt and nonribo.txt as your starting centers. Are all the ribosomal genes in the same cluster?
a. Yes
b. No

19. If your answer to the previous question is yes, you can skip this question.
If your answer to the previous question is no, list all the ribosomal genes that are in the cluster that is different from the majority of the ribosomal genes (by their index in ribo.txt. The first gene in that file is indexed as number 1).
Answer:

20. What percentage of genes in each cluster are ribosomal genes? (enter two % values, separated by a comma)

Answer:

21. K=2, choose two random data point as your starting centers.
What percentage of genes in each cluster are ribosomal genes? (enter two % values, separated by a comma)
Answer:

22. Comparing your results from choosing the first data point as the starting centers with those from choosing two random data point as the starting centers, are the clustering assignments for each gene the same?
a. Yes

b. No

23. What can you say about K-means clustering based on the question 22?

Answer:

24. Do K-means clustering on the same dataset for 20 times with K=2 and random starting centers. Are there any ribosomal genes that are often clustered into a different cluster from the majority of the ribosomal genes?

a. Yes

b. No

25. If there are not, you can skip this question. If there are, specify their index numbers.

Answer:

Submit:

- Answers to the short questions 1-25, source code, executable code and a readme file for each of K-nearest neighbor classification and K-means clustering to e-class.
- In the source code include comments and explanations of what is going on, especially if you expect partial credit for "almost working" code.
- In the readme file, include
  - Language used and instruction on how to run your program
  - Pseudocode description of the algorithm (one page or less).