CS 534: Machine Learning
Homework 5

(Due Nov 30th at 11:59 PM on Gradescope)

Submission Instructions:

The homework must be submitted electronically on Gradescope as a single submission. Each question must be tagged appropriately (you need to set which page the problem is on), otherwise you may lose credit. The answer must also be explicit, if you are exporting an iPython notebook, you will need to add remarks to denote your final answer for the question.

The code can be done in Python, R, or MATLAB. The use of any other programming language must be confirmed with Huan (TA). Source code must be submitted separately on Canvas (the deadline is 15 minutes after the Gradescope submission) and zipped into a single file. There is no need to attach the data, and you can assume the data will be in the same directory as the code. The source code should be executable and there should be no explicit path names for the data files (`joyce/CS534/HW5/train.csv` is an example of hard-coding).

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 $\epsilon/a = 0.01$ and $d = 1$, 10, 100, and 1000.

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

2. **(3 + 5 + 5 + 2 + 10 = 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 this processed data. Report how many components were needed to capture 95% of the variance in the normalized data. Discuss what characterizes the first 3 principal components (i.e., which original features are important).

   (c) Normalize the data (where applicable) and run PCA on the normalized data. Report how many components were needed to capture 95% of the variance in the normalized data. Discuss what characterizes the first 3 principal components (i.e., which original features are important).

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

   (e) Run NMF on the normalized data using $R = 3$. Discuss what characterizes the 3 components. How much variance does it capture?

3. **(2+15+3=20 points) Predicting Loan Defaults with k-Nearest Neighbors**

   Consider the Loan Defaults dataset from Homework #3, `loan_default.csv`. You will be using k-NN to predict whether or not a customer will default on a loan.

   (a) Preprocess the dataset for k-NN. What did you do and why?

   (b) Build a k-NN nearest neighbor classifier on the dataset. What was your model assessment and selection strategy (e.g., what were you hyperparameters)? What were the optimal hyperparameters?

   (c) Evaluate the best k-NN model using the hyperparameters from (b). In other words, how well does it do on the test data?

4. **(2+20+3=25 points) Predicting Loan Defaults with Neural Networks** Consider the Loan Defaults dataset from Homework #3, `loan_default.csv`. You will be using neural networks to predict whether or not a customer will default on a loan. Note that neural networks can be quite expensive you might want to use a beefier machine to do this.

   (a) Preprocess the dataset for neural networks. What did you do and why?

   (b) Build a feedforward neural network on your dataset. How did you select hyperparameters and what were the optimal hyperparameters?

   (c) Evaluate the best neural model using the hyperparameters from (b). In other words, how well does it do on the test data?

5. **(7 + 5 + 8 = 20 points) Model Bake-off for Predicting Loan Defaults**

   We will consider a new test subsample `loan_default_2.csv` from the Lending Club Loan Data to compare various models and determine the 'best' model for this dataset. All the models will be evaluated on this new test subsample in terms of AUC, F_1, and F_2.

   (a) Select the top 5 features from the dataset (`loan_default.csv`) and build an unregularized logistic regression model. Specify how you select the top 5 features. How well does it perform on the new test subsample?

   (b) Starting from Homework 3, you've used decision trees, random forests, SVM, k-NN and neural networks. For each of these 5 models, re-train the model using the optimal hyperparameters from your homeworks (you should avoid re-tuning to save yourself time) on all of the 2850 training data `loan_default.csv`. How well does it perform on the new test subsample?

   (c) Using the model from (a) as the baseline, how does all the more 'complex' models compare to the baseline? Compare and contrast the various models with respect to interpretability, computation time, and predictive performance. If you were the loan officer using this model, which one would you use to guide whether or not to give a loan?