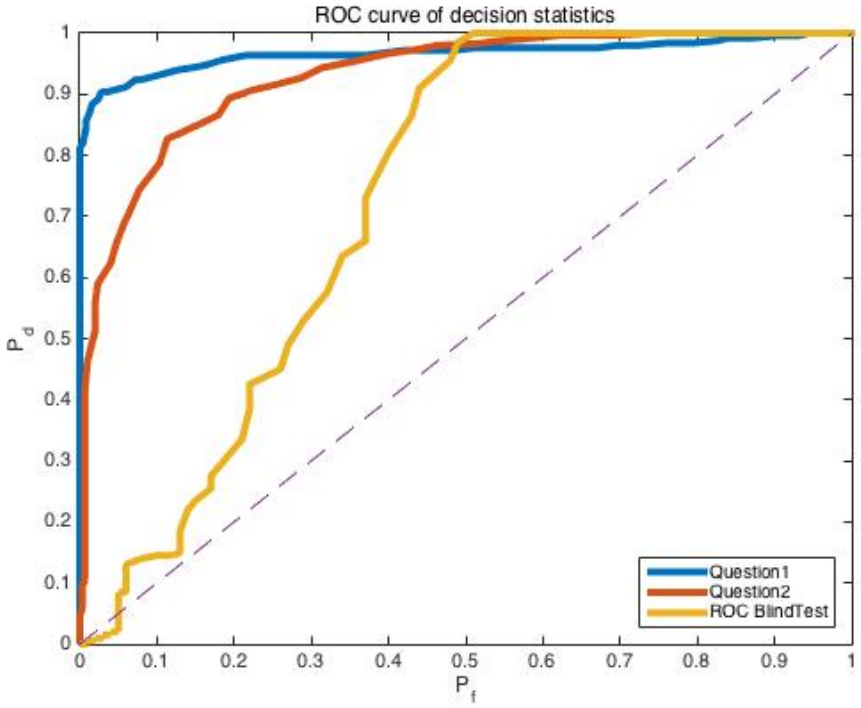
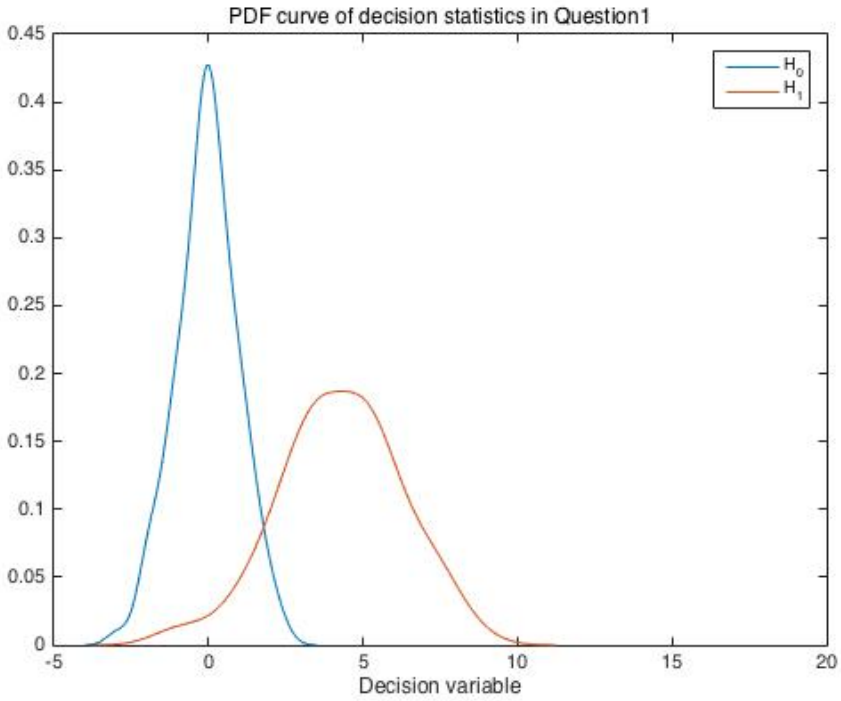
ECE681 HW1 Report

SHENGXIN QIAN

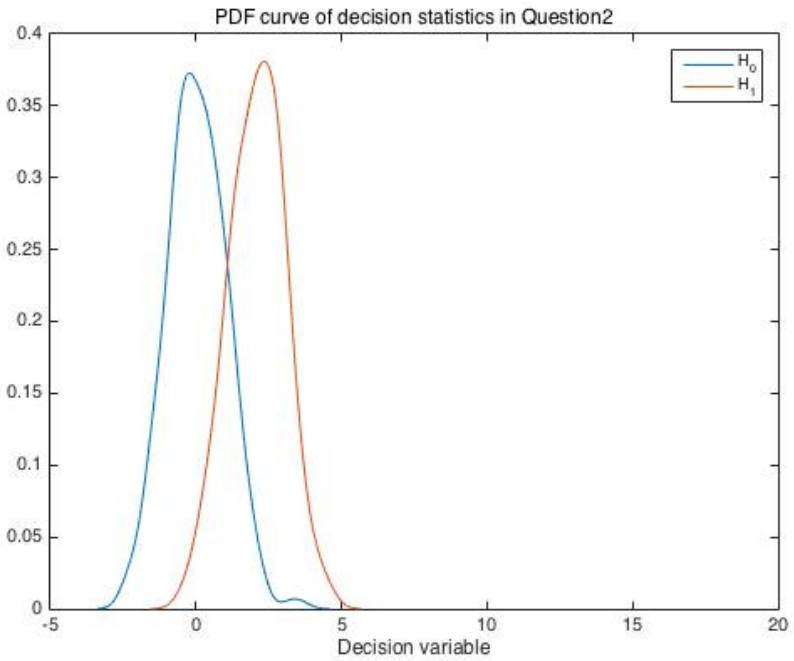
1. Analysis on ROC Curves



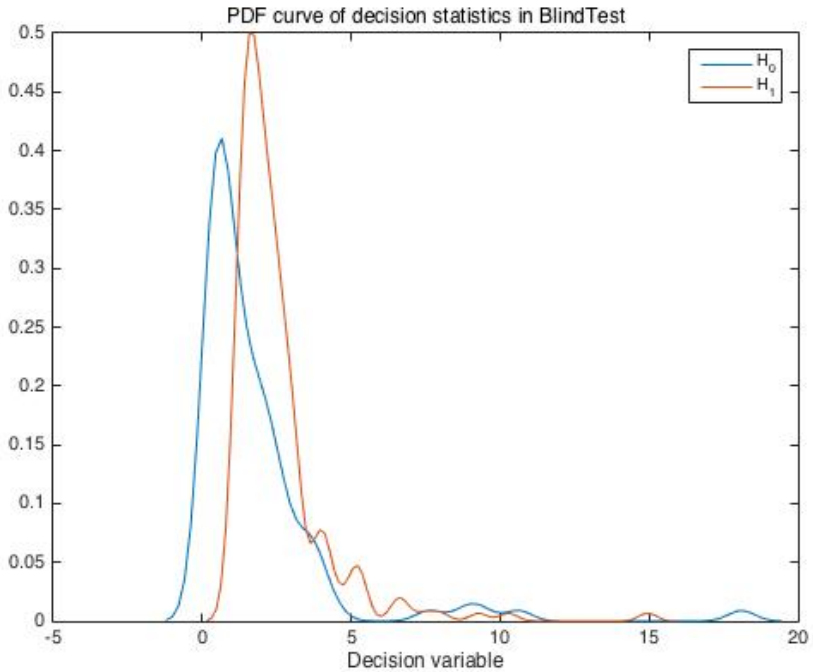


In Q1, the H_0 decision statistics are drawn from N(0,1) distribution. The H_1 decision statistics are drawn from N(4,4) distribution. The pdf curves in the picture above clearly represents normal distribution and mean of H_0 is around 0, the

mean of H_1 is around 4.



In Q2, the H_0 decision statistics are drawn from N(0,1), the H_1 decision statistics are drawn from N(2,1). The pdf curve in the picture above clearly shows what we designed.



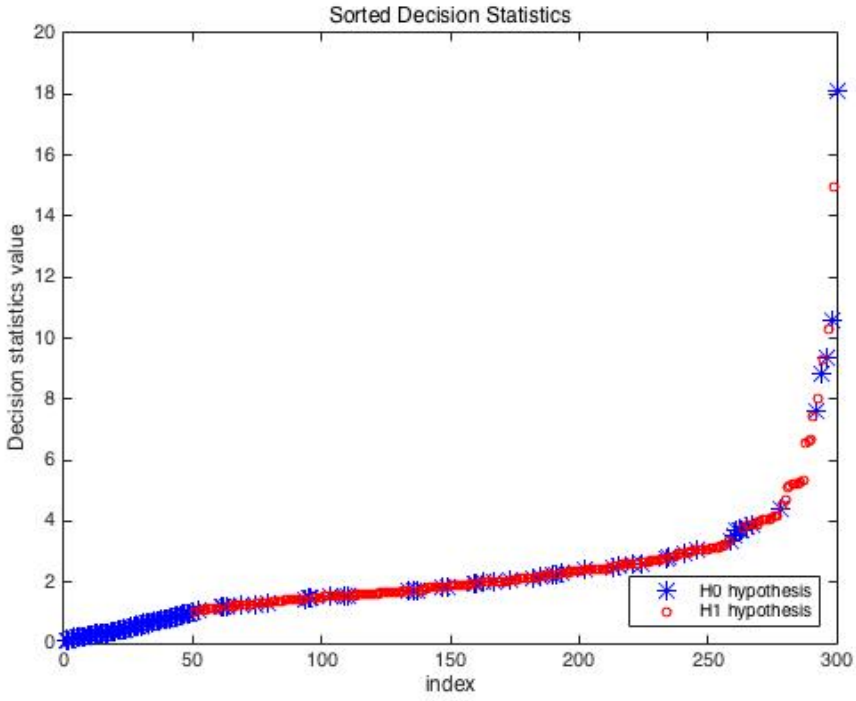
In Blind test, it is obvious that the decision statistics did not come from a standard normal distribution but similar to it. When we compare these PDF curves with their ROC curves, we can clearly recognize the relationship between these two kind of

curves. If the distribution is similar to normal distribution, the discriminability of two different hypotheses could be roughly defined as the distance between peaks of two curves or the overlap area between two PDF curves.

In Q1, the distance between two peaks is around 4.5 which is also the largest peak distance in three tests and it has the best discriminability. It could be supported be the blue curve in ROC picture. It is closer to the left and up side. When we consider about the Q2, the distance between two peaks is around 2.5 and the ROC curve of Q2 has less AUC than the ROC curve of Q1 which could also support each other about the discriminability of decision statistics in Q2.

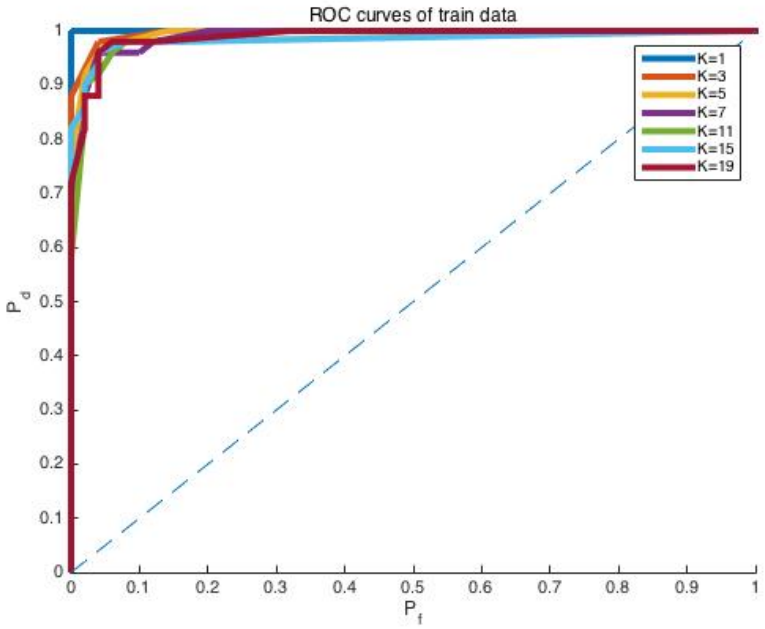
When we consider about the decision statistics in Blind test, the distance of two peaks is less than 2 and their PDF curves have more overlap area than the other which means the ROC curve of this set of decision statistics would be much closer to diagonal. Our estimation could get support from the shape yellow ROC curve.

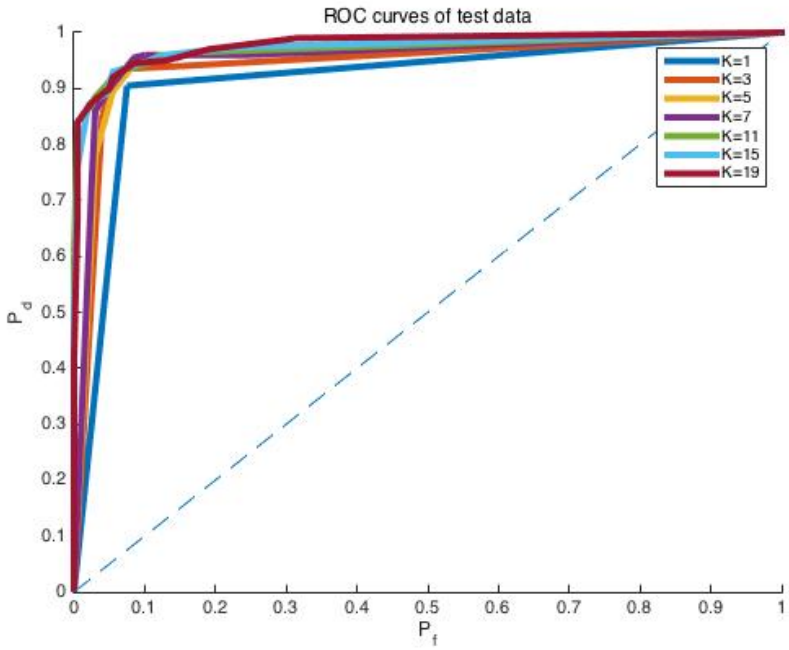
Another important thing is that the distribution of Pd or Pf on PDF curve could affect the distance between the ROC and up side or that between the ROC and left side. For instance, the ROC curve of Blind test has a bizarre part where Pf is around 0.05. Normally, ROC curve will not be below diagonal because most pdf of H1 hypothesis is at the right side of pdf of H0 hypothesis. But, in blind test case, where decision variable is around 18, the pdf of H0 hypothesis has the first small peak. It means the Pf grows faster than Pd at the most left part of ROC curve. That is why the bizarre part exists. More than that, when decision variable ranges from 5 to 20, the Pd and Pf grows at similar rate. That is why ROC curve is mush closer to diagonal than the other part when Pf is less than 0.5. On the contrary, the pdf of Q1 or Q2 corresponds to standard normal distribution and the shape of pdf is symmetric. That is why their' ROC curves will have similar distance to left and up side.



We could also see the discriminability of decision statistics in Blind test from the picture above. The H1 hypothesis part has some overlap part with H0 hypothesis part which may make the ROC curve much closer to diagonal. But there is no overlap between 0 to 2, which will make corresponding part of the ROC curve much closer to the up side. The shape of ROC curve in Blind test would get explanation from this sorted decision statistics picture.

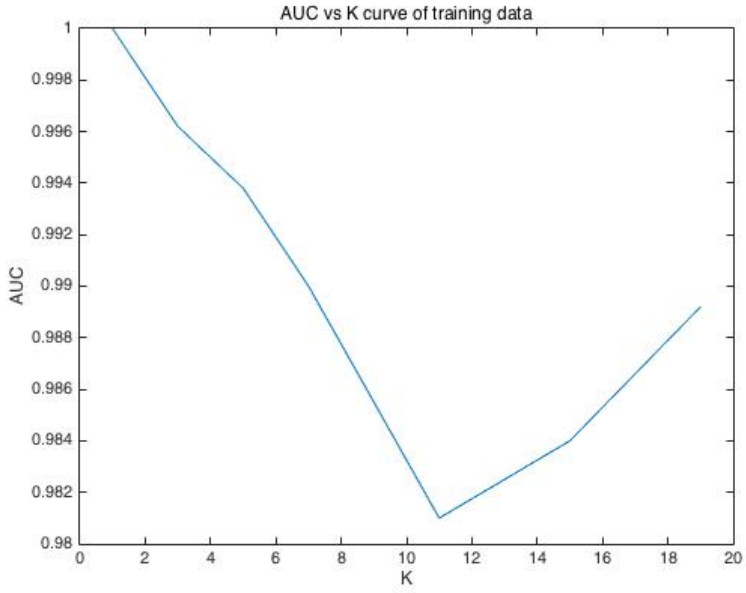
2. Analysis on KNN classification

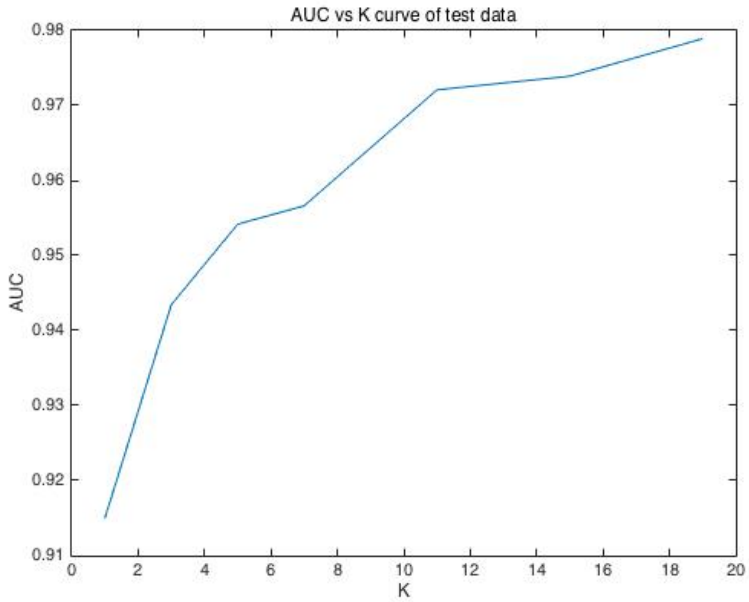




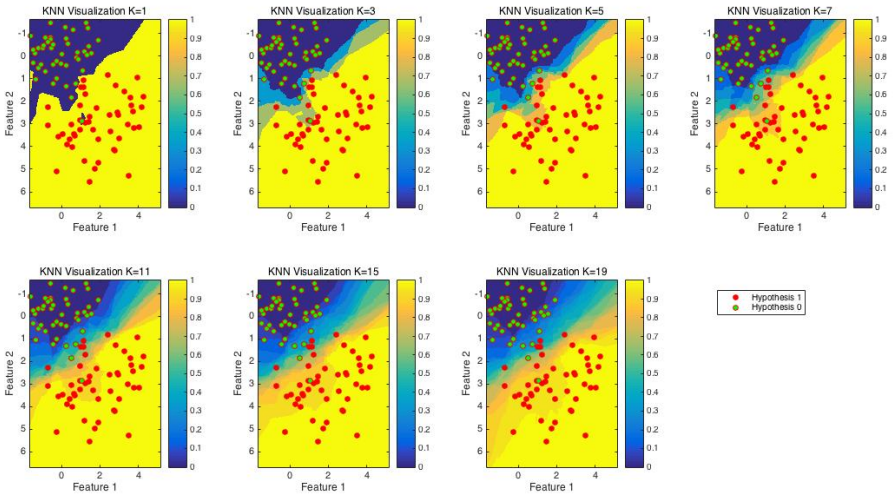
The ROC curve of train data clearly corresponds to our expectation. When K=1 and testing training data, the KNN classifier will get 100% success. When the number

of K gets higher, the accuracy of KNN classifier will drop a little. When the K is too high, many nodes in the other class would be counted which would cause ROC curve has stairs but the performance (AUC) of KNN classifier will be better. The ROC curve of training data explains our expectation very well. When we consider about the ROC curve of test data, it also makes sense. Because when K=1, the KNN classifier is too relative to training data which could bring some noise in the classifier. The classifier should only represent the general distribution but not some random noise. That is why the classifier has the worst performance when K=1 with another group of independent test data. When we increase K, the performance will be better. Although the ROC curve will have some stairs with training data when K is too high. But a higher K will cause higher performance when we test with independent data.





We could get intuitive conclusion about the relationship between the K and the accuracy of KNN Classifier. The AUC vs K curve of training data and test data could perfectly explain our rough theory above. However, we could not use these curves to choose the right K. I could easily conclude that high K is better from the graphs above. But this not right because the higher K is, the more information would be filter. In this case, our training data is a well organized data and corresponds to a simple distribution. If our training data is not so “standard”, there will be a lot of noise. If we filter too much “noise”, we may not have enough useful information left. More than that, we also need to consider the shape of ROC apart from AUC. A smooth ROC curve is what we need, a curve with too many stairs is also not we want even the AUC value is very high. **That is why I finally choose K=15.**



The visualization of KNN Classifier and training data distribution was shown above.

The distribution of Blind test point and test data was shown on the KNN Classifier visualization with K=15.