Assignment (MiniProject) 2

Machiry Aravind Kumar

UCSB

1 Dataset

1.1 Seeds Data Set

This dataset is from UCI Archive: `https://archive.ics.uci.edu/ml/datasets/seeds`. Dataset contains various Geometric parameters of wheat kernels measured using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The problem is to classify these into one of the 3 different classes viz Kama, Rosa and Canadian. For this Assignment I selected samples for only 2 classes: Kama (Class label: 1) and Rosa (Class label: 2). Available Features in the dataset are:

- area A.
- perimeter P.
- compactness C.
- length of kernel.
- width of kernel,
- asymmetry coefficient
- length of kernel groove.

To use LDA, I selected 2 features viz **area** and **perimeter**. Distribution of the corresponding features is as shown in figure: 1.

2 Parametric Values for the Datasets

2.1 Mammographic dataset

Mean For 2 classes :49.71317829, 62.25909091.
Standard deviation:185.59215191, 152.4828719.
The graph is same as the graph for Feature 2 in Figure **??**.

2.2 Wine dataset

Mean:1115.71186441, 519.50704225.
Standard deviation: 48239.7305372, 24367.26403491. The graph is same as the one in Figure **??**.

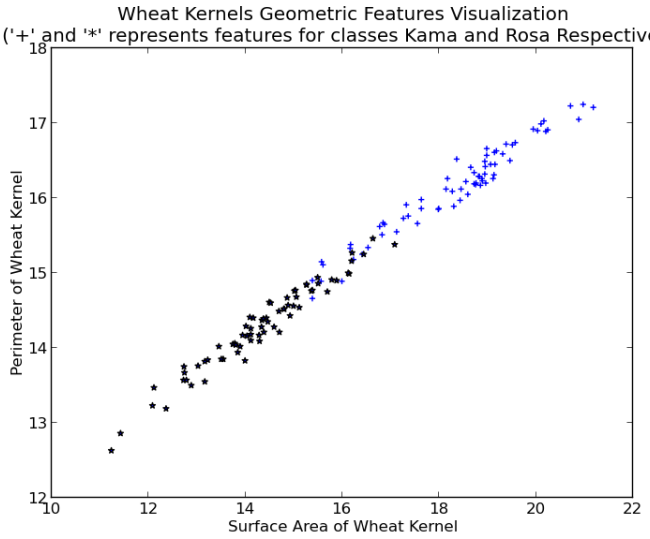

Figure 1: Dataset Features Visualization

3 Computing best dichotomy parameters

I used a modified binary search to find the right value for dichotomy, the code for which is present at: `parametric_est.py`. Values for the 2 datasets is as shown below:

3.1 Mammographic dataset

Best feature Dichotomy value: 58.25.

3.2 Wine dataset

Best feature Dichotomy value: 980.37109375.

4 Classification Accuracy

The accuracy for different values of n for n-fold validation on different datasets is shown below:

4.1 Wine dataset

The results for different values of n are as shown in Table 1

4.2 Mammographic dataset

The results for different values of n are as shown in Table 2

N	Mean Accuracy
3	0.923
4	0.931
5	0.923
6	0.922
7	0.923
8	0.923
9	0.923
10	0.923

Table 1: Cross Validation Results for Wine

N	Mean Accuracy
3	0.672
4	0.678
5	0.671
6	0.678
7	0.677
8	0.678
9	0.676
10	0.676

Table 2: Cross Validation Results for Mammographic