1st update on the cardiovascular diagnostic project

(Course Project – FA20 ECE532)

https://anabaa.github.io/ECE532_FALL20_PROJECT_NabaaAli/

- Progress done by 11/17/2020

- Data samples are loaded from the .csv file and checked for duplicate samples (24 duplicates were found).
- The zero entries in the labels vector are replaced with '-1', such that a sample with no cardiovascular risk is now labeled with '-1' instead of '0'.
- The dimension of the features matrix X is 66976 x 11 and the y vector is 66976 x 1.
- Rank(X) = 11, Hence X is full rank and thus invertible.
- The SVD decomposition of X showed that most of the information is gained using the first five singular values as seen in the project results page.
- The mean is subtracted from all the samples for the three linear regression classifiers.

  The goal of this project is to train classifiers using ML techniques to predict cardiovascular disease. Three classifiers were trained using three different algorithms as follows:

  - The least squares classifier performed on with an average error rate of 35.54 % and the residual $||Xw - y||_2^2 = 10232$ (Cross-validation with 7 sets, 69976 samples in total)
  - Ridge regression was used to train a classifier with a minimal sensitivity to noise and small disturbances. Cross validation was used to choose the optimum penalty (λ value). The average error rate is equal to 35.517 % and the residual $||Xw - y||_2^2 = 10357.8$
  - To obtain a sparse solution, LASSO classifier was trained using iterative thresholding and cross validation to choose the optimum penalty (λ value). The error rate for this classifier is 40.3% and the squared residual is 10773.389. Although the average error % and the squared residual both increased compared to ridge regression and LS, Interestingly the sparsity of the weights vector w, ranged from 1 to 4. In other words, 4 features are sufficient for classification purposes with minimal accuracy setbacks.
  - First-hand results are available at https://github.com/Anabaa/ECE532_FALL20_PROJECT_NabaaAli/blob/gh-pages/results.md
  - Future work involves utilizing K-nearest neighbours, SVM and a neural network with 2 layers.