**CSE3315 Assignment 2  
Classification and Clustering for the Microarray Data Analysis**

**Answers**

12191656 이채연

**Part b**

**Varying K in K nearest neighbor classification**

1. K=1  
Answer: 0.97

2. K=5  
Answer: 0.97

3. K=20  
Answer: 0.97

4. K=50  
Answer: 0.97

5. K=100  
Answer: 0.98

6. How is the performance affected by different values of K?  
Answer: K=1일때 K=5일때 K=20일때 K=50일때는 accuracy에서 큰 차이가 없지만 K=100일때 accuracy가 조금 높아진 것을 알 수 있다.

7. Are there any non-ribosomal genes that are consistently misclassified as ribosomal? If yes, list the gene numbers.  
Answer: Yes / 814, 846, 847, 853, 854, 1440, 1451, 1454, 1459, 1460, 1462, 1464, 1466, 1467, 1470, 1471, 1472, 1474, 1475, 1480, 1484, 1495, 1501, 1508, 1509, 1511, 1512, 1513, 1515, 1516, 1517

**Part c**

**Varying p: Sensitivity versus Specificity Tradeoff**

8. P=5%  
1. Sensitivity = 0.99  
2. Specificity = 0.93

9. P=25%  
1. Sensitivity = 0.95  
2. Specificity = 0.96

10. P=50%  
1. Sensitivity = 0.85  
2. Specificity = 0.98

11. P=75%  
1. Sensitivity = 0.71  
2. Specificity = 1.00

12. P=90%  
1. Sensitivity = 0.15  
2. Specificity = 1.00

13. P=100%  
1. Sensitivity = 0.00  
2. Specificity = 1.00

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

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

16. When might you be more interested in having high sensitivity? (Choose all that apply)  
Answer: b, c

17. When might you be more interested in having high specificity? (Choose all that apply)  
Answer: a, d

**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?   
Answer: b

19. K=2. 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: 121

20. K=2. What percentage of genes in each cluster are ribosomal genes? (enter two % values, separated by a comma)  
Answer: 21.66%, 0.05%

21. K=2, choose two random data point as your starting centers.  
What percentage of genes in each cluster are ribosomal genes? Answer two % values, separated by a comma.  
Answer: 20.80%, 0.05%

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?  
Answer: b

23. What can you say about K-means clustering based on the question 22?  
Answer: Starting point를 무엇으로 하는지에 따라서 clustering 결과가 조금 달라질 수도 있다.

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?  
Answer: a

25. If there are not, you can skip this question. If there are, specify their index numbers.  
Answer: 121