ELL 409/784 2020 Assignment 1

Prathosh A. P.

November 18, 2020

1 Instructions and Questions

1. This assignment has 3 problems. One each on Binary and multi-class classification and one on regression.
2. You are expected to try all the classification algorithms taught in the course so far: Bayes' classifier (with different class conditional densities such as Gaussian, GMM etc.), Naive Bayes, Max. Likelihood, MAP and Parzen Window density estimates, K-Nearest Neighbours, Linear models, Generalized linear models with different kernels, logistic classifier/regressor.
3. You should try different regularizers such as L2, L1, Elastic net with all the above models.
4. You are also expected to try different loss functions (MSE, MAE, Cross-Entropy).
5. Compute the Accuracy, Precision, Recall and F1 score, and plot the RoC curves with different hyper-paramters.
6. Observe and plot the Bias-Variance trade-off curves with different amount of training data and hyper-parameter changes.
7. All your codes should be in Python. You may use Numpy/Pandas etc. but not specified ML libraries such as scikit-learn.
8. All the algorithms have to be coded up from scratch (including the stochastic Gradient descent, EM for GMM etc.)
9. The number of experiments that you do is not limited. The more experiments/models you try, the more marks you get. (You are welcome to use data vizulization techniques that are not taught).
10. You have to finally submit a 4-page write up in IEEE format (use Overleaf and Latex), with all observations, graphs, results etc.
11. You also should submit your codes - It is recommended that you use Jupiter notebooks and share them directly (easier to evaluate).

2 Problem 1 (Binary Classification)

Data Link: `https://www.dropbox.com/s/t7ycfw00mc755cg/health_data.csv?dl=0`

Data consists of four columns:

1. age: Age of the patient (in years)
2. restbps: Resting blood pressure (in mm Hg on admission to the hospital)
3. chol: serum cholestoral in mg/dl
4. category: 0 indicates the patient is healthy, 1 indicates the patient suffers from heart disease

3 Problem 2 (Regression)

Data Link: `https://www.dropbox.com/s/8tqk3cavdbbe3nb/weather_data.xlsx?dl=0`

Data consists of seven columns:

1. dewptc: Dewpoint in C
2. hum: Humidity
3. wndspd: Wind Speed
4. pressure: Air pressure
5. rain: Binary variable indicating if it rained
6. smoke: Binary variable indicating if there was smoke
7. temp: Temperature in C

4 Problem 3: Multi-class Classification

Data Link: `https://www.dropbox.com/s/z9ebwa49koaqs7i/Medical_MNIST.zip?dl=0`

Dataset contains 6 categories as listed below:

1. AbdomenCT
2. BreastMRI
3. ChestCT

4. CXR
5. Hand
6. HeadCT

Design a 6-class classifier. Report 5-fold classification train and test accuracy with 80-20 split. Plot the 5-fold confusion matrix for both training and test split. Compute per-class precision, recall and F1 score. Finally, compute the macro-F1 score. Discuss the findings.