Semi-supervised Learning Programming

Problem Description This project involves exploration of different semi-supervised learning algorithms. Typically, in semi-supervised learning problems we assume that we have access to a large amount of unlabeled samples and a limited number of labeled instances. In this project, we ask you to apply three different semi-supervised algorithms to a dataset of credit card applications for classification task and compare the results with fully supervised models. Also, we ask you to repeat each experiment for varying proportion of labeled data and report the results.

Models In this project, you will explore a semi-supervised algorithm: Self-Training below. Implement the algorithms and compare their performance on the test set:

Self-Training (Self-learning) i s a heuristic-based algorithm. Apply this method using four different base classifiers: Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR). We also explore Label Propagation (LP). You can implement all such models using scikit-learn packages.

Then, compare the performance of these semi-supervised models against the supervised base classifiers, trained on the same set of labeled data.

Dataset & Evaluation: We have chosen a credit card application dataset from UCI repos-itory: https://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval). This dataset includes financial information from 690 samples with 14 variables. The at-tributes are a combination of numerical and categorical values and includes a few missing values t hat you need to handle. You can divide data into t raining and testing s et by 70/30 ratio. We expect you to experiment all models with varying proportion of labeled data start-ing from 1% and repeat for 2%, 10%, and 15%. For each experiment, you need to use stratified random sampling so that the ratio of positive and negative instances are the same across labeled (L) and unlabeled (U) data.

To evaluate the performance of each semi-supervised and supervised model use the general classification metrics such as accuracy, precision, recall, F1-measure, and AUC.