CS 4044 Pattern Recognition

2018-2019 Winter Semester

Assignment 1

Please use the “Bank Marketing” UCI dataset https://archive.ics.uci.edu/ml/datasets/Bank+Marketing and try to predict the marketing outcome (Y/N).

1. Create a parametric model using Logistic Regression (with linear as well as polynomial features)
2. Compare its accuracy/precision/recall etc. with a non-parametric model built using SVM
3. Perform PCA and find the number of dimensions required to capture 80% of variance. Assess the performance improvement.

Please submit your findings as a one page document. Please upload the final version of the source code as well.

Hints

- Refer the Data set Information and Attribute Information from UCI site to get a good understanding of the data set and its attributes
- Spin out a TEST set and keep aside

  import sklearn.model_selection
- Understand attribute types, ranges, outliers, nulls, correlation with other attributes etc.
- Numeric values are not normalized & standardized - all are having different ranges, different standard deviations; some in the range of 0-100, few others in the range 0-1 and some others in the range 0-1000. If there are very high outlier (default values for certain attributes) values, standardize them.

  import sklearn.preprocessing (MinMaxScaler, StandardScaler)

  import sklearn.pipeline
- Try linear models

  import sklearn.linear_model (LinearRegression, Ridge, Lasso, ElasticNet, LogisticRegression)

  import sklearn.preprocessing (PolynomialFeatures)
- Access models

  import sklearn.metrics
- Try SVM

  import sklearn.svm
- If you are trying ANN, use encoders for categorical attributes
- For PCA

  import sklearn.decomposition