Connor Lydon
CPSC393-01
Professor LaHaye
2322488

Support Vector Machines Writeup

What is a SVM: A Support Vector Machine (SVM) is an algorithm that classifies (or regresses) data through creating a line through the data. This line acts as a classification boundary to decide if a point lies in a given class. This boundary is created through fitting parallel lines that classify the data and have the most distance between them, the decision boundary is the line between those support vectors. Depending on the parameters, the support vectors will allow for some misclassifications and will change the strength of each point in the data. Though not all data is linearly separable, transformations can be used to turn data into linearly separable data. The decision boundary in the data is a hyperplane because it can work on n dimensions.

How SVM is used in this data: The desired goal is to classify this dataset into iris-setosa and non-iris-setosa. Plotting the data by all combinations (see code pdf) of two variables indicates that the data is linearly separable. This means that a SVM will be able to use a linear division between the data to allow for future classifications. Running a Support Vector Classified (SVC) with a linear kernel indicates that it does a great job fitting to the data. In 5 shuffled K-Fold iterations, it does a near perfect job in each fold. Though Gamma and Lambda C shouldn't need much tuning, I still ran a Grid Search that searched on different magnitudes of Gamma and Lambda as well as searching across the common kernel functions.

Lambda is a variable that indicates the amount of misclassifications the model is allowed to make in the training. This variable exists to get a decision boundary that is able to work with future data. If the model is focused on creating a decision boundary that correctly classifies everything then it will be so specifically fit to the data that given new, different data, it won't be able to make accurate predictions in the data. Gamma is a variable that indicates the impact of a single variable on the training of the model. Gamma deals with the curve of the decision boundary in certain kernels.

Our SVM perfectly classifies all the points except for one. For this reason the Lambda (Cost) is a bit higher to allow for the model to make a decision boundary that has high margins, but ignores this point.

Future uses of SVMs & capabilities: SVMs have capabilities to work with data that isn't so clean and to work with data that isn't linearly separable. By using different values of C (lambda) and different kernel functions, SVMs can be an effective classifier and show visually how it separates the data.