Support Vector Machine

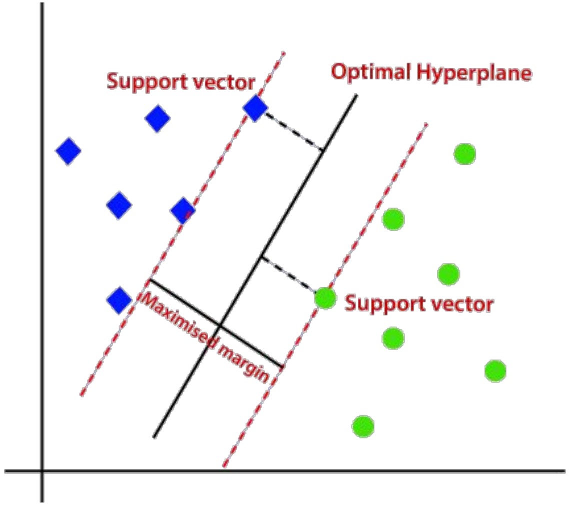
Stanley Perez, Yiqing Guo, Dean Papadopoulos, Seungmin Kim

Overview

- Support Vector Machine Explained
- Pros and Cons
- Data Processing Steps
- Hyperparameters

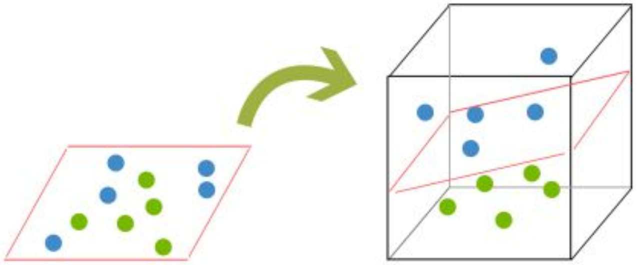
What is Support Vector Machine Algorithm?

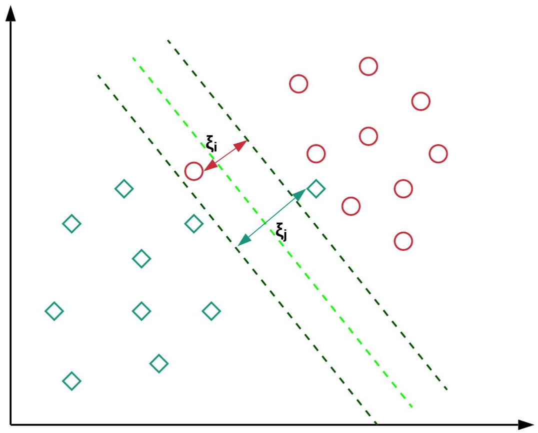
- SVM is a supervised algorithm used for classification and regression, often referred to as a "maximum margin classifier".
- The "Decision Boundary" is the optimal line/ plane.
- A support vector is defined by the data point(s) of a classification closest to the opposite classification.
- Support vectors define the algorithm's Margin.



Explanation Continued

- Soft Margins are used to account for outliers.
- Cross validation is used to determine the best soft margin.
- When a soft margin is used to determine decision boundary it is called a support vector classifier.

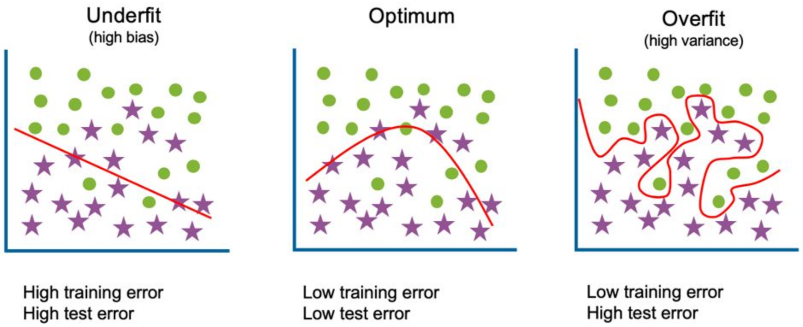




Advantages and Disadvantages

- Provides a clear separation of data between classes
- SVMs are effective when there is a high number of features
- Uses a subset of the training data to support the hyperplane
- Allows for kernel functions to be used depending on the data set

- Low performance if classes are not distinct
- No direct probability estimates from the model
- If the number of features is greater than observations, overfitting will be an issue



Data Cleaning

- sklearn.svm.SVC() takes in two arrays
  - Training samples: X of shape (# of samples, # of features)
  - Class labels: Y of shape (# of samples)
- Needs to be arrays with all numerical values -> use impute or drop NaN to get rid of NaN values
- What are class labels?
  - Numerical values that separate observations into separate categories
  - No upper limit to amount of different class labels

Hyperparameters

- C
- Gamma
- Kernel

C

In a SVM you are searching for two things:

1. a hyperplane with the largest minimum margin
2. a hyperplane that correctly separates as many instances as possible.

The problem is that you will not always be able to get both things. The c parameter determines how great your desire is for the latter. To the left you have a low c which gives you a pretty large minimum margin (purple). However, this requires that we neglect the blue circle outlier that we have failed to classify correct. On the right you have a high c. Now you will not neglect the outlier and thus end up with a much smaller margin.

Sources

- https://scikit-learn.org/stable/modules/svm.html
- https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe
- https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
-