COSC 681 Bioinformatics

Project 3 Gene Expression Analysis

2016/12/15

William Gillespie

**Part 1: Supervised Classification**

1. **Apply holdout method to divide samples into training set and testing set.**

I chose 3 different holdouts for the train/test parts: 50/50, 65/35, 75/25. I chose these because they each give a decent amount of training data while still having at lease 10 samples in the test data.

1. **Apply SAM to select informative genes**

I used the MeV software to select 78 informative genes. I chose 78 because that was the lowest number above 75 that I could select with the slider bar. The requirements were to select around 25-75. I wanted to select as many as possible while still being around that range.

1. **Use KNN classification methods; repeat 3 times with different holdout**

I used the standard KNN classifier in Matlab. The code for this are in the Part1\_Supervised\_Classification directory under KNN\_Pearson and KNN\_Euclidean. I performed two distance metrics with the three holdouts totaling 6 different classification runs. The distances I chose were ‘correlation’ and ‘Euclidean’.

1. **Summarize your analyzing results – at lease errors and consistency**

I will list the error percentages in the following table:

|  |  |  |
| --- | --- | --- |
|  | Error Percentages for KNN |  |
|  | **Distance: Pearson** | **Distance: Euclidean** |
| Holdout 50/50 | 16.13 % | 19.35 % |
| Holdout 65/35 | 18.18 % | 18.18 % |
| Holdout 75/25 | 18.75 % | 18.75 % |

For the most part, these are fairly low error rates. I’m not sure what the optimal error rates are, but I think this is quite remarkable given that the data started with 2000 features.

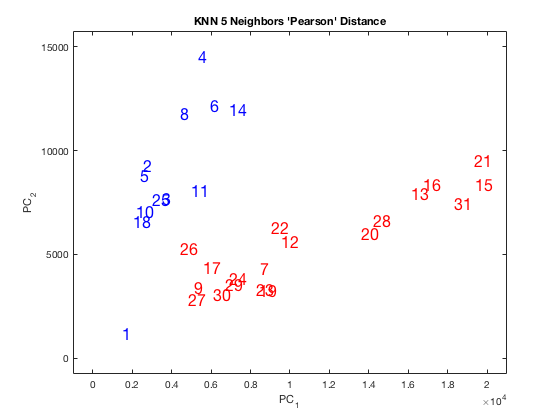
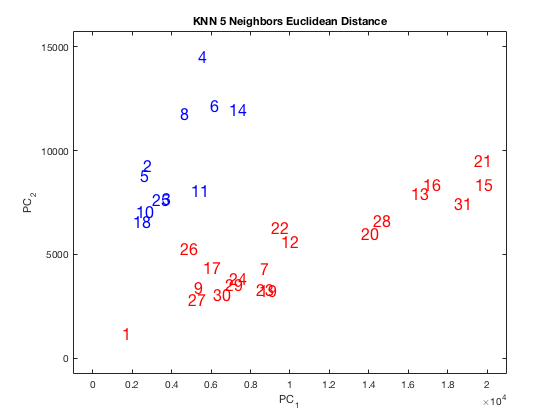
One can see that for the holdouts of 65/35 and 75/25, the Euclidean and Pearson distance measurements give the same error percentage. For the holdout of 50/50, Pearson gets a better error percentage. Such a difference is not that significant because it is due to one or two record misclassifications. With such a small test set, differences of the error percentages that small are not significant.

One thing that is strange is that for the Pearson distances, the error percentage goes up as the training data grows. This should not happen. If anything, the error percentage should go down as the training data grows. This is no doubt a symptom of the small test set. If the small test set has some data points that are hard to classify, then that will bump up the error percentage.

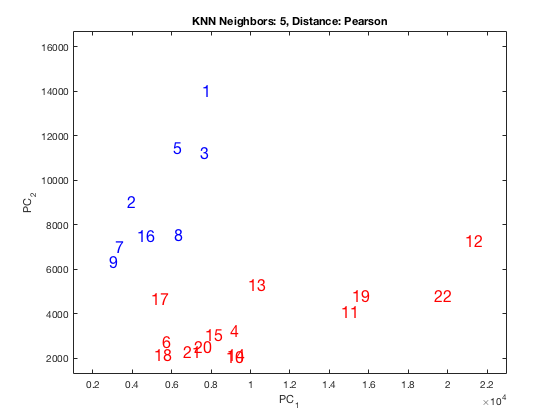
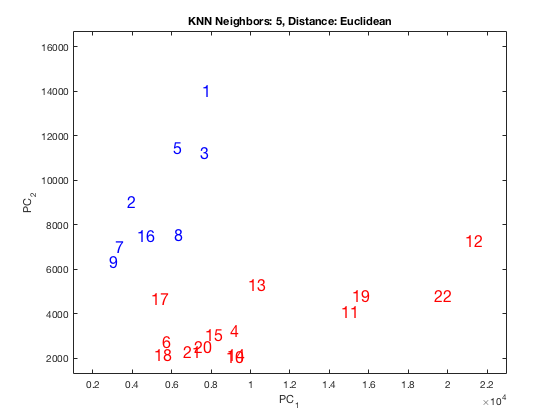
1. **Display at least tree view and colored scatter plot**

**Scatter plots (Euclidean on first column, Pearson on second column):**

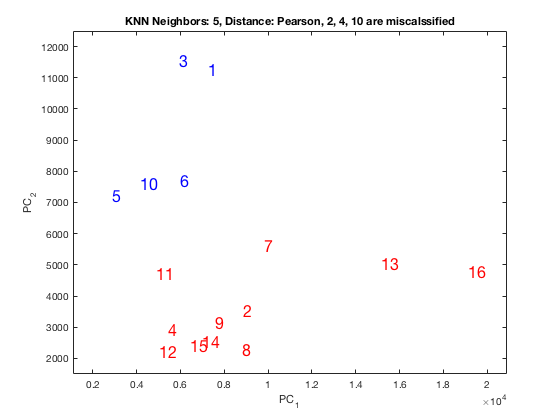
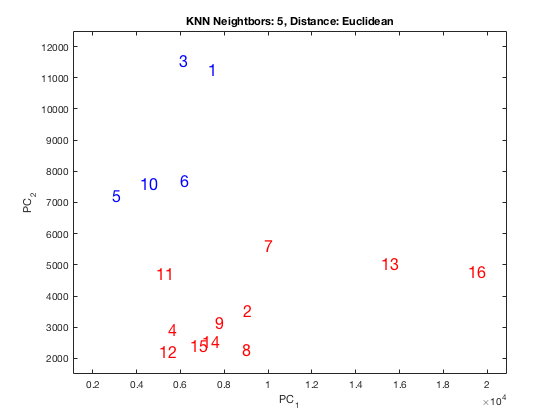
Holdout 50/50

****

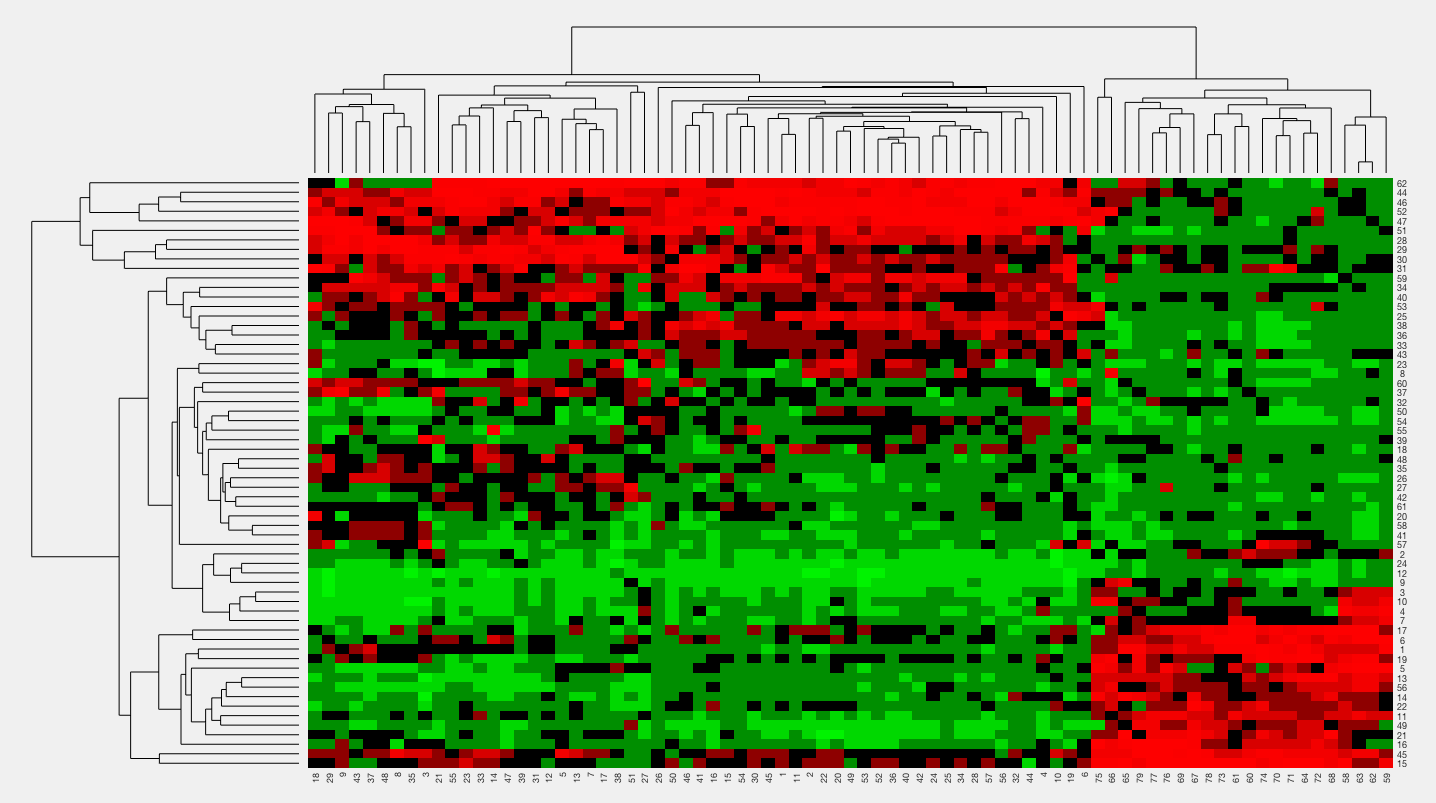
Holdout 65/35

****

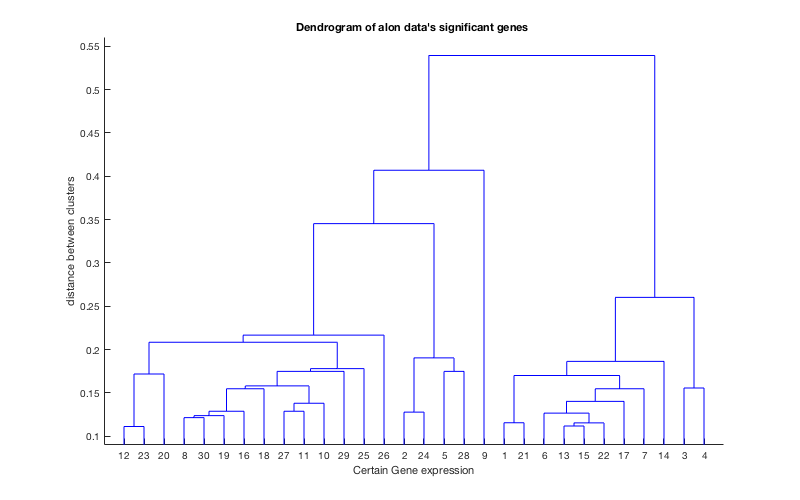
Holdout 75/25

****

**Tree View**



Dendrogram (Just for fun)



**Part 2: Unsupervised Clustering**