**Problem 1**

Suppose we set up the following notations:

x as a random variable regarding the three events, and

c(x) as the classifier function, technically c(x) = +1 or -1.

We then have the following probability table:

|  |  |  |
| --- | --- | --- |
| P(x = A) = 0.1 | P(c(x) = +1| x = A) = 0.9 | P(c(x) = -1| x = A) = 0.1 |
| P(x = B) = 0.6 | P(c(x) = +1| x = B) = 0.3 | P(c(x) = -1| x = B) = 0.7 |
| P(x = C) = 0.3 | P(c(x) = +1| x = C) = 0.8 | P(c(x) = -1| x = C) = 0.2 |

So, for the Bayesian optimal classifier, it makes the decision based upon which class is more likely given specific observation x.

Therefore, according to the probability table above, the Bayes optimal decisions are:

If x is event A, it’s classified as +1;

If x is event B, it’s classified as -1;

If x is event C, it’s classified as +1;

Now coming to the expected loss, we calculate the probability that the observation is misclassified:

Using the 0-1 loss, we can get that essentially the expected loss is equivalent to

**Problem 2**

According to Bayes classification rule, we make the classification decision based upon the posterior probability, given observation x and classifier function c(x):

By Bayes rule, we can have

Here in this question, we have two potential outputs for c(x): c1= +1 and c2=-1

And P(x) is a constant that we can ignore.

So we are now technically comparing

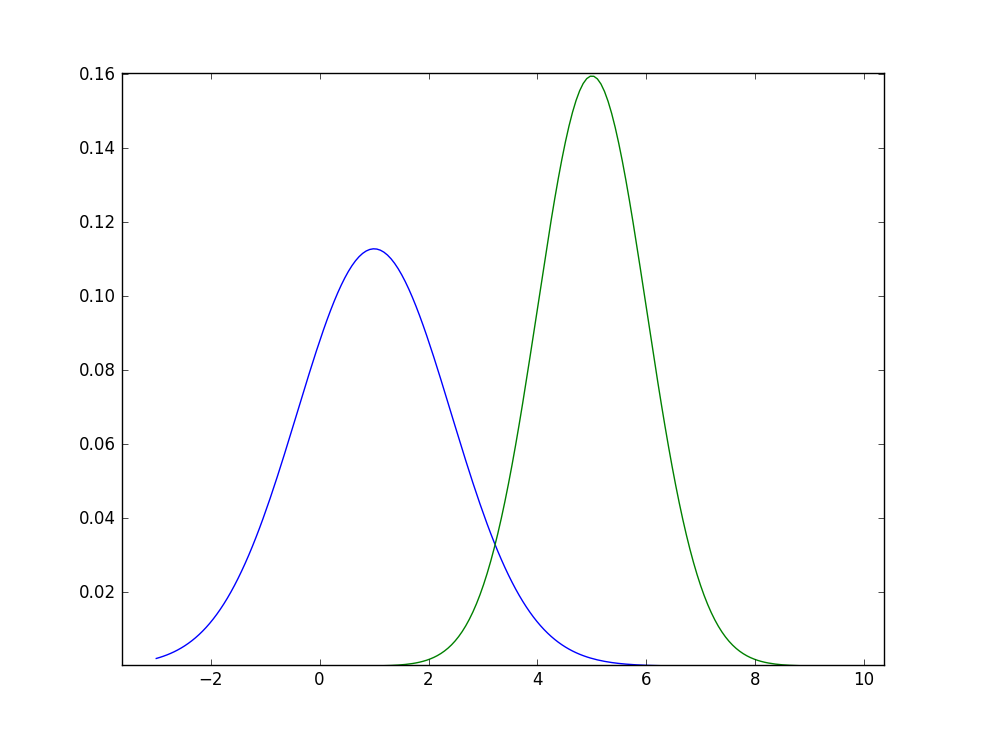
or

and

or

for which the conditions are given as

Thus we can have the plot for and



where the two curves join roughly at .

So when x is less than x0, the decision goes to class c1 +1. Otherwise, the decision goes to c2 -1.

The Bayes risk or the expected loss is represented by the joint area under the two pdf curves, which can be solved by

**Problem 3**

Technically, as k approaches infinity, kNN’s error rate/expected loss will be equivalent to that of Bayes optimal, which is the ideal error rate a classifier can reach. So when k=3, and assuming sufficient samples, we can make use of the Formula 2.125 in the Theodoridis book so that

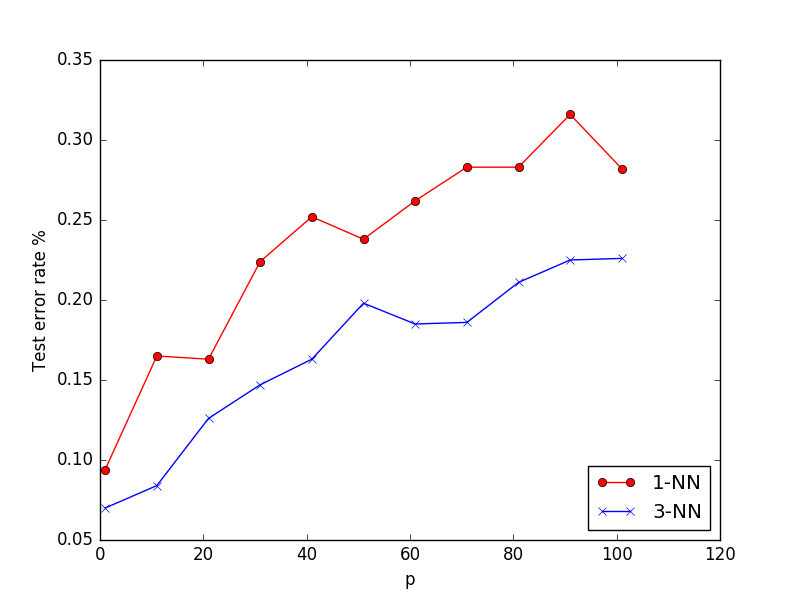
The empirical loss of the 3-NN algorithm depends on the actual training data. But unlike 1-NN, especially near the class decision boundary, the points are probable to be misclassified, whose class is now determined by 3 neighbors instead of only 1. So in the empirical loss function

for some individual points the part will no longer be 0, which means the overall will become closer to the Bayes optimal loss.

Problem 4

For this problem, I personally chose to use Python to substantiate 1-NN and 3-NN classifiers. Several 3rd party libraries are also used, notably NumPy and SKLearn for building the classifier, and Matplotlib for plotting. The source file is attached at the end of this homework report, and also available at

The relationship between the data dimensionality p and the classifier error rate is shown below, as one of the many trials.



We can observe that in general, 3-NN has a much better classification accuracy than 1-NN. Also shown in both classifiers, as the dimensionality of the dataset grows, the classification generally becomes increasingly inaccurate.

**Problem 5**

The VC-dim for disks in R2 is 3. The proof is shown in two major steps.

Step 1. The VC-dim for disks is at least 3, since any 3 points in 2-D will form a triangle which always has a circumscribed circle. Thus 3 points can be easily shattered by relocating this circumscribed circle.

Step 2. Now, for 4 points, there are 3 scenarios.

1. The 4 points are in a line. The labeling series +-+- is impossible.
2. The convex hull of the 4 points is a triangle. Labeling the point inside the triangle differently from the other 3 is impossible since the circle disk is also convex.
3. The convex hull of the 4 points is not a triangle but a quadrilateral. In this case it is impossible to label differently between the two disjoint sets, each of which contains two diagonal points. If such is possible, then there will be two different disks whose symmetric difference has 4 disjoint regions, which is impossible for disks. Like this:

X O

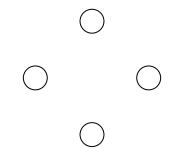
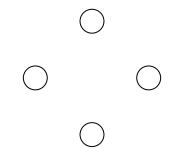
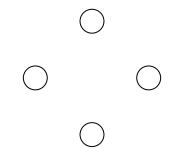
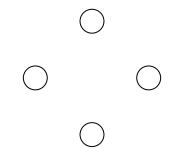
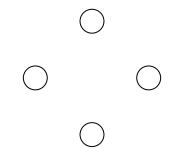
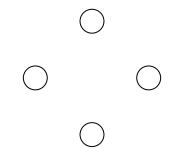
O X

Therefore, the VC-dim is less than 4, as is proved above.

Hence we have proved the VC-dim for disks is exactly 3.

The VC-dim for axis parallel rectangles in R2 is 4.

That 4 points can be shattered by an axis parallel rectangle is quite obvious.



(Above is showing the representative cases)

But now if there are 5 points, we can always create the smallest enclosing quadrilateral by using 4 points assigned as +1. There will always be at least one point classified as -1 which is contained within the enclosing quadrilateral and thus can never be shattered. Therefore, the VC-dim is less than 5.

Hence we have proved the VC-dim for axis parallel rectangles is exactly 4.