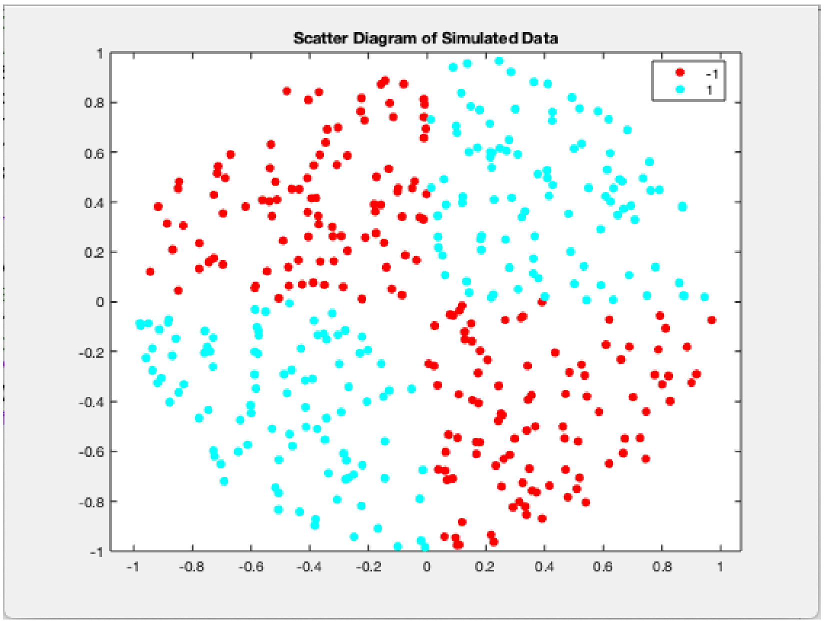
HOMEWORK 1 - VIVEK YADLAPALLI (Personal SEED - 220)

Two problems are solved using the Support Vector Machine (SVM) model.
An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes
For inseparable classes, the objective is the same, but the algorithm imposes a penalty on the length of the margin for every observation that is on the wrong side of its class boundary.

Question 1

First, a random set of 320(100+220) points (random seed) in a unit circle are generated. In this classification problem the first and third quadrants belong to the positive class (points in blue), and those in the second and fourth quadrants to the negative class (points in red).

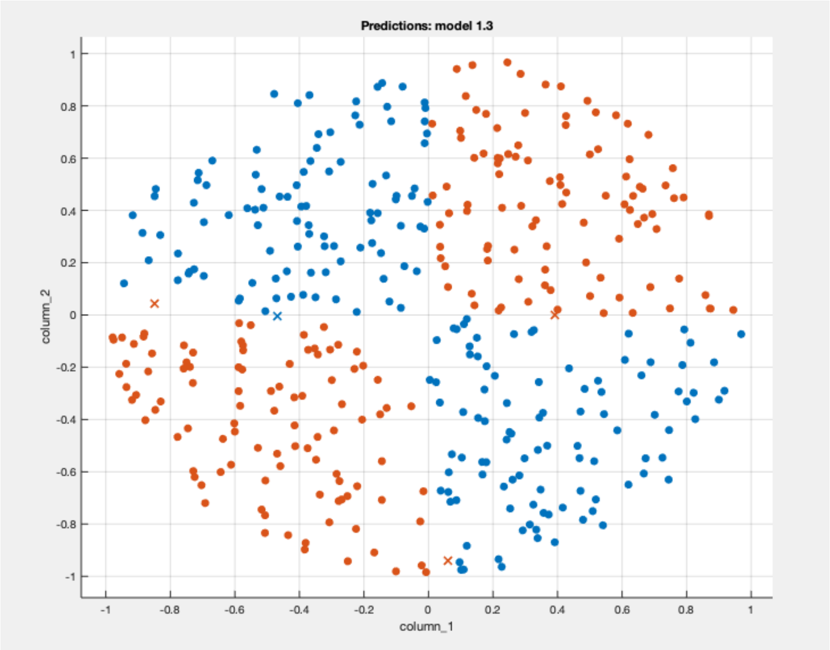


Training model

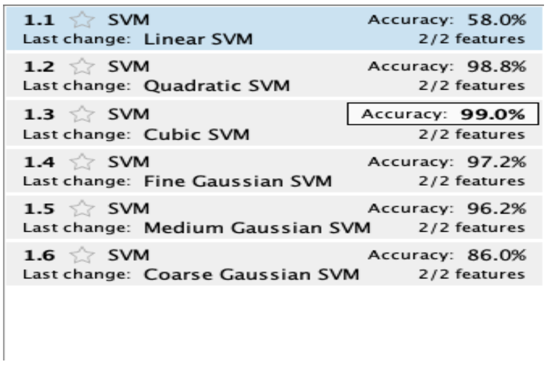
Using the generated sample, we do a full SVM training for it with 5-fold cross validation. 5-fold cross validation means that the original sample will be split into k subsamples of equal size. One of the k subsamples is reserved for data validation to test the model while the remaining k-1 subsamples are used to train the model. This cross-validation process is repeated k times so that each of the k subsamples is used exactly once as a validation data.
The SVM classification score for classifying observation x is the signed distance from x to the decision boundary. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

Obtained model

Having completed the training of the model, we obtain the model that contains the optimized parameters from the SVM algorithm, allowing to classify new data. The cubic SVM shows the highest accuracy of 99.0% (4 points out of 320 were predicted incorrectly).



It should be noted that the linear SVM has only 53.2% accuracy that is not surprising since it is not possible to separate the data points by a straight line.



1.1 ☆ SVM Accuracy: 58.0%
Last change: Linear SVM 2/2 features

1.2 ☆ SVM Accuracy: 98.8%
Last change: Quadratic SVM 2/2 features

1.3 ☆ SVM Accuracy: 99.0%
Last change: Cubic SVM 2/2 features

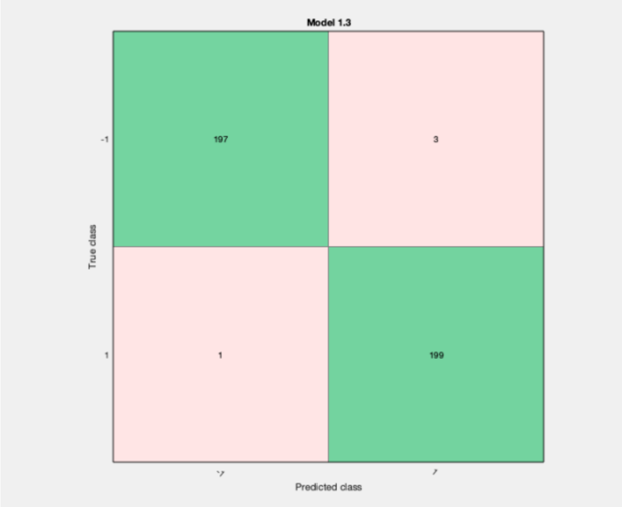
1.4 ☆ SVM Accuracy: 97.2%
Last change: Fine Gaussian SVM 2/2 features

1.5 ☆ SVM Accuracy: 96.2%
Last change: Medium Gaussian SVM 2/2 features

1.6 ☆ SVM Accuracy: 86.0%
Last change: Coarse Gaussian SVM 2/2 features

Confusion matrix and Model Accuracy (in-sample)

Using the confusion matrix allows to understand how the currently selected classifier performed in each class. The confusion matrix helps you identify the areas where the classifier has performed poorly, in particular, see how many observations the model predicted as positive and it is false and how many observations the model predicted as negative and it is false.



It can be seen that the cubic SVM model performed in-sample relatively well. 3 observations the model predicted as positive and it is false and 1 observations the model predicted as negative and it is false. 197 observations the model predicted as positive and it is true and 199 observations the model predicted as negative and it is true.

The accuracy of the model tested on the out-of-sample data is 98.5% (only the class of 6 points was predicted incorrectly).

Question 2

The classification problem: given a set of observations (vectors in the feature space) assign each point to one of two classes.

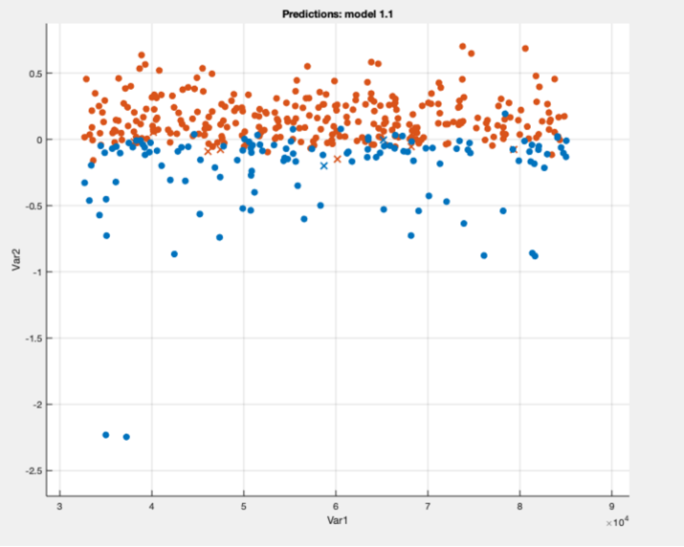
In this classification problem we have a class 1 assigned for all credit ratings except ‘CCC’ and class -1 for ‘CCC’. From the whole provided we created a sample a random subsample of 520(300+220) observations with label 1 and 130 observations with label -1.

Training model

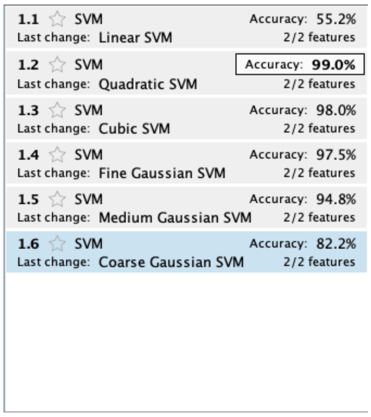
Using the generated subsample, we do a full SVM training for it with 5-fold cross validation. 5-fold cross validation means that the original sample will be split into k subsamples of equal size. One of the k subsamples is reserved for data validation to test the model while the remaining k-1 sub-samples are used to train the model. This cross-validation process is repeated k times so that each of the k sub-samples is used exactly once as a validation data.

The SVM classification score for classifying observation x is the signed distance from x to the decision boundary. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

Obtained model



Having completed the training of the model, we obtain the model that contains the optimized parameters from the SVM algorithm, allowing to classify new data. The quadratic SVM shows the highest accuracy of 99.0%.



1.1 ☆ SVM Accuracy: 55.2%
Last change: Linear SVM 2/2 features

1.2 ☆ SVM Accuracy: **99.0%**
Last change: Quadratic SVM 2/2 features

1.3 ☆ SVM Accuracy: 98.0%
Last change: Cubic SVM 2/2 features

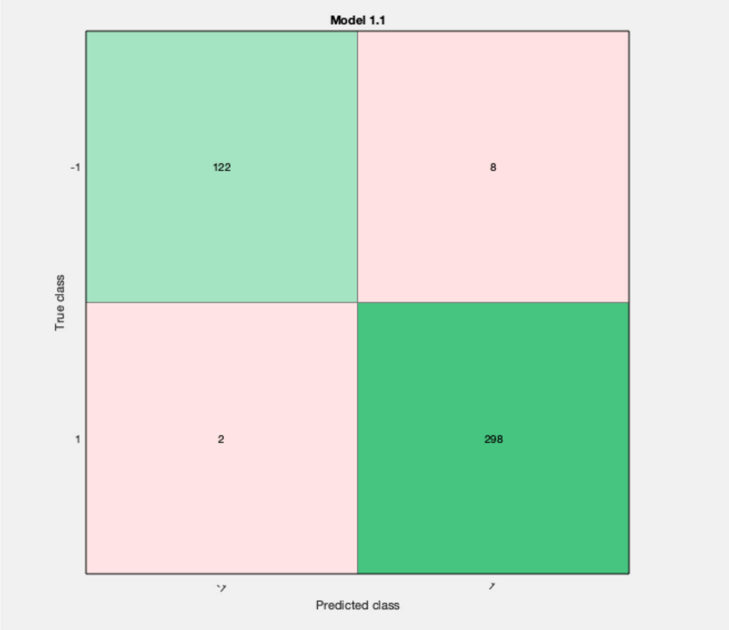
1.4 ☆ SVM Accuracy: 97.5%
Last change: Fine Gaussian SVM 2/2 features

1.5 ☆ SVM Accuracy: 94.8%
Last change: Medium Gaussian SVM 2/2 features

1.6 ☆ SVM Accuracy: 82.2%
Last change: Coarse Gaussian SVM 2/2 features

Confusion matrix and Model Accuracy (in-sample)

Using the confusion matrix allows to understand how the currently selected classifier performed in each class. The confusion matrix helps you identify the areas where the classifier has performed poorly, in particular, see how many observations the model predicted as positive and it is false and how many observations the model predicted as negative and it is false.



It can be seen that the model in-sample performed relatively well. 8 observations the model predicted as positive and it is false and 2 observations the model predicted as negative and it is false. 122 observations the model predicted as positive and it is true and 298 observations the model predicted as negative and it is true.

The accuracy of the model tested on the out-of-sample data is 99.0% (only the class of 9 points was predicted incorrectly).