SVM Assignment Questions

Q1. How is Soft Margin Classifier different from Maximum Margin Classifier?

Ans. The **Maximal-Margin Classifier** is a hypothetical classifier. It is the perfect classifier to best explain how SVM (Support Vector Machines) work in practice. The numeric input variables in the data-set form a n-dimensional space. A **hyperplane** is the line that splits the input variable space into different segments. The distance between the hyperplane and the closest data points of the dataset are referred to as the **margin**. The Maximum margin classifier is the line that maintains the maximum possible equal distance from the nearest points of both the classes in the n-dimensional space.

There can be many lines (hyperplanes) possible for the same data set with different values of margins. The line with the maximum margin would be considered the best fit for the given data, which is essentially what the maximum margin classifier does.

The Advantages of maximum margin classifier are:

1) Model becomes less biased
2) Number of training errors are reduced.

In practice, it is not possible the divide the points using just one line as data is messy. Therefor the constraint of maximizing the margin that separates the different classes must not be applied. This is often called the **Soft-Margin classifier**. This change allows some points in the training data to violate the separating line in order to have a better separation of data and capture pattern of data rather than overfitting on the dataset. An additional set of coefficients are introduced in the Soft-margin classifier that gives the margin some more space from the closest points. These coefficients are sometimes called slack variables. These slack variables are explained in the subsequent questions.

Q2. What does the slack variable Epsilon (ε) represent?

Ans. In Soft-margin classifier we introduce an additional set of coefficients that give the margin some room in each dimension. These coefficients are known as slack variables. However, this increases the complexity of the model as now there are more parameters for the model to fit to the data. Thus, lower slack values are preferred to higher values. When slack value is zero (**slack = 0**) implies a correct classification, but if slack greater than 1 (**slack>1**) implies an incorrect classification and when slack value is within 0 and 1(**0<=slack<=1**) classifies the point correctly but violates the margin.

Slack variable $\varepsilon = 0$: For points which are at least at a distance of more than M which means that the points are at a safe distance from the hyperplane, the value of the slack variable is 0.

Slack variable $0 < \varepsilon < 1$: For points which are correctly classified but violate the margin or falls inside the margin, this means that the value of its slack ϵ is between 0 and 1.

Slack variable $\varepsilon > 1$: Finally, for points that are incorrectly classified the slack value is greater than 1 (i.e. it violates the hyperplane), the value of epsilon (ϵ) > 1.

Q3. How do you measure the cost function in SVM? What does the value of C signify?

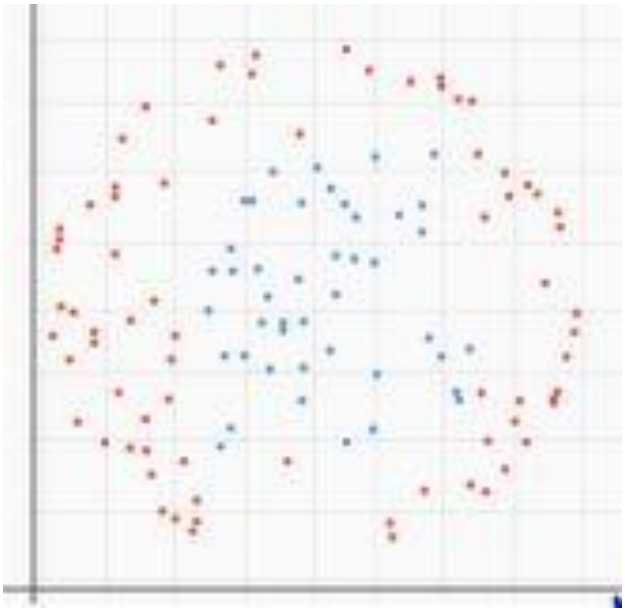
Ans. In the SVM algorithm, we are trying to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is known as the hinge loss. Its given as

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we will then have to calculate the loss value. We also need to add a regularization parameter to the cost function. The regularization parameter is mainly to balance the margin maximization and loss and provide a tradeoff between the two. After adding the regularization parameter, this is what the cost function looks like

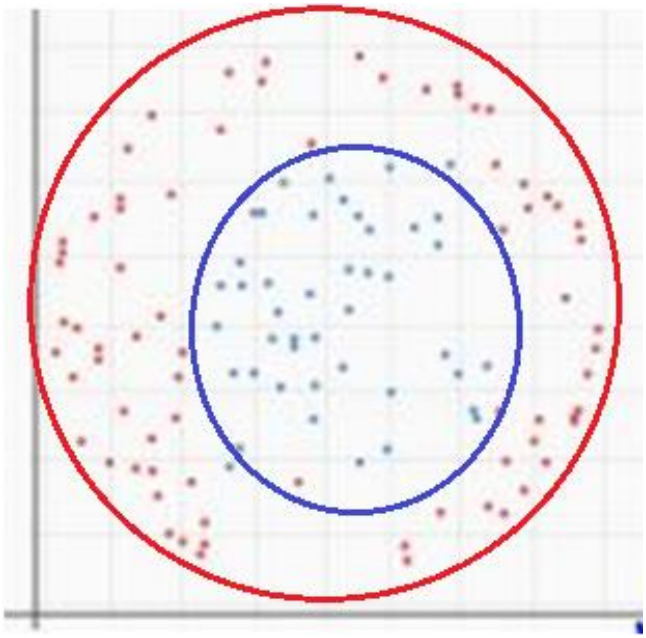
$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

C is a regularization parameter that controls the trade-off between achieving a low training error as well as maintaining a low testing error which is the ability to generalize your classifier so that it can work with unseen data. It defines the amount of violation of the margin that can be allowed across all dimensions. When C=0 there is no violation allowed which makes the classifier Maximal-Margin Classifier. The larger the value of C more are the number of wrongly classified points. Therefore, an optimum value of C must be used to maximize margin as well keep loss to minimum.



Q4. Given the above dataset where red and blue points represent the two classes, how will you use SVM to classify the data?

Ans. It is very evident from the above image that it is impossible to classify the dataset into two classes (red and blue) using a linear hyperplane. Hence, neither of the linear techniques, Maximum Margin Classifier or Soft Margin Classifier cannot be used. Another way to do it is to transform the data from 2-D attribute space to 3-D feature space. If we look closely you can see that the data can be classified as rings (ellipse) as shown below.



Thus after transformation we can apply the Maximum Margin or Soft Margin classification

Q5. What do you mean by feature transformation?

Ans. In practicality we will rarely have data which can be easily classified using any of the linear techniques. Like the diagram in the above question. In such cases we are going to need to **transform the points from 2-d attribute space to 3-D feature space**. In such a case the data is not linearly separable. However, this assumption is not correct. It is important to note here that the data is not linearly separable but only when it's plotted in two dimensions. Even if your original data is in two dimensions, there's no rule that prevents anyone from transforming it before feeding it into the SVM. This transformation will make the data linearly separable and this can be fed to SVM. One possible transformation would be, for instance, to transform every two-dimensional vector (x1, x2) into a three-dimensional vector. Which basically transforming data to a higher dimension.

For example, we can do what is called a polynomial mapping by applying the function $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined by:

$$\phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

This is what is known by Feature transformation.