CS 534: Machine Learning
Homework 5

(Due Dec 1st at 11:59 PM on Canvas)

Submission Instructions: The homework should be submitted electronically on Canvas. The code can be done in any language that `Euler` supports.

1. **(5 + 3 + 2 = 10 pts)** (**Illustrating the "curse of dimensionality"**)

   For a hypersphere of radius a in d dimensions, the volume is related to the surface area of a unit hypersphere (S) as

   $$V = \frac{S \times a^d}{d}.$$

   (a) Use this result to show that the fraction of the volume which lies at values of the radius between $a - \epsilon$ and a, where $0 < \epsilon < a$, is given by $f = 1 - (1 - \epsilon/a)^d$. Hence, show that for any fixed ϵ, no matter how small, this fraction tends to 1 as $d \to \infty$.

   (b) Evaluate the ratio f numerically by plotting the results for different values of $\epsilon/a = 0.01, 0.05, 0.1$ and $d = 1$, 10, 100, and 1000.

   (c) What conclusions can you draw from the plot?

2. **(3 + 5 + 5 + 10 + 2 = 25 pts) PCA & NMF**

   Load the college dataset `Colleges.txt` provided.

   (a) Preprocess the data by removing missing data and properly dealing with categorical data.

   (b) Run PCA on the data. Report how many components were needed to capture 95% of the variance in the data. Discuss what characterizes the first 3 principal components (i.e., which original features are important).

   (c) Discuss why you should normalize the data before performing PCA.

   (d) Run NMF on the data using $R = 3$. What is the squared error of the decomposition? How does this compare to PCA's squared error? Discuss what characterizes the 3 components.

   (e) Compare and contrast your experience with PCA and NMF.

3. **(5 + 15 + 5 = 25 pts) Simulated Neural Network**

   Create a feedforward neural network that learns the $y = x^2$.

   (a) Generate several examples such that $y \geq x^2$ are positive and $y < x^2$ are negative. How many samples do you think constitute a good training dataset?

   (b) Train a neural network using the samples from part (a). Specify the hyperparameters for your model. How did you arrive at these values?

   (c) Evaluate your final neural network on test data. Plot the shape of the decision boundary learned by plotting the predicted positive values.

4. **(5 + 25 + 10 = 40 pts) Thyroid Prediction with Neural Networks**

We will be using the hyperthyroid dataset, and evaluating the model on misclassification rate, F_1 score, and F_2 score. Note that neural networks can be quite expensive – you might want to use a beefier machine to do this.

(a) Preprocess the dataset for NN, what did you do and why?

(b) Build a feedforward neural network on your dataset. How did you select hyperparameters (what was your model assessment strategy)?

(c) Evaluate your final neural network. How does it compare to the results on random forest and support vector machine in terms of the performance metrics, number of parameters, and computation time?