CS 534: Machine Learning

Homework #4

Due: April 4th at 11:59 PM on Canvas

Submission Instructions: The homework should be submitted electronically on Canvas. The code can be done in Python, R, or Matlab with the hints mostly focusing on those who will be programming in Python. Also be sure to attach the code to ensure you'll get partial / full credit if the answer is not what was expected.

1. **(5 points) Variance of Correlated Samples**

   Assume that we have B identically distributed random variables where $x_i \sim N(m, \sigma^2)$ for all i and x_i and x_j are correlated with a correlated coefficient of ρ. Show that the variance of the average is:

   $$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

2. **(7 + 7 = 14 points) Kernel Methods**

   (a) Assuming that $\mathbf{x} = [x_1, x_2], \mathbf{z} = [z_1, z_2]$ (i.e., both vectors are two-dimensional) and $\beta > 0$, show that the following is a kernel:

   $$k_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta\mathbf{x} \cdot \mathbf{z})^2 - 1$$

   Do so by demonstrating a feature mapping $\phi(\mathbf{x})$ such that $k_\beta(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$.

   (b) One way to construct kernels is to build them from simpler ones. Assuming $k_1(\mathbf{x}, \mathbf{z})$ and $k_2(\mathbf{x}, \mathbf{z})$ are kernels, then one can show that so are these:

   i. (scaling) $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})k_1(\mathbf{x}, \mathbf{z})$ for any function $f(\mathbf{x}) \in \mathbb{R}$

   ii. (sum) $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$

   iii. (product) $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z})\ k_2(\mathbf{x}, \mathbf{z})$

   Using the above rules and the fact that $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top\mathbf{z}$ show that the following is also a kernel:

   $$\left(1 + (\frac{\mathbf{x}}{||\mathbf{x}||_2})^\top(\frac{\mathbf{z}}{||\mathbf{z}||_2})\right)^3$$

3. **(4+5+5+5+6 = 25 points) Credit Card Default with Support Vector Machines**

   We will be using the credit card default dataset provided on the UCI Machine Learning repository (`http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`). The problem is to determine whether or not there will be a default payment. Partition the data into 50%–50% stratified train–test split (i.e., the ratio of positive to negative class should remain the same in both datasets) that you will use for the rest of the homework. We want to evaluate our models on misclassification rate, F_1 score, and F_2 score. You may find it helpful to preprocess the data for computation purposes. If you do, specify how you plan to preprocess the data and why you chose this.

(a) Name a real–world scenario where one would prefer optimizing the model using the F_2 score over the F_1 score for predicting credit card default?

(b) Build a linear SVM. How did you choose the optimal parameter(s) for your model?

(c) Build a SVM with polynomial kernel. How did you choose the optimal parameter(s) for your model?

(d) Build an RBF-kernel SVM. How did you choose the optimal parameter(s) for your model?

(e) Report the evaluation metrics for the training and test set for all of the above. How do they compare? What can you say about the models?

4. **(5+5+5+10= 25 pts) Credit Card Default with Neural Networks**

(a) Build a MLP with 1 hidden layer with 20 neurons and the sigmoid activation function. How did you choose the learning rate?

(b) Build a MLP with 1 hidden layer with 20 neurons and the rectified linear unit activation function. How did you choose the learning rate?

(c) Build a MLP with 1 hidden layer with 20 neurons and the hyperbolic tangent activation function. How did you choose the learning rate?

(d) Scale (where applicable) the data to have zero mean and standard deviation of 1. Re-run the previous 3 parts. How does the pre-processing step compare in terms of computation time and the evaluation metrics?

5. **(15 pts) Credit Card Default with k-Nearest Neighbors**

Use k-NN with k = 1, 3, 5, 15, 25. You may find it helpful to preprocess the data for computation purposes. If you do, specify how you plan to preprocess the data if you are doing it differently from Problem 3. How does it perform on the evaluation metric?

6. **(16 points) Credit Card Default Classification using Ensemble Model**

Build an ensemble model based on the classifiers you've learned in class thus far. To achieve full credit, your ensemble should have an accuracy above 0.82 on your test set.