East West University
Department of Computer Science and Engineering

CSE 303: Statistics for Data Science
LAB 06
Course Instructor: Dr. Mohammad Rezwanul Huq

Datasets Description for plotting using Matplotlib

1. Wine Quality Dataset

The Wine Quality Dataset involves predicting the quality of white wines on a scale given chemical measures of each wine.

It is a multi-class classification problem, but could also be framed as a regression problem. The number of observations for each class is not balanced. There are 4,898 observations with 11 input variables and one output variable. The variable names are as follows:

1. Fixed acidity.
2. Volatile acidity.
3. Citric acid.
4. Residual sugar.
5. Chlorides.
6. Free sulfur dioxide.
7. Total sulfur dioxide.
8. Density.
9. pH.
10. Sulphates.
11. Alcohol.
12. Quality (score between 0 and 10).

The baseline performance of predicting the mean value is an RMSE of approximately 0.148 quality points.

Download link:
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

2. Pima Indians Diabetes Dataset

The Pima Indians Diabetes Dataset involves predicting the onset of diabetes within 5 years in Pima Indians given medical details.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 768 observations with 8 input variables and 1 output variable. Missing values are believed to be encoded with zero values. The variable names are as follows:

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 65%. Top results achieve a classification accuracy of approximately 77%.

Download Link: https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv

3. Banknote Dataset

The Banknote Dataset involves predicting whether a given banknote is authentic given a number of measures taken from a photograph.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 1,372 observations with 4 input variables and 1 output variable. The variable names are as follows:

1. Variance of Wavelet Transformed image (continuous).
2. Skewness of Wavelet Transformed image (continuous).
3. Kurtosis of Wavelet Transformed image (continuous).
4. Entropy of image (continuous).
5. Class (0 for authentic, 1 for inauthentic).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 50%.

Download Link:
https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt

4. Abalone Dataset

The Abalone Dataset involves predicting the age of abalone given objective measures of individuals.

It is a multi-class classification problem, but can also be framed as a regression. The number of observations for each class is not balanced. There are 4,177 observations with 8 input variables and 1 output variable. The variable names are as follows:

1. Sex (M, F, I).
2. Length.
3. Diameter.
4. Height.
5. Whole weight.
6. Shucked weight.
7. Viscera weight.
8. Shell weight.
9. Rings.

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 16%. The baseline performance of predicting the mean value is an RMSE of approximately 3.2 rings.

Download Link:
https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data

5. Boston House Dataset

The Boston House Price Dataset involves the prediction of a house price in thousands of dollars given details of the house and its neighborhood.

It is a regression problem. There are 506 observations with 13 input variables and 1 output variable. The variable names are as follows:

1. CRIM: per capita crime rate by town.
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of nonretail business acres per town.
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5. NOX: nitric oxides concentration (parts per 10 million).
6. RM: average number of rooms per dwelling.
7. AGE: proportion of owner-occupied units built prior to 1940.
8. DIS: weighted distances to five Boston employment centers.
9. RAD: index of accessibility to radial highways.
10. TAX: full-value property-tax rate per $10,000.
11. PTRATIO: pupil-teacher ratio by town.
12. B: 1000(Bk – 0.63)^2 where Bk is the proportion of blacks by town.
13. LSTAT: % lower status of the population.
14. MEDV: Median value of owner-occupied homes in $1000s.

The baseline performance of predicting the mean value is an RMSE of approximately 9.21 thousand dollars.

Download Link:
https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data