SVM Lab Notes

Ahman Smith

Kernel

- Different ways to split or classify your data

Using Social Network Ads CSV dataset, reading it in with pd.read_csv function.

Splitting the variables X and Y into independent and dependent variables. We are trying to predict the salary of the individuals within the dataset, so that will be our Y.

Split data up in typical fashion, using 25% of data for test size.

Important Note About Scaling

Feature scaling is especially important within this dataset, because age and salary are not comparable and would thus create an inaccurate distribution. We need to fix it! We're going to use the sklearn preprocessing standardscaler, which uses the standard deviation. We applied feature scaling on X_train and X_test via sc.fit_transform(). This will scale and normalize our data.

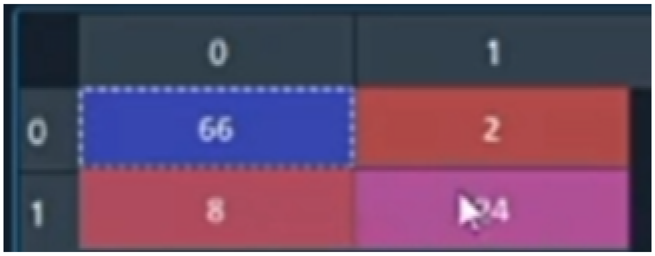
Applying SVM

We import SVM from sklearn.svm, and then use classifier class. We're using a linear kernel for this lab.

Confusion Matrix

Import confusion matrix from sklearn.metrics, compare the y_test data to y_pred data in order to see how off we've been.

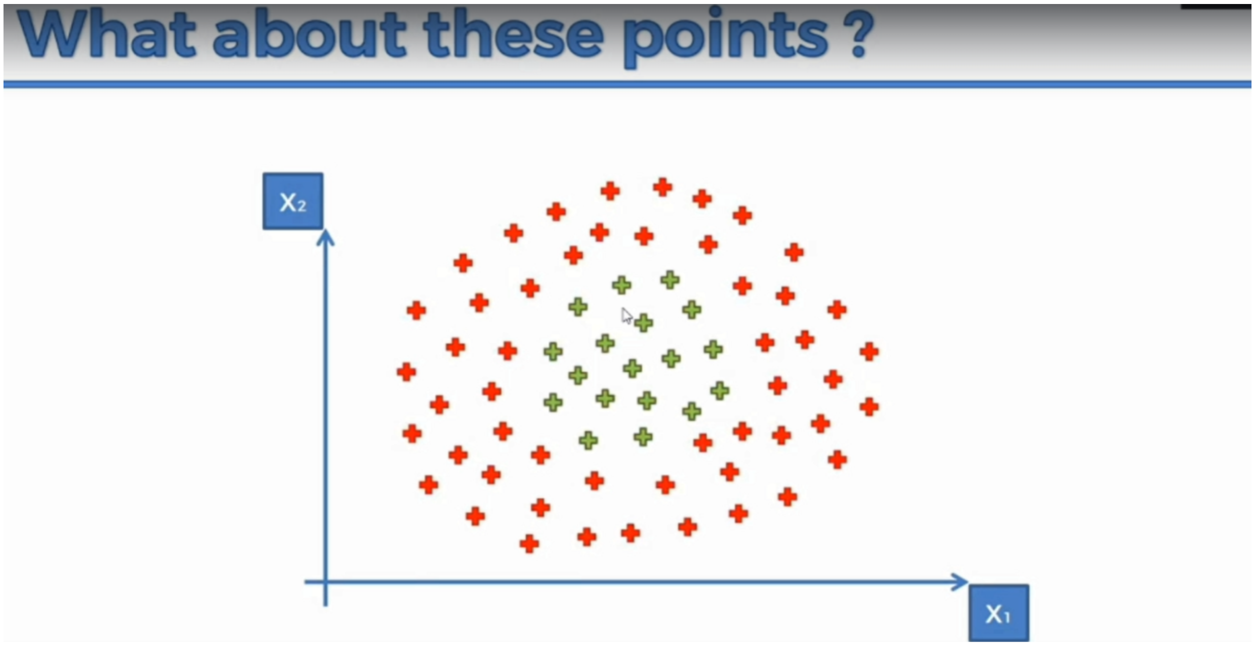
After printing it out,



	0	1
0	66	2
1	8	24

We get an accuracy of 90%, given that the diagonal adds to 90 (66+24). The other squares are inaccurate predictions.

Other Situations



What about these points ?

In these situations, we clearly do not want to use a linear kernel. The data is not linearly separable!

Different Kernels try and make it linearly separable using their approaches.