IEE 520/BMI 555 Fall 2020 Project

Due by 11:59 PM (MST) on December 1, 2020.

INDIVIDUAL PROJECT—YOUR WORK IS TO BE YOUR OWN.

Objective

Use the data provided with this project. One file contains the **labeled data** that you should use for model building. A second file contains **unlabeled data** that will be used to evaluate your model.

Build a classification model for this data based on the methods described in the course. You may choose any classifier and use any options covered in the course. It is often better to work to tune a particular classifier than to superficially jump to many classifiers.

Data Description

The labeled and unlabeled data sets were selected randomly from the original data set.
Columns x1 through x12 are coded categorical attributes and x13 through x22 are numerical.
The *instance* column is available to order data and should not be used in models. Missing values are encoded as *None*.

Evaluation

You will be primarily evaluated by the performance of your model, but a written report is also required. Your model will be scored as the ***balanced error rate***. This is the average of the error rate on each class. This is *different* from the overall (weighted) error rate. This balanced error rate encourages models to predict each class equally well. You should work to minimize the error rate on each class. For example, consider the following confusion matrix:

	Prediction			
Actual	Class 1	Class 0	Error rate	
Class 1	850	50	50/900 = 0.055	
Class 0	50	50	50/100 = 0.5	

Overall error rate = 100/1000 = 0.1
Balanced error rate = (0.5 + 0.055)/2 = 0.28

A simple adjustment is to sample from the training data to create a new data set with equal rows from each class that is used to train your model. As mentioned in class, you might *upsample* (select the same instance more than once) to create equal instances for the majority class or *downsample* to create equal instances for the minority class. In many cases, this extra step **might not** even be needed.

Predictions and Report

Upload two files.
1) Submit a written report with a brief description of 1) Prepare: any preprocessing of the data, 2) Methods: methods and parameters you tried, 3) Evaluate: how you evaluated your methods, 4) Selected Model: your final model, and why you choose your final model and parameter settings. The description of your final model and parameters needs to provide sufficient detail to be reproduced. Typically five pages are sufficient. Name your file IEE520BMI555Report2020*yourfullname*.

2) Submit predictions for the **unlabeled** data in a comma separated values (CSV) file with two columns: the first column is the *instance* number, the second column is the predicted class. No headings. Name your file BMI555IEE520Results2020*yourfullname*.csv.