Assignment (MiniProject) 3

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.

2 Running SVM

I used SVM package from `sklearn` module. I used following kernel functions: `linear,rbf` and `sigmoid` with different values of Penalty Parameter(C) which trades off misclassification of training examples against simplicity of the decision surface, in other words this parameter controls over fitting, larger value of C results in over fitting and Kernel coefficient (gamma) which defines how much influence a single training example has. Its clear that, we need to select a kernel function which gives highest accuracy with least C and gamma. The error rates of using different kernel functions with different C and gamma are as shown in Table 1. I selected kernel function `linear` with $C = 1.4$ and $gamma = 0$ as it gives the best result satisfying our constraints.

Kernel Function	**C**	**gamma**	**Mean Accuracy**
rbf	0.1	0	50.0
sigmoid	0.1	0	93.57
linear	0.1	0	94.28
linear	1	0	94.28
linear	1.1	0	94.42
linear	1.2	0	95.0
linear	1.3	0	95.71
linear	**1.4**	**0**	**96.42**
linear	1.5	0	96.42
linear	10	0	96.42

Table 1: Effectiveness of various Kernel Functions

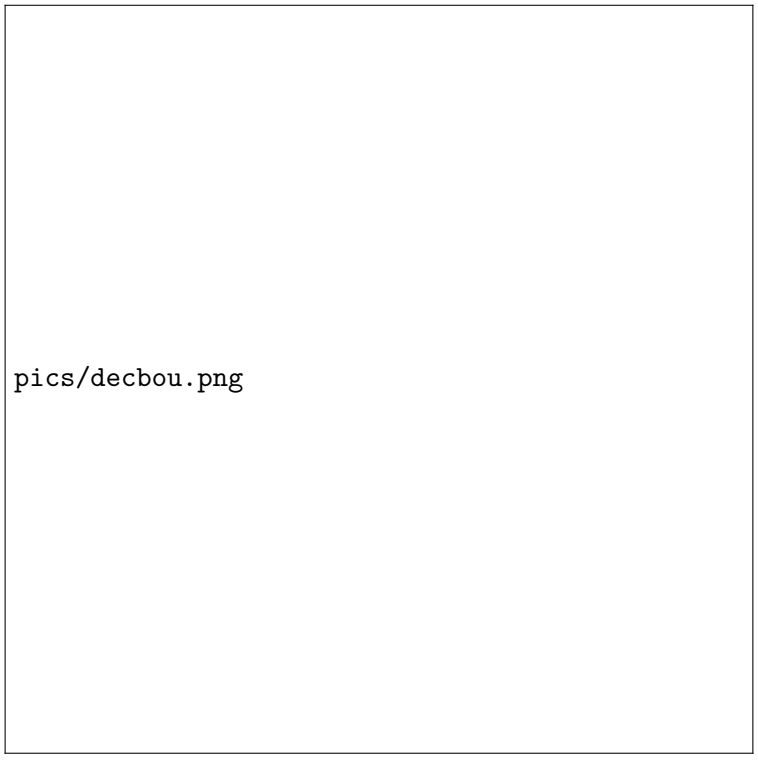

Figure 1: Dataset Features with Decision Line

N	Mean Accuracy	Mean Error
3	91.3	8.63
4	91.99	8.00
5	91.42	8.57
6	91.16	8.83
7	91.42	8.57
8	90.97	9.02
9	90.97	9.02
10	91.42	8.57

Table 2: Cross Validation Results for Seeds Dataset using LDA

3 Running LDA classifier on the Dataset

I ran LDA function from sklearn.ida module with default arguments as shown below:

```
classifier = LDA()
classifier.fit(features, classes)

weights = classifier.coef_[0]
dec_s = weights[0] / weights[1]
x = np.linspace(10, 22)

Line Equation:
y = dec_s * x    (classifier.intercept_[0] / weights[1])
```

It provided weights and intercept using which I was able to get line equation of the decision boundary. The resulting line equation for Decision boundary is: **y = -0.011*x - 15.02**. Figure 1 shows the features with decision line.

4 Classification Accuracy

The results for different values of n are as shown in Table 2. The results are pretty accurate with 91% Accuracy and doesn't vary much with n, which proves that features are good for the classification. It also proves that a linear classifier is good fit for this classification problem.