DHRUV GARG 16 BCE 1190

SUPPORT VECTOR MACHINE USING SKLEARN

LAB 5

Support vector machine theory:

Support Vector Machine (SVM) is a supervised learning algorithm which can be used for both classification or regression. In SVM we plot all the data points in the n-dimensional space, with the value of each feature being a coordinate. Then, we perform classification by finding a hyper-plane which will separate the classes well. We try and find a plane which separates the classes as well as has the **largest possible margin on both sides** with respect to the **support vectors**. Also, SVM has a feature of ignoring the outliers, ie. SVM is robust to outliers. For classifying data that is not linearly separable, the SVM has a technique called the **kernel trick**, which converts a non-separable problem to a separable problem. Important parameters in SVM:

a. **Kernel:** various options available – linear, rbf and poly.

b. **Gamma:** Higher value of gamma will try to exact fit the training data which could lead to overfitting and cause generalization error.

c. **C:** This is the penalty parameter of the error term. It controls the trade off between smooth decision boundary and classifying the training points correctly.

PART 1: Lab implementation

First, we created 50 sample points in two clusters using **make_blobs**. The two clusters were linearly separable and we tried different line equations that could act as the decision boundary. We chose the decision boundary that gave the largest margin on both sides. Next we used sklearn's SVC and varied the parameters to get the optimal decision boundary. We then implemented the same for non-linearly separable data using **make_circles**. The model was fit for this data using the 'rbf' kernel and it did a fairly good job. Lastly we used **sklearn's SVR** to fit data with noise using linear, poly and rbf kernels. **Rbf** performed the **best**, followed by poly. 1.

PART 2: ABSENTEEISM FROM WORK Dataset

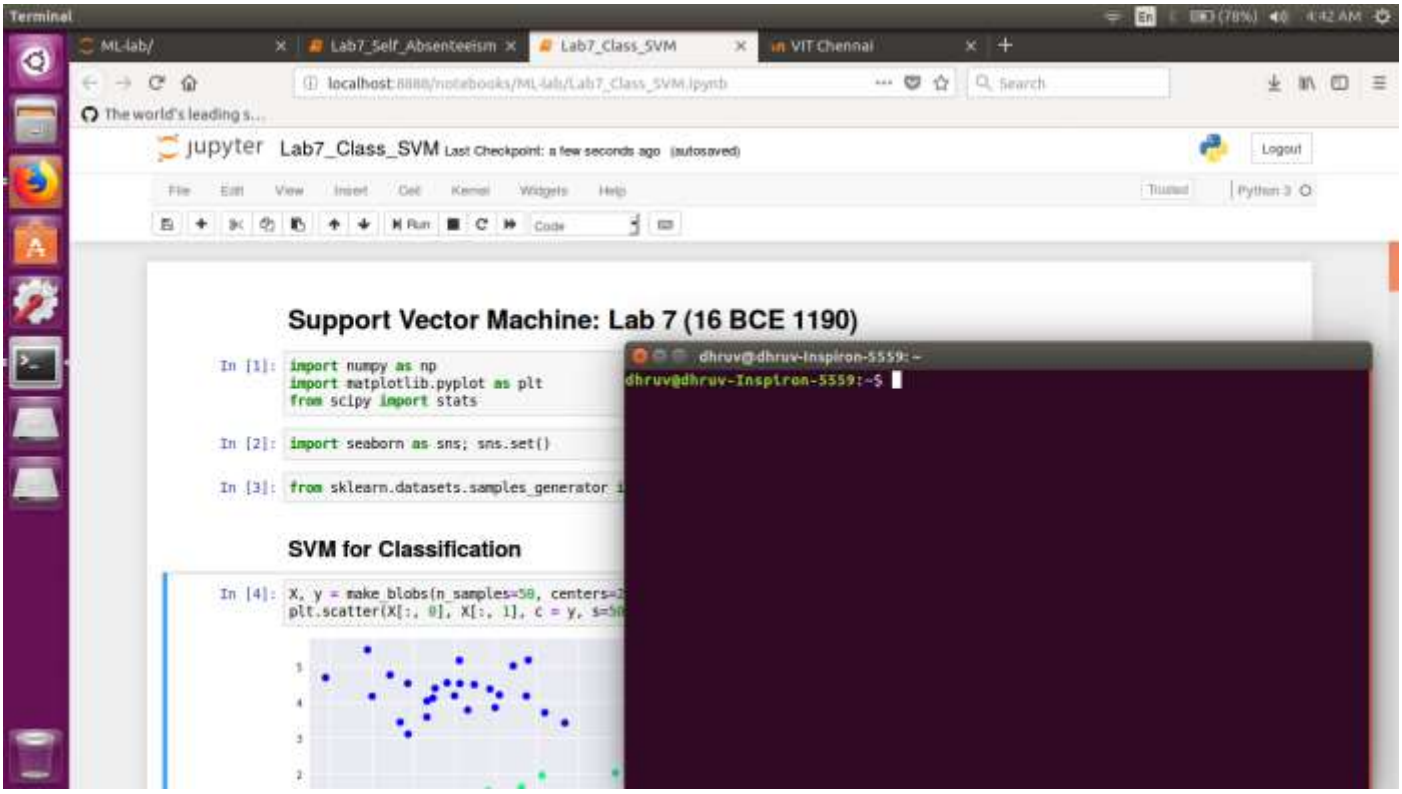
For classification I performed **multi-class classification** on categories of education. The rows were extracted and the attributes that linearly separated the pet classes were identified. In this, I wanted to perform classification between 3 classes. I first visualized the way in which the different kernels created decision boundaries using **meshgrid**. **Rbf** kernel performed the **best**. Next, I used GridSearchCV on the list of parameters chosen, and observed the accuracy scores. The accuracy of the model increased for higher values of C and gamma. **Best classifier** was found with the following parameters: C=10, gamma=1 and kernel=poly, with an **accuracy of 93%**. For regression I used the columns of Service time and age. Using **sklearn's SVR**, it was seen that the rbf kernel did a better job than the poly kernel. Next, I used GridSearchCV on the list of parameters chosen, and observed the accuracy scores. The accuracy of the model was seen to have no obvious correlation with c or gamma parameters. **Best classifier** was found with the following parameters: C=10, gamma=1 and kernel=rbf.

PART 3: Interactive SVM

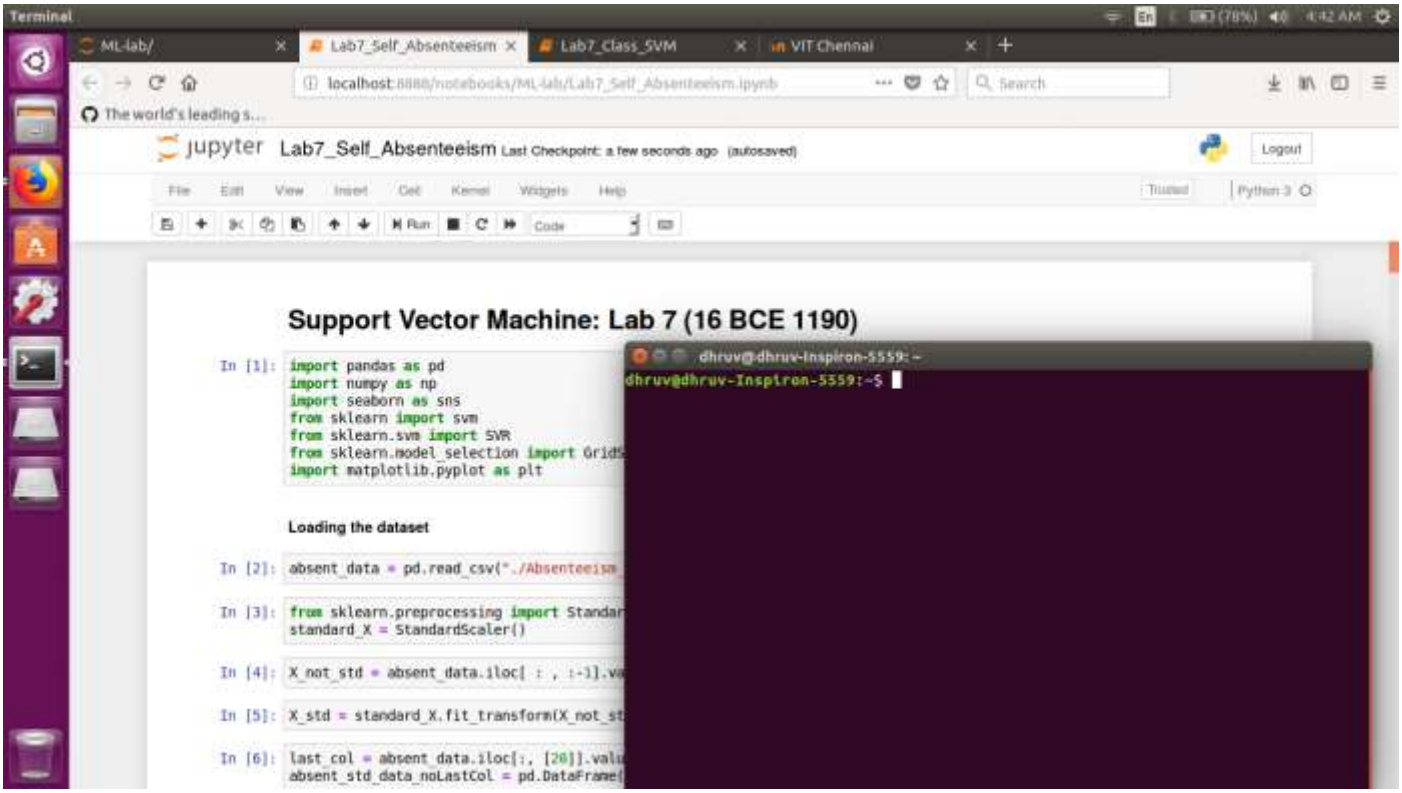
I chose the **CIRCLES dataset** and observed the plots on changing noise, cost and gamma. **Rbf was the only kernel to perform reasonably well.**

SCREENSHOTS

Lab implementation



Chosen dataset – Absenteeism at work



DHRUV GARG 16 BCE 1190

Interactive SVM