Assignment (MiniProject) 1

Machiry Aravind Kumar

UCSB

1 Datasets

I selected 2 datasets for this homework.

- Mammographic Mass Data Set
  This dataset is from UCI Archive: `https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass`. Mammography is the most effective method for breast cancer screening available today. The provided dataset has 961 samples, with 5-features and a classification of benign(0) or malignant(1). Features in the dataset are:
  - BI-RADS assessment: 1 to 5 (ordinal, non-predictive!)
  - Age: patient's age in years (integer)
  - Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
  - Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
  - Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
  - Severity: benign=0 or malignant=1 (binominal, target class)

  Based on the distribution of features as shown in Figure 1.

- Wine Dataset
  Again, this dataset is from UCI archive: `http://archive.ics.uci.edu/ml/datasets/Wine`. This data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Thus dataset has 13 features and for this homework I used Feature 13 (Proline value) and corresponding distribution is as shown in Figure 2. Also, This dataset has 3 classes (3 types of wines) but because of the restriction, I selected data for only 2 classes (i.e 2 types of wines: 1 and 2)

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 1.

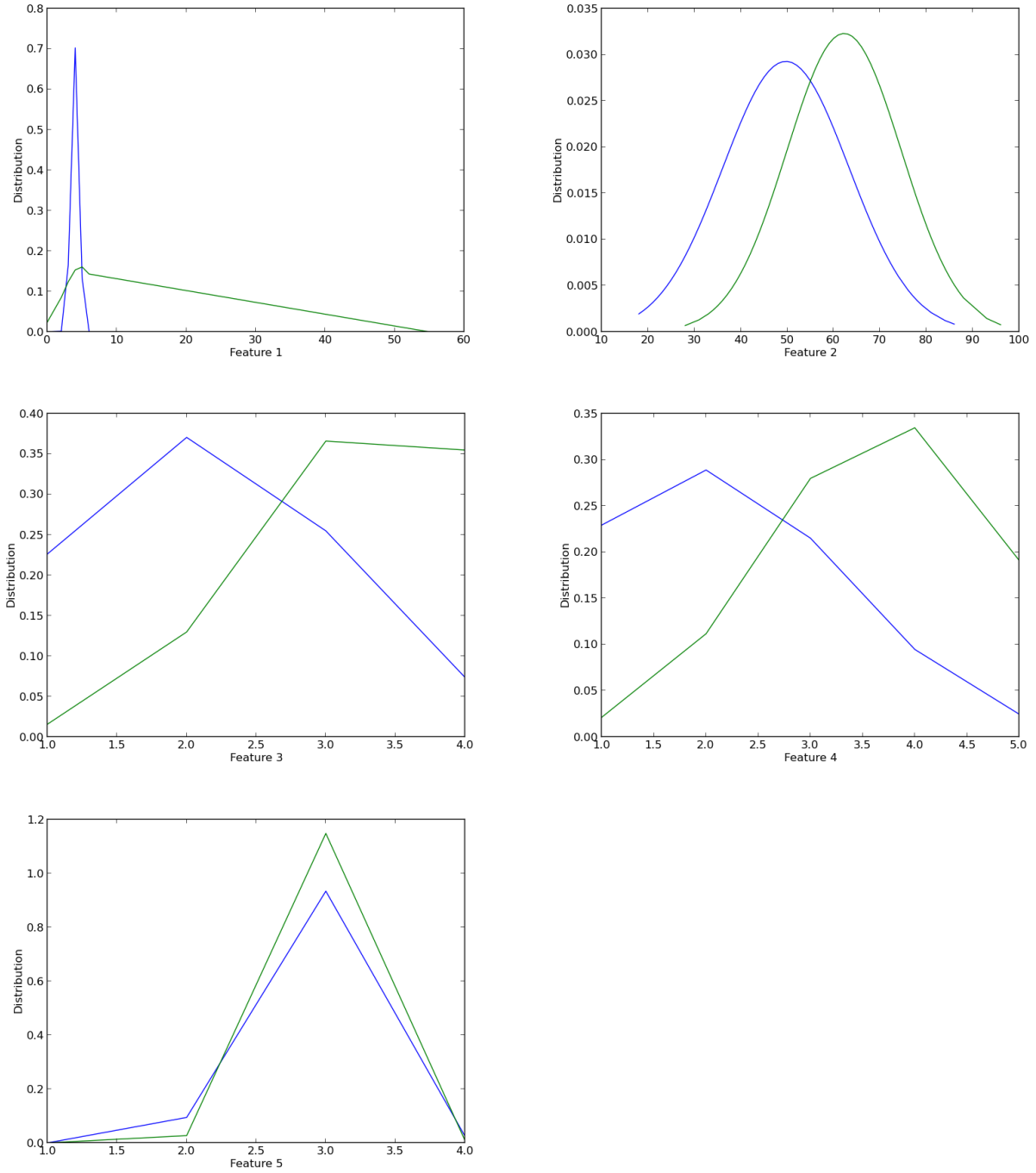

Figure 1: Mammographic features distribution

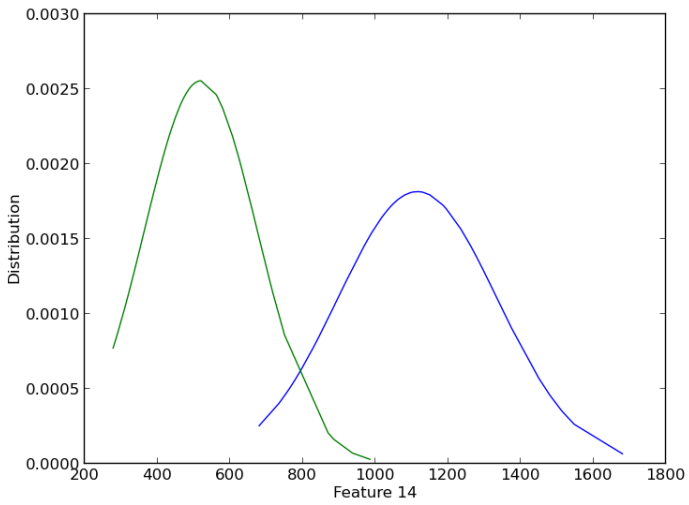

Figure 2: Feature Proline distribution

2.2 Wine dataset

Mean:1115.71186441, 519.50704225.
Standard deviation: 48239.7305372, 24367.26403491. The graph is same as the one in Figure 2.

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

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

4.2 Mammographic dataset

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