CS 5785: Applied Machine Learning **(Due: 09/26/19)**

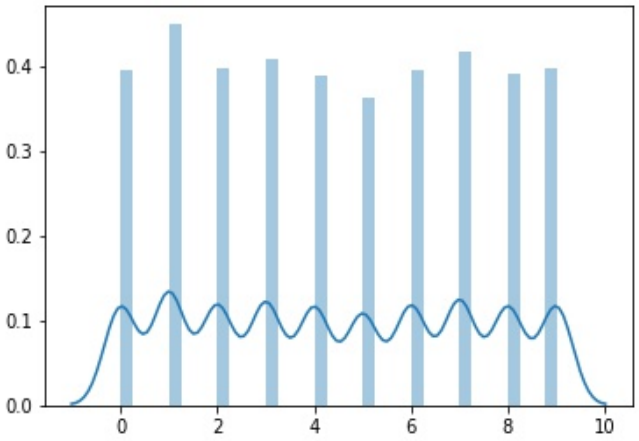
Homework #1

Instructor: Serge Belongie *Name:* Yanlin Chen, Yang Ma, *Netid:* yc2565, ym473

Part 1: Programming Exercises

1. (b) MNIST digits are displayed as follow.

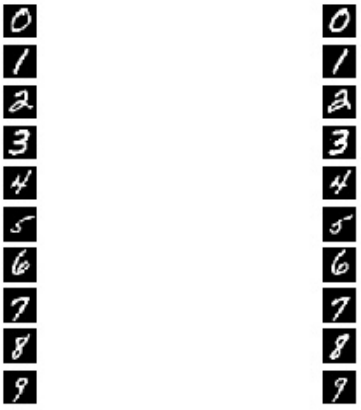
(c) The prior probability of the classes in training data is not uniform across the digits. It is not even.

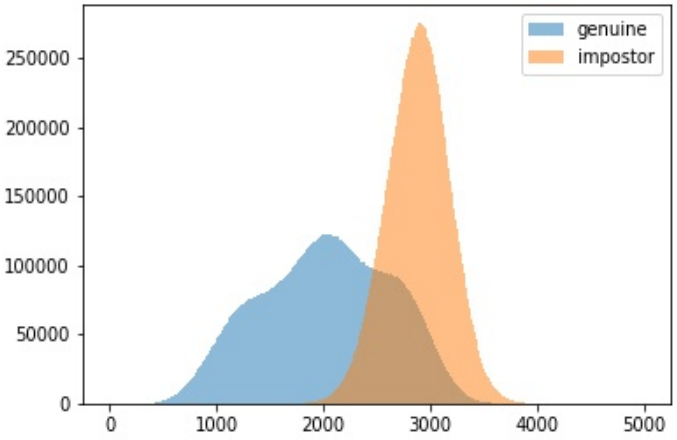


(d) The best match for each sample digit is as follow.

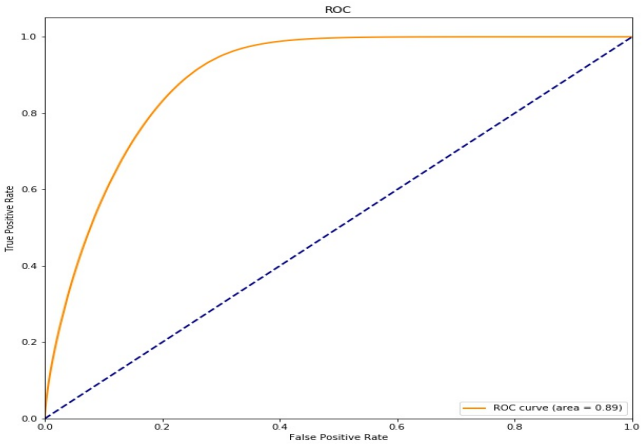
(e) The histograms of the genuine and impostor distances are as follow.

(f) The ROC curve is as following.
AUC: 0.8941213517035006
EER: 0.18898105539642254




genuine
impostor
250000
200000
150000
100000
50000
0
0
1000
2000
3000
4000
5000

When a classifier simply guesses randomly, then the error rate would be 0.5 (the worst)



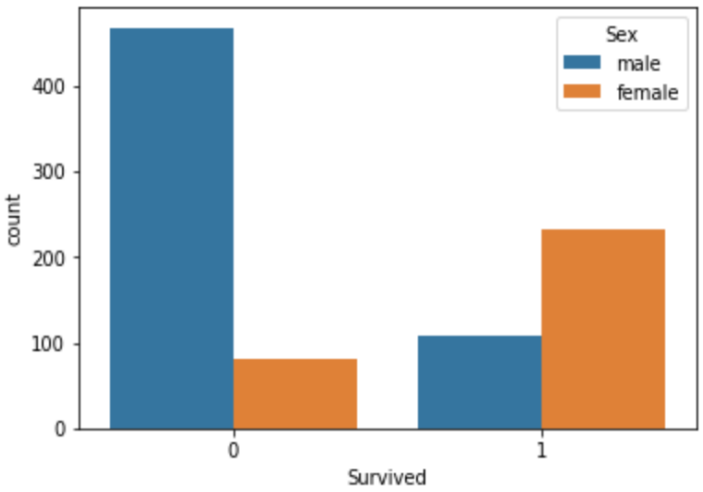
(h) With an increasing K from 1 to 10, the average accuracy firstly increase and then decrease. The best K in this experiment is 4.

K	Average Accuracy
1	0.9655
2	0.9655
3	0.9665
4	0.967666667
5	0.966366667
6	0.966233333
7	0.963533333
8	0.964133333
9	0.962466667

(i) The following confusion matrix is computed under k=3. According to this matrix, digits [8,9,2,3] are particularly tricky to classify.

```
[[2919    2    3    1    1    5    9    2    0    2]
 [   0 3332    9    0    6    0    2    7    1    3]
 [  25   27 2764   11    2    0    3   55    5    3]
 [   4   12   15 3000    1   35    2   22   22    8]
 [   3   25    1    0 2860    0    5    5    1   71]
 [  12    4    2   37    6 2574   26    4    8   11]
 [  18    6    0    0    3   11 2870    0    3    0]
 [   0   34    3    1    8    0    0 3080    3   43]
 [  10   32    9   38   14   47   15    7 2757   40]
 [   6    9    3   19   26   12    0   41    4 2853]]
```

2.
(b) We ignored Name and Ticket features and chose all of the other features, because the name of passengers and the ticket number are irrelevant to decide the probability of survival. However, other features impact the result. For example, after plotting a box graph of the number of males/females survived/died for training data, we found that more females survived and more males died.



The box plot of the number of people from different classes survived/died shows that passengers belonging to class 3 died the most.

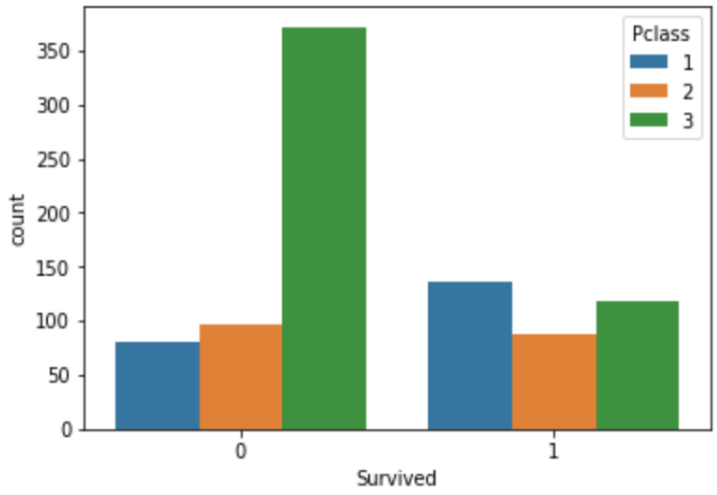
From the plot below, we can refer that people had 0 SibSp died the most.

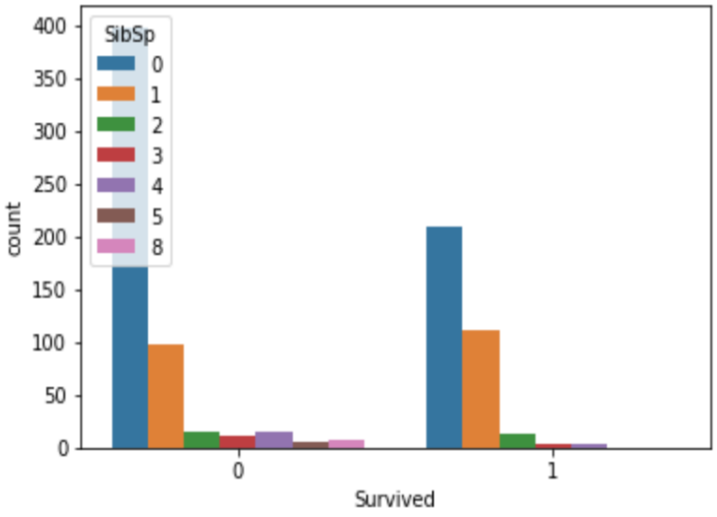
From the plot below, we can refer that people had 0 parch died the most.

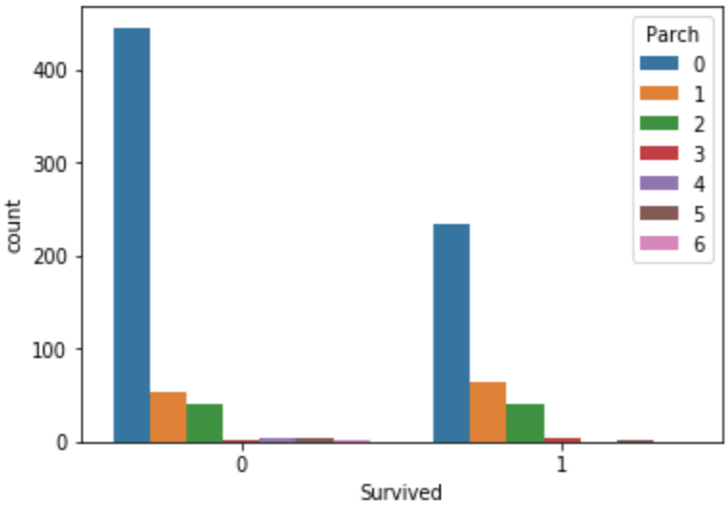
The other data of features could be visualized in the same way. From these plots, we can say that they are all relevant to the probability of survival.

Part 2: Written Exercises

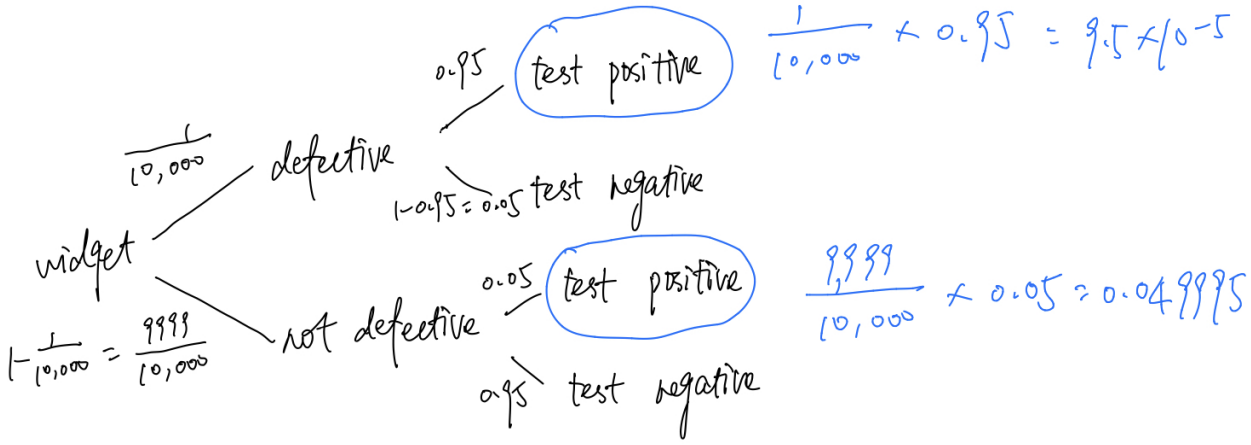
1. Based on the definition of variance,
$var(X-Y) = E[(X-Y-E[X-Y])^2] = E[(X-\mu_x-Y-\mu_y)^2] = E[(X-\mu_x)^2+(\mu_y-Y)^2+2(X-\mu_x)(\mu_y-Y)] = E[(X-\mu_x)^2+(\mu_y-Y)^2-2(X-\mu_x)(Y-\mu_y)]$
Now, applying the linearity of expectation, we have
$var(X-Y) = E[(X-E[X])^2] + E[(Y-E[Y])^2] - 2E[(X-\mu_x)(Y-\mu_y)] = var(X) + var(Y) - 2cov(X,Y)$


Pclass
1
2
3
count
0
50
100
150
200
250
300
350
0
1
Survived


SibSp
0
1
2
3
4
5
8
count
0
50
100
150
200
250
300
350
400
0
1
Survived

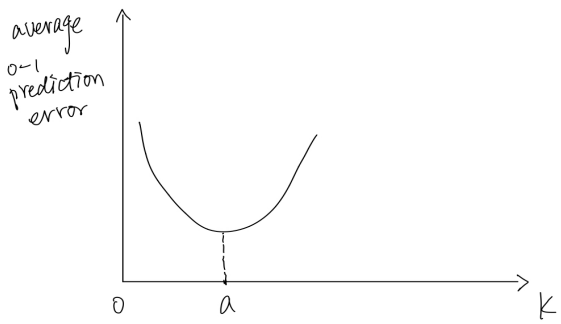

Parch
0
1
2
3
4
5
6
count
0
100
200
300
400
0
1
Survived

2.
(a). B: the widget is actually defective, A: the test shows a widget is defective
Given the formula of conditional probability:
$p(B|A) = p(B)p(A|B)/p(A)$
Based on the given information,
p(B) = 1/100,000, p(A—B) = 0.95
To compute p(A), we drew a graph shown below.



There are possibilities that a widget can be detected as positive, which are the widget is actually defective and the widget is actually not defective. Based on the graph, p(A) = 9.5e-5 + 0.049995. Then $p(B|A) = (1/100,000 * 0.95)/(9.5e-5 + 0.049995) = 1.8966e-4$.
(b). From the graph shown above, the probability that good widgets are tested as defective(positive) is 0.049995. The probability that a defective widgets are tested as good(negative) is $1/10,000 * 0.05 = 5e-6$. Therefore, the number of good widgets are thrown away per year is $10million * 0.049995 = 499950$. The number of bad widgets are still shipped to customers each year is $10million * 5e-6 = 50$.

3.
(a) When k gets smaller, the average 0-1 prediction error will increase but when $k = 1$, the average 0-1 prediction error will be zero.
(b) When k = n, the average 0-1 prediction error will be very large because the model will be too simple and ignore the useful information. Let's say the appropriate k should be a. When $n < k < a$, the average 0-1 prediction error will get smaller and smaller. When k = a, the average 0-1 prediction error will be minimized. However, when $a < k < n$, the average 0-1 prediction error will increase because of over-fitting and noise. The model will be too complicated at this time. The graph of the average 0-1 prediction error is roughly U-shape.



(c) We recommend 7. Every bucket has about 14% of data. The data of every bucket is not extremely large or small. Therefore, it's less likely to have much noise and the over-fitting. Moreover, the computation is not very complex because every time 7 buckets need to be computed.
(d) Neighbors that are closer to the data point weights more. For example, when k = 2, a and b are the nearest

neighbors of c. Moreover, a is closer than b and they are not from the same class. Then, we should predict c as a member of a's class because a is closer than b.
(e) 1. KNN is a "lazy learner" that does not learn a discriminative function from the training data.
2. Euclidean distance does not work well in high dimensions because all vectors are almost equidistant to the search query vector.