Introduction to Machine Learning
Lab Assignment 2: Classical Classification Techniques

Due February 13, 2018 at 11:59pm

Summary: In class, we have discussed classification models that you can use with predefined features. In this assignment, you will demonstrate you understand how to apply, evaluate, and analyze these models. A successful submission of your project will consist of two contributions. First, it should include the source code of your implementation (i.e., portions indicated by "Code"): a hyperlink to a website where your code is publicly available is sufficient. Second, it should include a PDF document with all requested experimental results and analysis (i.e., portions indicated by "Write-up").

1. **Construct Datasets for Training and Evaluation [5 points]**

   (a) Load a real dataset of your choice that is designed for the classification problem; e.g., iris data, wine data, breast cancer data. (Code)

   (b) Create a 80/20 train/test split of the dataset. (Code)

2. **Optimize Hyperparameter(s) for Each Classification Model [60 points]**: Perform stratified 10-fold cross-validation on the training dataset and use the "accuracy" measure to assess performance when evaluating each model.

   (a) Decision tree: find the optimal hyperparameters for the split criterion (i.e., test "gini" and "entropy") and tree depth (i.e., test at least 5 different values) when training a decision tree. Report the optimal hyperparameters found and how many hyperparameter combinations you tested in total. (Code and Write-up)

   (b) K-Nearest Neighbors (K-NN): find the optimal hyperparameters for the distance metric (i.e., test "Euclidean" and "Manhattan") and number of nearest neighbors (i.e., test at least 5 different values) when using k-Nearest Neighbors. Report the optimal hyperparameters found and how many hyperparameter combinations you tested in total. (Code and Write-up)

   (c) Support Vector Machine (SVM): find the optimal hyperparameters for the polynomial degree, kernel bandwidth (i.e., `gamma`), and regularization parameter (i.e., `C`) when training a kernel SVM with a polynomial kernel. You must evaluate all possible combinations of at least 3 degree values (for the polynomial degree), at least 5 `gamma` values, and at least 5 `C` values. Report the optimal hyperparameters found and how many hyperparameter combinations you tested in total. (Code and Write-up)

3. **Comparative Analysis of Optimized Classification Models [35 points]**

   (a) Retrain each of the three models (i.e., Decision tree, K-NN, and SVM) on all the training data using the optimal hyperparameters found in part 2. (Code)

   (b) Train a Gaussian Naive Bayes model on the the training data. (Code)

   (c) Report the predictive performance on the test dataset for each of the four models from parts (a) and (b) with respect to each of the following evaluation metrics: accuracy, precision, and recall.

   (d) Also visualize the predictive performance of the each of the four models from parts (a) and (b) by showing the resulting confusion matrix for each model. (Code and Write-up)

(e) Write a discussion analyzing and comparing the performance of the four models from parts (a) and (b). For example, which method(s) perform the best/worst? What do the different performance metrics tell you about the results? Why do you suspect the algorithms performed as they did? Your discussion should consist of two to four paragraphs. (Write-up).

How to Submit Lab Assignment 2: Please submit a pdf that provides hyperlinks to your code or answers to the questions, as deemed appropriate for the task. The pdf file should be named using your first and last name; i.e., firstname_lastname.pdf. The material you submit must be your own.