CS 6375
ASSIGNMENT __Project Status Report__

Names of students in your group:
Xinyang Zhu
Ye Yao

Number of free late days used: _____0_____

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Problem

Your goal is to predict the binary class heart_disease_present, which represents whether or not a patient has heart disease:

- 0 represents no heart disease present
- 1 represents heart disease present

Dataset

There are 180 instances in the train dataset and 90 instances in the test dataset.
There are 13 features in the dataset. They are described below.

- slope_of_peak_exercise_st_segment (type: int): the slope of the peak exercise ST segment, an electrocardiography read out indicating quality of blood flow to the heart
- thal (type: categorical): results of thallium stress test measuring blood flow to the heart, with possible values normal, fixed_defect, reversible_defect
- resting_blood_pressure (type: int): resting blood pressure
- chest_pain_type (type: int): chest pain type (4 values)
- num_major_vessels (type: int): number of major vessels (0-3) colored by flourosopy
- fasting_blood_sugar_gt_120_mg_per_dl (type: binary): fasting blood sugar > 120 mg/dl
- resting_ekg_results (type: int): resting electrocardiographic results (values 0,1,2)
- serum_cholesterol_mg_per_dl (type: int): serum cholestoral in mg/dl
- oldpeak_eq_st_depression (type: float): oldpeak = ST depression induced by exercise relative to rest, a measure of abnormality in electrocardiograms
- sex (type: binary): 0: female, 1: male
- age (type: int): age in years
- max_heart_rate_achieved (type: int): maximum heart rate achieved (beats per minute)
- exercise_induced_angina (type: binary): exercise-induced chest pain (0: False, 1: True)

Metric

The metric in this competition is logarithmic loss, or log loss, which uses the probabilities of class predictions and the true class labels to generate a number that is closer to zero for better models, and exactly zero for a perfect model.

Techniques

Regression

Bayesian regression, logistic regression, support vector machines, stochastic gradient descent, nearest neighbors regression, gaussian processes regression, decision trees, neural network, ensemble methods (bagging, random forest, AdaBoost, gradient boosting)

Experimental Methodology

Pre-processing

One hot encoder, normalizer, column transformer, pipeline, covariance matrix

Create training, validation and test datasets

Random splitter, grid search cross validation

Coding language/technique

The coding language we use is Python 3. The packages we use include numpy, pandas and sklearn.

Preliminary Results