Assignment Part-2

Answer1:

The Support Vector Machine (SVM) is a linear classifier. Hyperplane is a line which separates the data into its classes. There could be many possible lines which perfectly divide the two classes.

The best line is the one that maintains the largest possible equal distance from the nearest points of both the classes.It is called as Maximal margin classifier.

The two major constraints necessary while maximising the margin are:

1.The standardisation of coefficients such that the summation of the square of the coefficients of all the attributes is equal to 1.

2.Along with the first constraint, the maximal hyperplane should also follow the constraint :

$$l_i X(W.Y_i) \geq M$$

The Soft Margin Classifier overcomes the drawbacks of the Maximal Margin Classifier by allowing certain points to be misclassified. You control the amount of misclassifications using the cost of misclassification 'C' , where C is the maximum value of the summation of the slack variable epsilon.

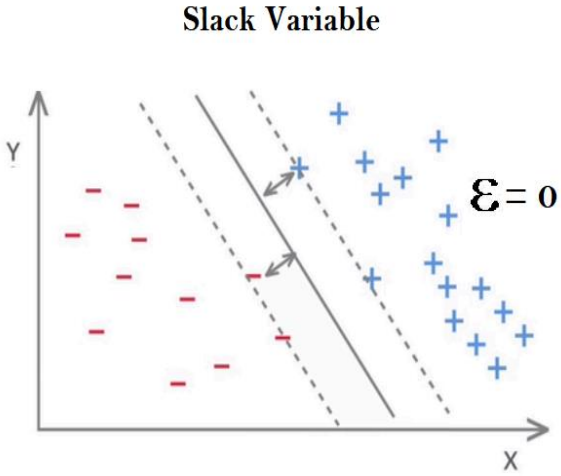
$$\sum \epsilon_i \leq C.$$

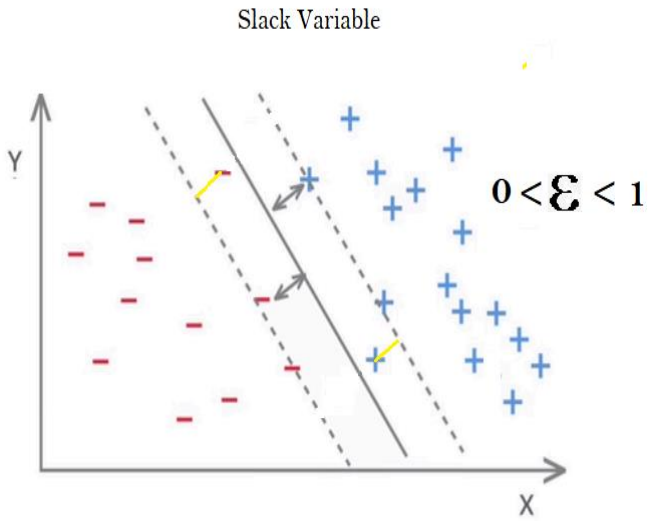
If C is high , a higher number of points are allowed to be misclassified or violate the margin. In this case, the model is flexible, more generalisable and less likely to overfit.

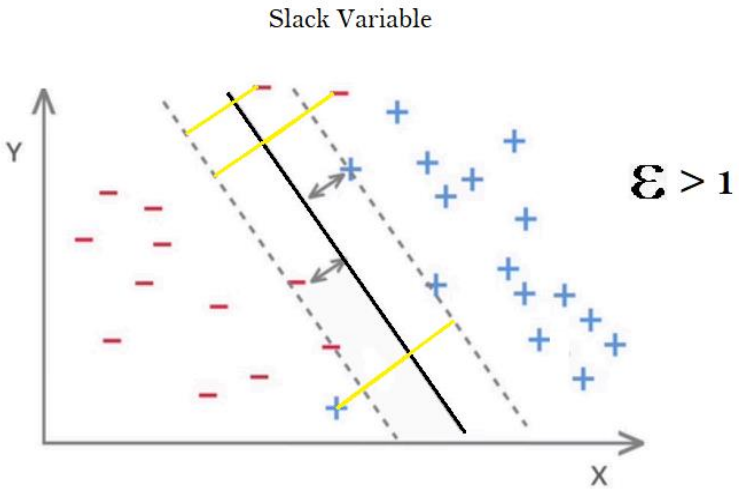
Answer2:

A slack variable indicated by (ε) is used to control misclassifications and it tells you where an observation is located relative to the margin and the hyperplane while constructing a soft margin hyperplane.

For points that are correctly classified, i.e,each observation is on the correct side of the margin,(ε) value is 0. If a data point is either correctly classified but falls inside the margin or violates the margin, then the value of (ε) lies between 0 and 1. If a data point is incorrectly classified, then the value of (ε) is > 1.







Answer3:

The summation of all the epsilons of each data point is denoted by 'C' also known as the cost function. Generally, the notion of slack variable ε needs to be understood first. This is followed by comparing any two support vector classifiers by measuring the summation of epsilons ε of both the hyperplanes and choose the best one that gives you the least value of summation.

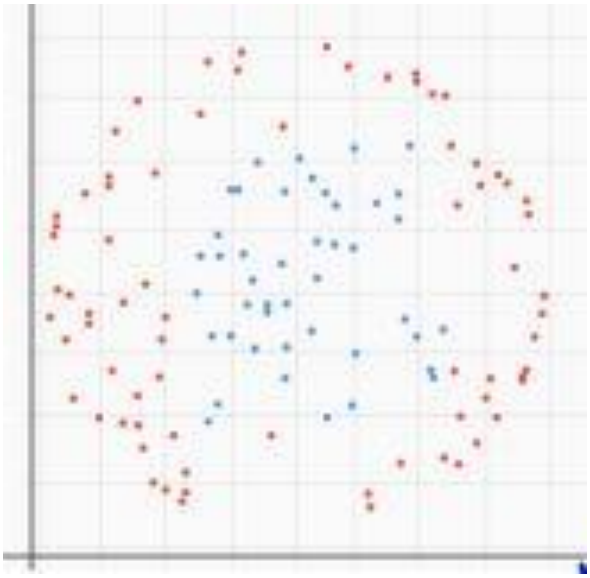
$$\sum \varepsilon_i > C$$

When C is large, the slack variables can be large, i.e. a larger number of data points are allowed to be misclassified or to violate the margin. The result is a hyperplane where the margin is wide and misclassifications are allowed. In this case, the model is flexible, more generalisable, and less likely to overfit. In other words, it has a high bias.

On the other hand, when C is small, many data points are not allowed to fall on the wrong side of the margin or the hyperplane. Hence, the margin is narrow and there are few misclassifications. In this case, the model is less flexible, less generalisable, and more likely to overfit. In other words, it has a high variance.

Assuming that the future dataset is unknown. While building a model, generally, it is a good idea to set the value of C to moderate due to the reason that if the value of C is very low, there will be no misclassifications, the model becomes less generalisable which may overfit the training data. Likewise, if the value of C is very high, many data points will be misclassified resulting in a bad model. If the SVM model is overfitting, add more training data points and increase the value of C.

Answer4:



The data points shown above are not linearly separated. Instead, they are intertwined and it is not possible to draw a linear hyperplane to separate the red and blue points. Kernels, enable the linear SVM model to separate nonlinearly separable data points by transforming nonlinear datasets into linear ones. This transformation is done by applying certain functions to the original attributes (X,Y) and create a transformed space or feature space (X', Y').

SVM models work well if the data is linearly separable. Hence, the transformation of the attributes is required to make the relationship between the variables linear when the dataset in hand is not linearly separated and instead the data points are intertwined. The final step after transformation is to run/execute the linear models.

The three most popular types of kernel functions are:

1.The linear kernel: This is the same as the support vector classifier, or the hyperplane, without any transformation at all
2.The polynomial kernel: It is capable of creating nonlinear, polynomial decision boundaries
3.The radial basis function (RBF) kernel: This is the most complex one, which is capable of transforming highly nonlinear feature spaces to linear ones. It is even capable of creating elliptical (i.e. enclosed) decision boundaries.

Answer5:

The process of transforming the original attributes into a new feature space is called Feature Transformation. However, As the number of attributes increases, there is an exponential increase in the number of dimensions in the transformed feature space.

Feature transformation results in a large number of features and hence makes the modelling computationally intensive.

Example: A data point (x,y) = (2,3), the transformed data point in the feature space say ($x,y,xy,x^2,y^2,1$) will be (2,3,6,4,9,1). This process is repeated for each data point. The last step is to run/execute the linear model on the transformed linear dataset.

Feature transformation results in a large number of features and hence makes the modelling computationally intensive.