Problem Set 1
Due on Friday, October 7, 2016 at 11:55 pm

How to Submit

Create one .zip file (**not** .rar) of your code and written answers and submit it via `ilearn.ucr.edu`. Supply all necessary matlab files. Include `parta.m` and `partc.m` that define functions (or scripts) that take no arguments and generate the results for parts a and c, respectively. Supply one text file that answers part b and part d and name it `ans.txt`. Do not give any directories in your zip file. *Each* file should (in comments if necessary) list

- Your name
- Your UCR student ID number
- The date
- The course (CS 229)
- The assignment number (PS 1)

Curse of Dimensionality

Many machine learning datasets have many features (the inputs are vectors with a large number of dimensions). This problem explores the proximity of points in such high dimensional spaces.

part a. [3 pts]

Consider you have collected m data points, each with d dimensions. For simplicity, we will assume these data points are drawn from a multivariate normal distribution with mean 0 and a covariance matrix that is the identity matrix.

What is the average distance from a point in the dataset to its nearest neighbor in the dataset?

To answer this question emperically, construct a function that will sample such a dataset and return the mean distance (the average is over all points in the dataset) from a point in the dataset to its nearest neighbor in the dataset.

This function's output will vary slightly, depending on the random draw. So, take the average output over 100 different randomly drawn datasets.

Plot this average distance to the nearest point as a function of m from 1 to 1000. On the same figure, show the function for $d = 2, 5, 10$ (as three separate curves). Label your curves and axes correctly.

part b. [2 pts]

Compare this result to Figure 2.6 and Equation 2.24 in the textbook. What does this tell you about learning functions in high dimensional spaces? Why do we care about the average distance to another point?

part c. [2 pts]

Plot the same plot as in part a. However, also add three additional curves: each is the same as one of original curves, but with the following change. Instead of using the identity matrix as the covariance matrix, use a matrix in which the diagonal elements are 1, but the off-diagonal elements are all 0.8;

part d. [2 pts]

Give a short interpretation (a few sentences) explaining the plots from part c.