Homework 3

Wei Jiang

EN.550.440 Computational Molecular Medicine *Due: 03.13.2016*

Problem 1

(a) The Bayes classifier is $F_B(x) = \underset{l}{\operatorname{argmax}}\, p(Y = l|X = x)$.
Since in our case, Y doesn't depend on x, we have $F_B(x) = \underset{l}{\operatorname{argmax}}\, p(Y = l) = l_{larger}$, where l_{larger} is the class label which has largest proportion in our entire data sample.
The Bayes error rate is : $e(F_B) = E[\delta(y \neq F_B(x))] = p(y \neq F_B(x)) = \frac{\sum_{i=1}^{n} \delta(y_i \neq l_{largest})}{N}$ where N is the entire sample size.

For example we have 25 individuals labeled as 1 in our dataset. So the Bayes error rate would be $\frac{25}{50} = 0.5$

(b) The cross-validation error rate is 0.5 which is consistent with the Bayes error rate in last question. Please see **HW3_problem1.py** for code.

(c) The error rate is still 0.5. So it's consistent with e^*.Please see **HW3_problem1.py** for code.

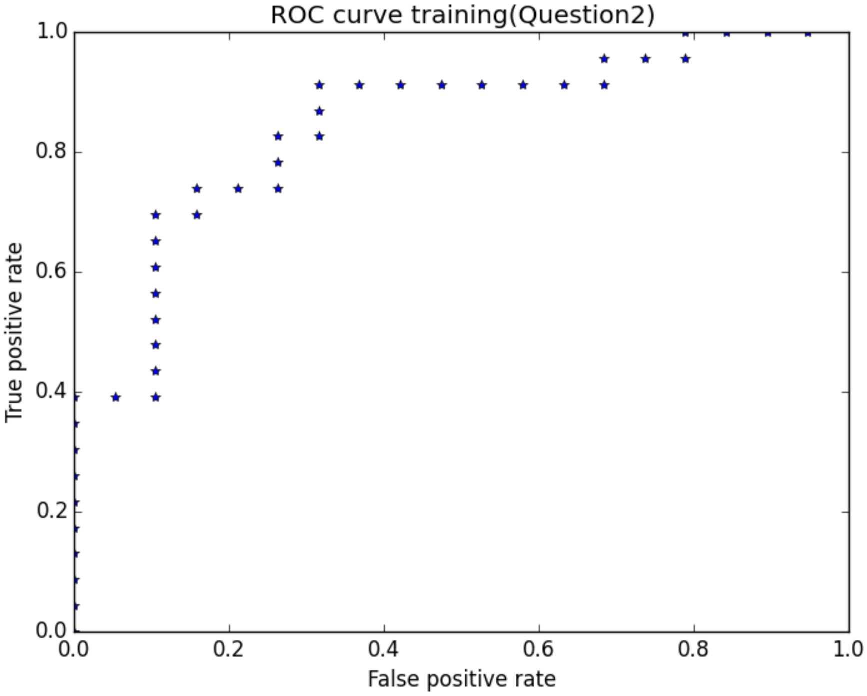
Problem 2

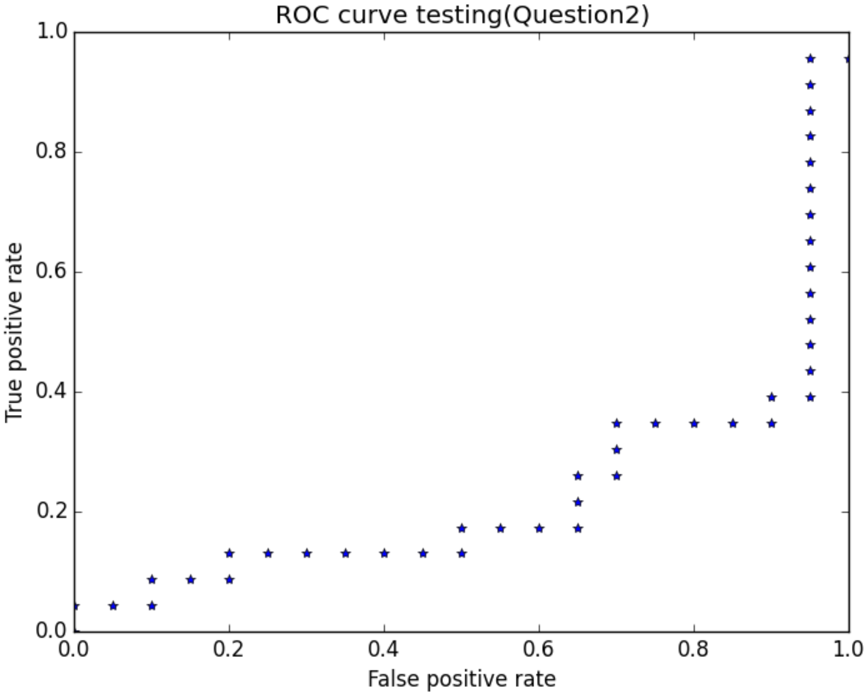
(a) Top 10 features are:
'ZRANB1' 'CDKN2C' 'GLUD2' 'C10orf4' 'LRRC27' 'GLUD1' 'HOXA3' 'CBARA1' 'HIF1AN' 'PTPN12'.

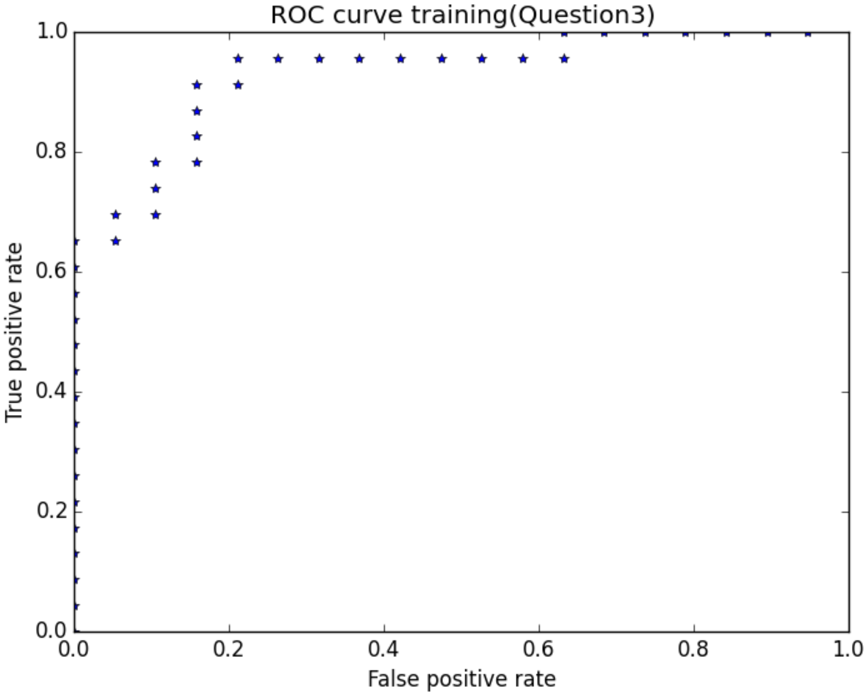
(b) The best classifier is as following:
Use feature 'CDKN2C' mRNA. If the feature value is larger than threshold 8.321, classify the patient as having class label 1. The error rate for training is 0.119

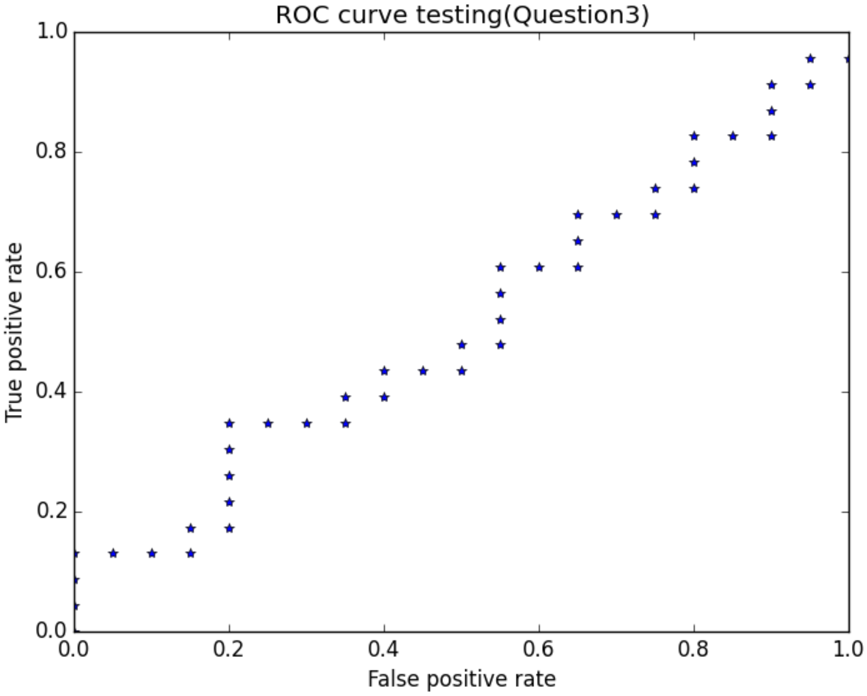
(c) The error rate on test data is 0.372.
The reason for the difference is that the classifier trained using the training data has high bias because the sample size of the training data is not big enough to represent the the true model or the test data. So the model basically suffers the problem of overfitting.

(d) The results are exactly the same as problem 2. The classifier is the same. The error rate for training and testing are also same with prblem 2.


ROC curve training(Question2)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve testing(Question2)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve training(Question3)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve testing(Question3)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Problem 3

(b) We used the entire dataset including the testing data to pick a set of the most significant feature. Then we search the most significant feature using only training data. This approach has no consistency. We should touch testing data during training. So the model has high performance on training data but poor performance on testing data.

(c) The difference is caused the model overfitted the training data because the training data could not represent the test very well.