Chiranjeev
P21CS007
chiranjeev.1@iitj.ac.in

Assignment - 3
Support Vector Machines

22-09-2021

Hard SVM

I. Dataset Description for Hard SVM:

a. Name of Dataset : Iris Dataset (link - Dataset link)

b. About Dataset: The Iris dataset consist of 3 classes but for Hard SVM we need binary classification so we reconstructed this dataset by removing some data points which belong to class 2(Iris Virginica) and kept 2 classes class 0 (Iris Setosa) and class1(Iris Versicolour). The dataset is also Linearly Separable.

c. Dataset features :

- Sepal length – in cm
- Sepal width – in cm
- Petal length – in cm
- Petal width – in cm

d. Data Pre-processing :

Found that there are not any Not Available (NA) values in the dataset.

e. Features Selected :

All features mentioned above are chosen for training and testing.

f. Target Value to be Predicted :

class -1 (formerly class 0) : Iris Setosa
class 1 (formerly class 1) : Iris Versicolour

II. Splitting the Dataset for Hard SVM:

Used train_test_split of sklearn to split the dataset into train and test.

Split the Dataset into: 70% - train set, 30% test set

III. Hard SVM implementation using CVXOPT Quadratic solver

A. **Formulation Used :** Dual Formulation

B. **Number of Support Vectors :** 4

C. **Training Data points which are Support Vectors :**

```
[[4.8 3.4 1.9 0.2 1. ]
 [4.5 2.3 1.3 0.3 1. ]
 [5.1 2.5 3.  1.1 1. ]
 [5.1 3.3 1.7 0.5 1. ]]
```

D. **Margin distance :** 0.81755

E. **Training Accuracy :** 100.0 %

F. **Testing Accuracy :** 100.0 %

G. **Confusion Matrix for test set :**

```
confusion_matrix:
 [[17  0]
 [ 0 13]]
```

Soft SVM

I. Dataset Description for Soft SVM:

a. **Name of Dataset :** Breast Cancer Wisconsin (Diagnostic) Dataset (link - Dataset link)

b. **About Dataset :** Features in the data are computed from a digitalized image of a fine needle aspirate (FNA) of breast mass that describe characteristics of the cell nuclei present in the image in the 3-dimensional space.

c. **Dataset features :**

```
id
diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave points_worst
symmetry_worst
fractal_dimension_worst
```

d. **Features Dropped :**
Feature ‘id’ is dropped as for classification task id is not an attribute of breast.
And also found that there are not any Not Available (NA) values in the dataset.

e. **Features Selected :**
All features mentioned above are chosen except ‘id’ for training and testing as all features describes the attributes of breast for cancer detection.

f. **Train-Test split:**
Train set : Test set = 80:20

II. Soft SVM implementation using CVXOPT Quadratic solver

A. Formulation Used : Dual Formulation

Regularization Parameter(C)	Number of Support Vectors	Margin distance	Training Accuracy	Testing Accuracy	Confusion Matrix for Test set
0.1	54	0.6936	98.2417%	98.2456%	confusion_matrix: [[41 2] [0 71]]
0.01	105	1.4587	96.7032%	97.3684%	confusion_matrix: [[40 3] [0 71]]
0.001	218	3.0388	81.0989%	81.5789%	confusion_matrix: [[22 21] [0 71]]

III. Soft SVM with SGD implementation

Regularization Parameter(λ)	Number of Support Vectors	Margin distance	Training Accuracy	Testing Accuracy	Confusion Matrix for Test set
0.1	82	1.0933	97.142%	98.245%	confusion_matrix: [[41 2] [0 71]]
0.01	12	0.3941	98.241%	98.245%	confusion_matrix: [[42 1] [1 70]]
0.001	17	0.4284	97.142%	97.368%	confusion_matrix: [[41 2] [1 70]]