DSC 190: Homework 2

January 12, 2021

Instructions

Please submit your solution **by the beginning of the week 4 lecture (9:30 AM PT, Jan 26)**. Submissions should be made on **gradescope**. Please complete homework **individually**. Please include the code of your solutions in the submission with a write-up describing how to run the code.

You are allowed to use any third-party libraries.

You will need the following files for this Homework:
Car-evaluation.csv (available on canvas)
wine.csv (available on canvas)

1. **Kernel Space Transformation (20 points):**

Kernel functions are the functions that map one space to a higher dimensional space. This question demonstrates the primary use of kernel functions in SVM. Consider four points: A $= (1, 1)$, B $= (2, 2)$, C $= (2, 1)$, D $= (1, 2)$ and two labels *red*, *blue*. Suppose A, B belong to the label *red* and C, D belong to the label *blue*.

(a) Are labels *red* and *blue* linearly separable? Explain your reasoning.

(b) Consider the kernel function: $\phi(x) = (\frac{x_1}{x_2}, |x_1 - x_2|, x_1 + x_2)$ where $x = (x_1, x_2)$. We can observe that this kernel function maps 2-dimensional space to 3-dimensional space. What are the points A, B, C, D after transformation?

(c) Are labels *red* and *blue* linearly separable after the transformation? Explain your reasoning.

2. **Naive Bayes (20 points):**

(a) Suppose we have a large training set with $100,000$ samples. There are 3 different classes and 4 different features in the training set. Each distinct features can take 5 different values. How many parameters do we need to estimate, if we don't make the naive Bayes assumption? How many parameters do we need to estimate, if we make the naive Bayes assumption?

(b) Shuffle the data with random seed 42, and split it into training, validation, and test splits, with a 60/20/20% ratio(e.g., use `random_state` in `sklearn.model_selection.train_test_split`). Train a Naive-Bayes classifier (e.g., `sklearn.naive_bayes.GaussianNB`) on the car-evaluation dataset using all the features and report the training/validation/test accuracy.

3. **SVM (20 points):** In this question, we will be using the Car Evaluation dataset. (Car-evaluation.csv' in Canvas)

   (a) Shuffle the data with random seed 42, and split it into training, validation, and test splits, with a 60/20/20% ratio(e.g., use `random_state` in `sklearn.model_selection.train_test_split`). Train a SVM classifier (e.g., `sklearn.svm.SVC`) with regularization parameter $\mathbf{C} = 1.0$ on the car-evaluation dataset using all the features and report the training/validation/test accuracy.

   (b) Consider values of $\mathbf{C}$ in the range $\{10^{-4}, 10^{-3}, ..., 10^{3}, 10^{4}\}$. Report (or plot) the train, validation, and test accuracy for each value of C. Based on these values, which classifier would you select (in terms of generalization performance) and why?

4. **Decision Tree and Random Forest (20 points):** In this question, we will be predicting wine quality for the Wine dataset. ('Wine.csv' in Canvas)

   (a) Shuffle the data with random seed 42, and split it into training and test sets, with 70/30 ratio. Train a Decision Tree classifier (e.g., `sklearn.tree.DecisionTreeClassifier`) on the training set only and report the confusion matrix on the test set using the trained model. Report the precision, recall and f1 score using the confusion matrix.

   (b) Using the same data and split, train a Random Forest classifier (e.g., `sklearn.ensemble.RandomForestClassifier`) on the training set only and report the confusion matrix on the test set using the trained model. Report the precision, recall and f1 score using the confusion matrix. Explain why Random Forest algorithm performs better than Decision Tree algorithm.

   (c) Suppose we have a dataset with 1000 negative samples and 10 positive samples, and a model was trained based on this dataset. Explain what could happen and why precision and recall would be a better evaluation metrics than accuracy in this case.

5. **Bias vs Variance (20 points):**

   (a) (5 points) State briefly what you understand by the bias-variance tradeoff.

   (b) (5 points) What happens to the bias and variance when the number of training samples increases?

   (c) (10 points) Suppose you decide to use the k-Nearest Neighbors (KNN) classifier for a classification problem. In this problem, assume that you are given a fixed number of training samples. You can choose either 1 or 100 as your value of k, the number of neighbors to be considered for classification. Which k would give you higher variance? Which k would give you higher bias? Explain.

Suggested reading for this question: http://scott.fortmann-roe.com/docs/BiasVariance.html