CptS 580: Advanced Topics in Machine Learning

Homework #2

Due: Tuesday, October 4, 2011

Description

In this homework assignment you will get hands on experience in working with dimensionality reduction techniques discussed in the class. You will also get to work with MATLAB a popular scripting language for machine learning. This software is available on the computers in the computing center in room 353. There are many MATLAB tutorials available online. A short list of links to good tutorials is provided in the supplementary files. There are three parts to this assignment. The first part of the homework does not require any programming. The second part focuses on unsupervised dimensionality reduction techniques, while the third part deals with feature selection.

Part 1

This part of the homework aims to give you an understanding of data distributions in high dimensions. Consider a hyper-orange of radius r (a hyper-orange is a d dimensional orange that tends to be a perfect hyper sphere is high dimensions). Assume that the thickness of the peel is ϵ.

a) Derive an expression of the fraction of the volume of the rest of the hyper-orange to the volume of its peel.
b) Evaluate this for $\epsilon = 0.01r$ and $d = 2, 10, 50, 100, 1000, \infty$.
c) Assume that the data samples from a particular application domain are uniformly distributed inside a hypercube. Based on the analysis done in the earlier problems, where in the hypercube do the samples lie in the high dimensional space?
d) If you have nothing to do and want an interesting problem to solve; Here is one… show that any two random vectors in a high dimensional space are almost orthogonal. Note that there are **no** bonus points for your proof.

Part 2

In this part of the assignment you will explore unsupervised dimensionality reduction (DR) techniques and apply it on a high dimensional dataset. You will be using the dimensionality reduction toolbox for MATLAB developed by Maaten et al. This is a free download and contains implementations of popular DR techniques. You will be conducting the experiments on the MNIST handwritten images dataset. The supplementary files associated with this homework contain two files – mnisttrain and mnisttest. Both of these files contain 100 randomly selected images for the digits 1, 2, 3, 7 and 8 (total of 500 images) to be used for training and testing. Select any two DR techniques from the toolbox (one from each linear and non-linear methods) and project the training data into two or three dimensions. Scatter plot the data points in 3

dimensions where each class has a different color or symbol. What observations can be made in terms of the distribution of the data in the low dimensional space? Using the transformation computed on the training data, project the test data into lower dimensions and classify it using a simple 1-NN. Do you think projecting the data into low dimensions will help to improve the classification performance? Vary the dimensionality of low-dimensional space and observe its effect on the classification performance.

To help you with the coding, the supplementary files for this homework contain a MATLAB script called part2.m, which is a skeletal code for running the DR techniques on the MNIST dataset.

Part 3

This part of the homework deals with exploring feature selection strategies. Select any feature selection strategy from the ORANGE module and apply it on the COLON CANCER dataset. This dataset contains expression levels of 2000 genes taken in 62 different samples. For each sample it is indicated whether it came from a tumor biopsy or not. The real challenge with this dataset is the shape of the data matrix. The number of data samples is less, but the number of attributes (gene expressions) is significantly large. The traditional approach of splitting the dataset into training and test set may not be a good option and you might have to rely on a leave-one-out cross validation approach. Note that the first column in the file corresponds to the label of the instance. Identify and reason out what features are important for the classification problem and comment on the challenges in selecting the features.

Turning in the assignment

The assignment should be mailed as a TXT or PDF file to cook@eecs.wsu.edu by 9.00am on the due date. Include your source code, sample output and summarized results and observations. No late assignments will be accepted.