PRML Homework Assignment I -Updated

(Due on Saturday, May 16, 2020.)

Please submit to our TA–Siwen Liu before 8:30 am.

For every 5 minutes beyond the deadline, 5 points will be deducted from your score.

Part I

Problem 1: The curse of dimensionality (15 points)

(a) (3 points) Describe the curse of dimensionality. Why does it make learning difficult in high dimensional spaces?

(b) (6 points) For a hypersphere of radius r on a space of dimension d, its volume is given by

$$V_d(r) = \frac{r^d \pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)},$$

where $\Gamma(n)$ is the Gamma function, and $\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx$. Consider a crust of the hypersphere of thickness ε. What is the ratio between the volume of the crust and the volume of the hypersphere? How does the ratio change as d increases?

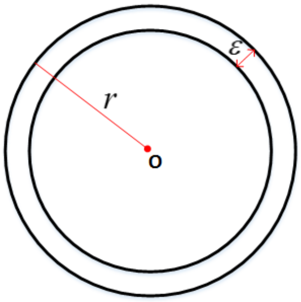

Figure 1: An illustration of the hypersphere and the crust when $d = 2$.

(c) (6 points) We assume that N data points are uniformly distributed in a 100-dimensional unit hypersphere (i.e. $r = 1$) centered at the origin, and the target point x is also located at the origin. Define a hyperspherical neighborhood around the target point with radius r'. How big should r' be to ensure that the hyperspherical neighborhood contains 1% of the data (on average)? How big to contain 10%?

Part II (Optional, Extra Credits)

Principle Component Analysis (PCA) and Fisher Linear Discriminant (FLD)(45 points)

In this problem, we will work on a set of data samples which contains three categories, each category contains 2000 samples, and each sample has a dimension of 2. Please download and uncompress hw1_partII_problem1.zip, and then we will have three text files contains the data of three categories, respectively. Note: Please do not use built-in functions of PCA and FLD in Matlab or Python.

(a) (0 points) Warming up. Plot the first 1000 samples of each category. Your result should be similar to Figure 2.

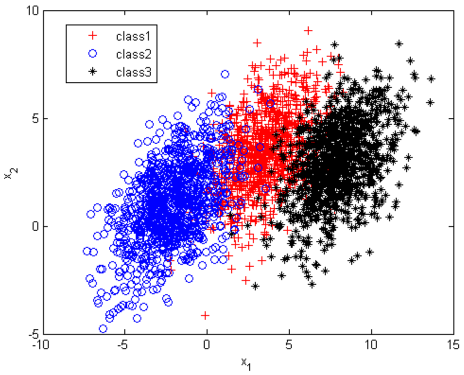

Figure 2: The first 1000 samples of each category.

(b) (10 points) Assume that the first 1000 samples of each category are training samples. We first perform dimension reduction to the training samples (i.e. from two dimensions to one dimension) with PCA method. Please plot the projected points of the training samples along the first PCA axis. Your figure should be similar to Figure 3.

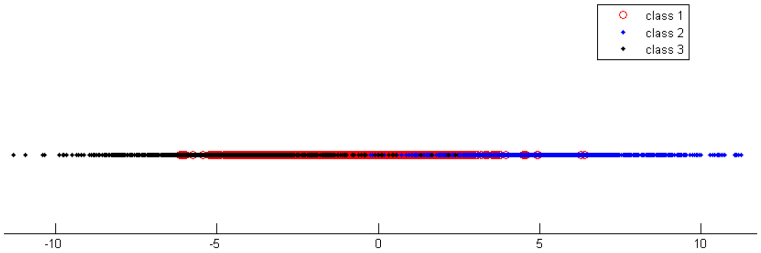

Figure 3: Projection onto the first PCA axis.

(c) (10 points) Assume that the rest of the samples in each category are target samples requesting for classification. Please use PCA method and the nearest-neighbor classifier to classify these samples, and then compute the misclassification rate of each category.

(d) (20 points) Repeat (b) and (c) with FLD method.

(e) (5 points) Describe and interpret your findings by comparing the misclassification rates of (c) and (d).

Submission Guidelines for Part II

1. Create a zip file, YourID_HW1.zip, which includes your source code and a short report that clearly illustrates the required results. The report has to be in the format of pdf or doc. In addition, please verify that all the files can be successfully extracted from the zip file before submission.