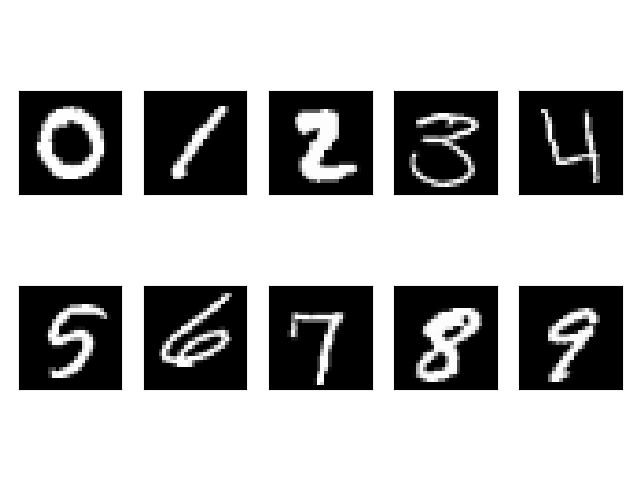
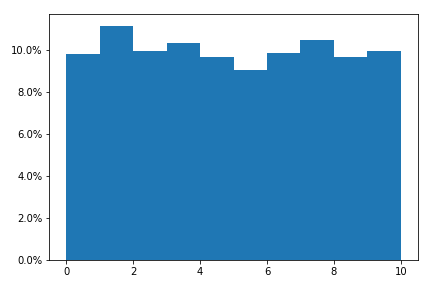
**Programming Exercise**

**1. Digital Recognizer**

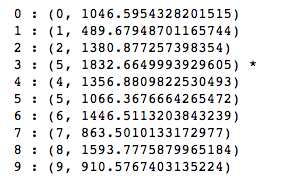
(a) (b) The following image displays one of each digits.



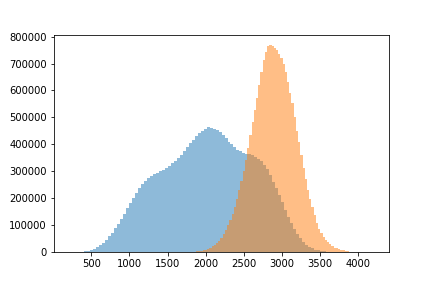
(c) It’s not uniform across the digits. Digit “1” is the most frequent and digit “5” is the least frequent.



(d) Among the example digits we chosen, nine out of ten digits have genuine match. But the nearest neighbor of digit 3 is digit 5.



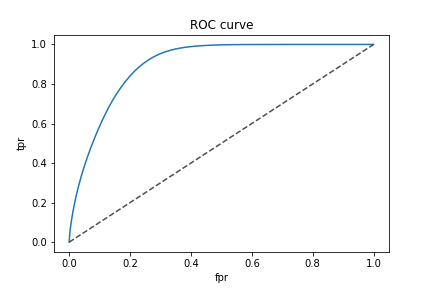
(e) The blue and yellow curve are the histogram of the genuine and imposter distances. On average, the imposter distance is larger than the genuine distance. But they also overlap quite a lot.



(f) The following graph is the ROC curve. Equal Error Rate is the point where false positive rate equals false negative rate, that is: . The EER=0.189







In the training data, we have 4684 digit 1 and 4132 digit 0. Because there are more digit 1s than digit 0s, an educated random guess will always be “1”, leading to a baseline error rate of 46.87%. The error rate of uneducated random guess is 50%.

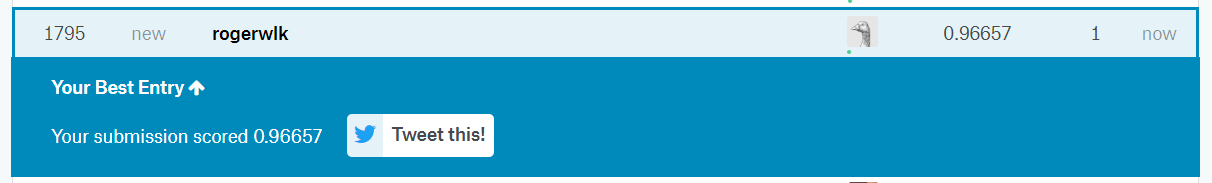
(g) (h) (i) We have implemented a classifier. The Confusion matrix is as follows. We also calculated the accuracy for each digit. The digit 8 has the lowest accuracy rate, which is 90.8%.



Rabel

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Actual  Predicted | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 2056 | 0 | 20 | 2 | 2 | 5 | 8 | 1 | 9 | 7 |
| 1 | 1 | 2324 | 43 | 16 | 18 | 13 | 5 | 37 | 32 | 11 |
| 2 | 1 | 2 | 1896 | 9 | 0 | 0 | 0 | 3 | 9 | 2 |
| 3 | 0 | 0 | 13 | 2126 | 0 | 22 | 0 | 0 | 41 | 18 |
| 4 | 0 | 3 | 4 | 3 | 1938 | 2 | 3 | 1 | 12 | 26 |
| 5 | 3 | 0 | 5 | 26 | 0 | 1745 | 17 | 1 | 40 | 6 |
| 6 | 8 | 0 | 4 | 4 | 9 | 21 | 2037 | 0 | 9 | 1 |
| 7 | 0 | 3 | 49 | 17 | 4 | 4 | 0 | 2103 | 6 | 39 |
| 8 | 0 | 1 | 8 | 15 | 0 | 3 | 0 | 0 | 1896 | 4 |
| 9 | 1 | 2 | 2 | 13 | 62 | 17 | 0 | 34 | 33 | 2004 |
| Accuracy | 0.993 | 0.995 | 0.928 | 0.953 | 0.953 | 0.953 | 0.984 | 0.965 | **0.908** | 0.946 |

(j) We got an accuracy of 0.96657 on the test data.



**2. Titanic**

(a)(b) 1. We choose variables sex, age, sibsp, and fare to conduct logistic regression.

2. We use median value of ages and median value of fares to fill out the blanks in the data. Because age and fare have lots of extreme values, median is more appropriate than average for the blank.

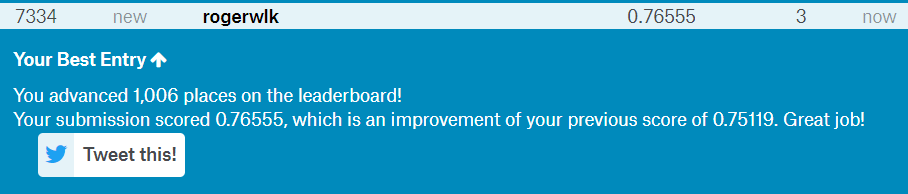
3. We convert sex to a dummy variable with 0 means ‘male’ and 1 means ‘female’.

4. We think that ticket class could be related to fare and fare provides more information, so we discard the “ticket class” variable.

5. We tried “parent children” variable and this variable is not significant at 95% level.

6. “cabin number” and “name” are not numerical value and each of them is different. They have no predictive value.

7. “port of embark” may have slight correlation to fare and this variable is not significant at 95% level.



**Written Exercise**

**Question 1**



**Question 2**

Define  as the real quality of widgets,  ;

Define  as the Test Result of widgets, .

Based on what we know about the quality control process and the widgets, we have:



(a) We need to calculate 

According to Byes rule:



Also,



We have:



(b) The probability of a widget is thrown away and is actually good, equals 



That is, about 499,995 good widgets are thrown away per year.

The probability of a negative is testing negative, which means it would be shipped to customers, but is defective is 



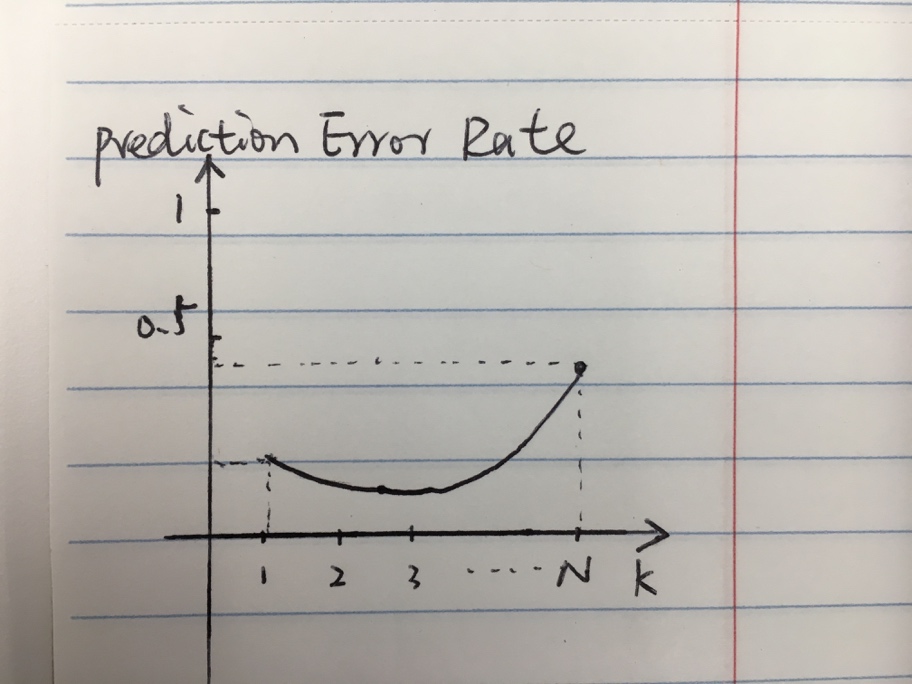
that is, about 5 bad widgets are still shipped to customers.

**Question 3**

(a) if equals 1 and we include  as part of the example training data, the prediction error will always be 0.

if  equals n, the majority of the example training data will always win. Let’s say the majority of training data is digit 1, then the predication error will be 

(b) If , variance of the test data is the primary reason the cause predication error. By increasing k slightly, we can reduce prediction error rate. In cases where is very large, the prediction error is close to the percentage of minor digits.



(c) When is large, the will include some neighbors that are relatively far away and are weak indicators of the class, leading to higher prediction error. One way to mitigate this is to weigh neighbors based on the distance. The larger the distance, the less relevant should the vote be.

(d) First of all, the assumption of classifiers is that features are independent. When the input dimensions increase, the probability that this assumption be violated is increased.

Secondly, high dimension input typically includes many irrelevant features, which are noise that make distance less reliable as indicator of similarity.

(e)

1. for the logistic regression

Confusion Matrix for the training data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Red | Predicted Blue | Total |
| Red | 6 | 3 | 9 |
| Blue | 3 | 8 | 11 |

Prediction accuracy is 14/20=0.7

Confusion Matrix for the test data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Red | Predicted Blue | Total |
| Red | 3 | 2 | 5 |
| Blue | 1 | 4 | 5 |

Prediction accuracy is 7/10=0.7

1. for the algorithm

Confusion Matrix for the training data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Red | Predicted Blue | Total |
| Red | 9 | 0 | 9 |
| Blue | 0 | 11 | 11 |

Prediction accuracy is 100%

Confusion Matrix for the test data:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Red | Predicted Blue | Total |
| Red | 1 | 4 | 5 |
| Blue | 2 | 3 | 5 |

Prediction accuracy is 0.4

1. For both the training data and test data, logistic regression outperforms . Explain why the difference occurs?



The 100% accuracy of  on training data means overfitting. On the test data, the variance becomes the dominate source of prediction error.