SUBJECTIVE QUESTIONS

Question 1

How is Soft Margin Classifier different from Maximum Margin Classifier?

Solution 1

Maximal Margin Classifier

1. The best line that maintains the largest possible equal distance from the nearest points of both the **c**lasses is referred to as a **maximal margin classifier**.

2. It perfectly separates the two classes that is it does not allow for any misclassification of data points

3. It has a limited applicability

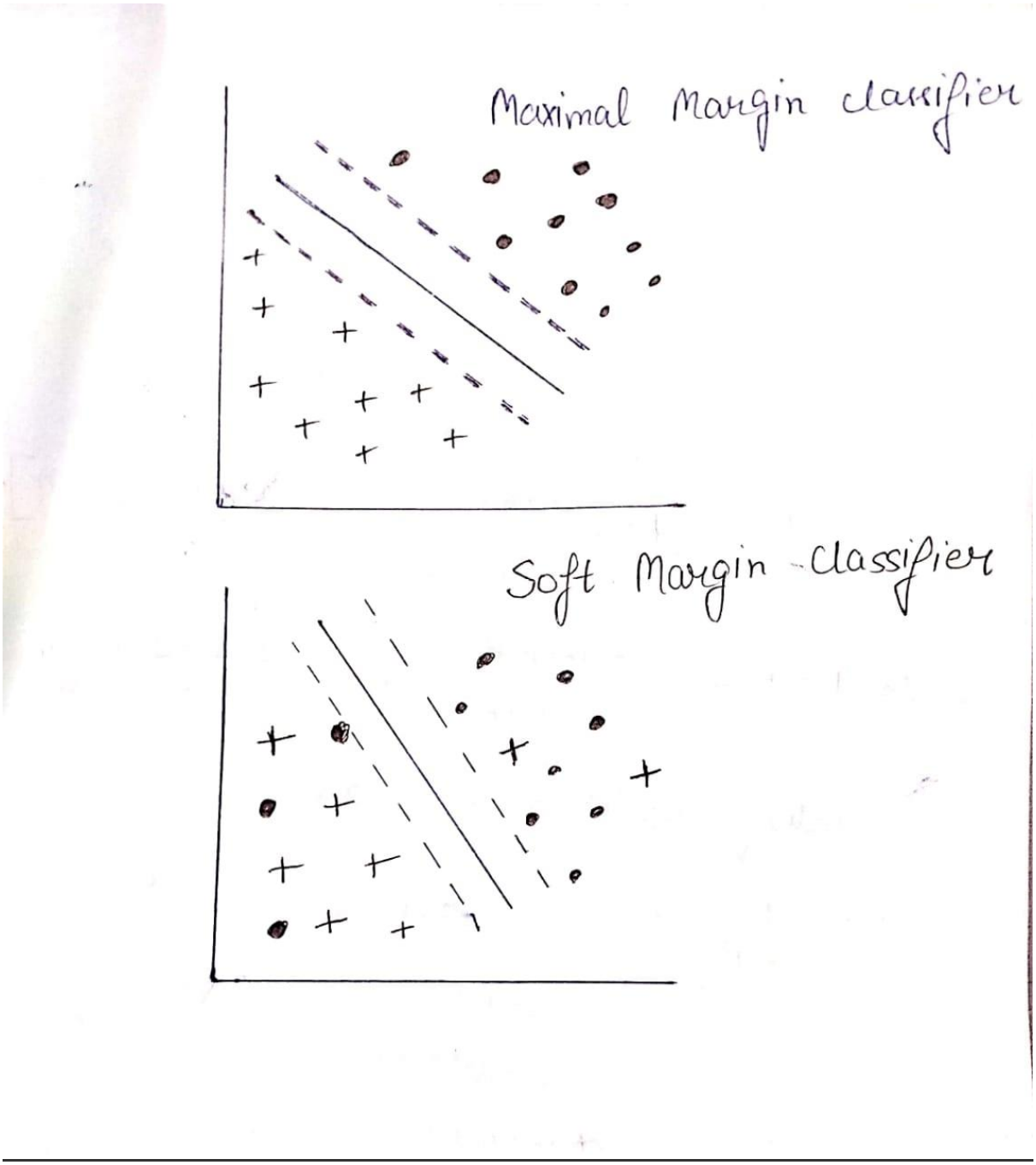
4. It cannot classify data points if they are **partially intermingled**

5. It is very sensitive to the training data, which may lead to over fitting in certain circumstances

Soft Margin Classifier

1. The constraint of maximizing the margin of the line that separates the classes must be relaxed in the soft margin classifier
2. It is meant for those data which cannot be separated into perfect classes. Hence, it allows for certain points to be misclassified to arrive at the best classifier
3. It is widely applicable as it handles majorly all types of datasets

4. It classifies data points even if they are **partially intermingled**
5. It allows some points in the training data to violate the separating line and not succumb to the perils of over fitting



Question 2

What does the slack variable Epsilon (ε) represent?

Solution 2

The value of ϵ defines a margin of tolerance where no penalty is given to errors. A slack variable is used to control **misclassifications**. It tells you where an observation is located relative to the margin and hyperplane.

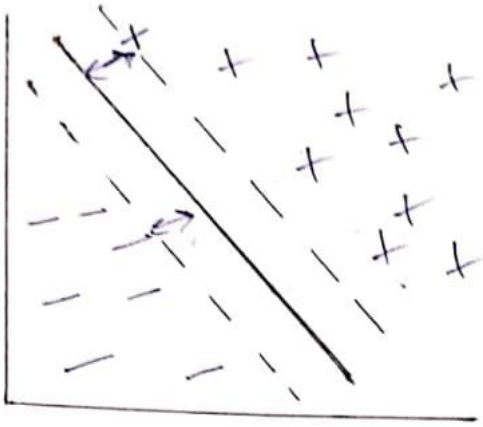
For points which are at a distance of more than M, i.e. at a safe distance from the hyperplane, the value of the slack variable is 0.

If a data point is correctly classified but falls inside the margin (or violates the margin), then the value of its slack ϵ is between 0 and 1.

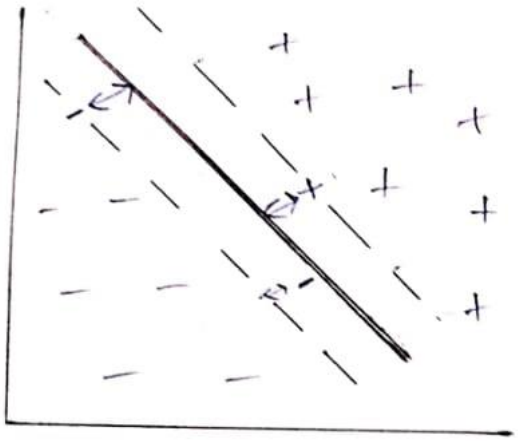
If a data point is incorrectly classified (i.e. it violates the hyperplane), the value of epsilon (ϵ) > 1.

Hence, **lower values of slack are better** than higher values and the purpose of the analysis is to determine a classifier while keeping in check the sum of the slack variables.

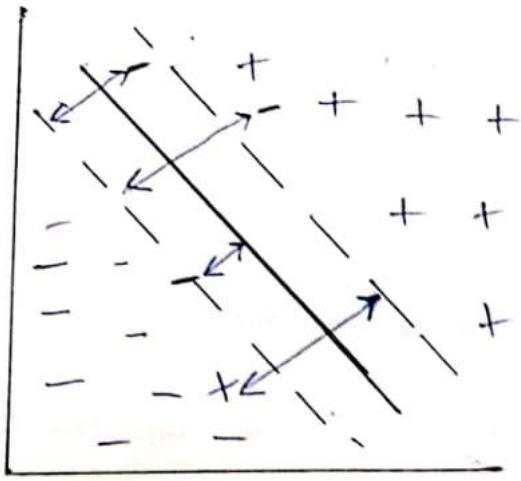
SLACK VARIABLE



$\varepsilon = 0$



$0 < \varepsilon < 1$



$\varepsilon > 1$

Question 3

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

Solution 3

The cost parameter in the SVM means: The trade-off between misclassification and simplicity of the model. It is one of the hyper parameters in SVM.

The cost parameter decides how much an SVM should be allowed to "bend" with the data. For a low cost, we aim for a smooth decision surface and for a higher cost, we aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

The summation of all the epsilons of each data point is denoted by cost or 'C', i.e.
$\sum \epsilon i \leq C$.

Misclassification can be controlled by the value of cost or C.

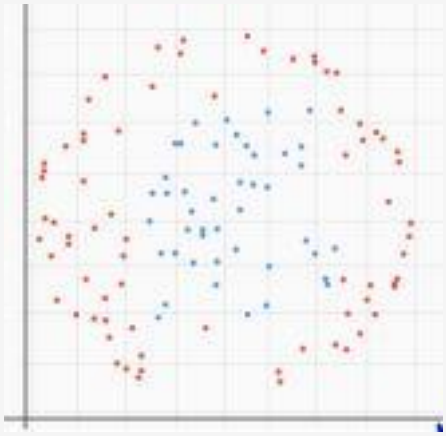
When **C is large**, the slack variables can be large i.e. you allow a larger number of data points to be misclassified or violate the margin.

In this case, the model is **flexible, more generalizable, and less likely to overfit**. In other words, it has a **high bias.**

When C is small, we force the individual slack variables to be small, i.e. we do not allow many data points to fall on the wrong side of the margin or the hyperplane.

In this case, the model is **less flexible, less generalizable, and more likely to overfit.** In other words, it has a **high variance**.

Question 4



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

Solution 4

We can figure out from the image that it is a non-linear distribution of data points.

And we know that SVM is a linear model.

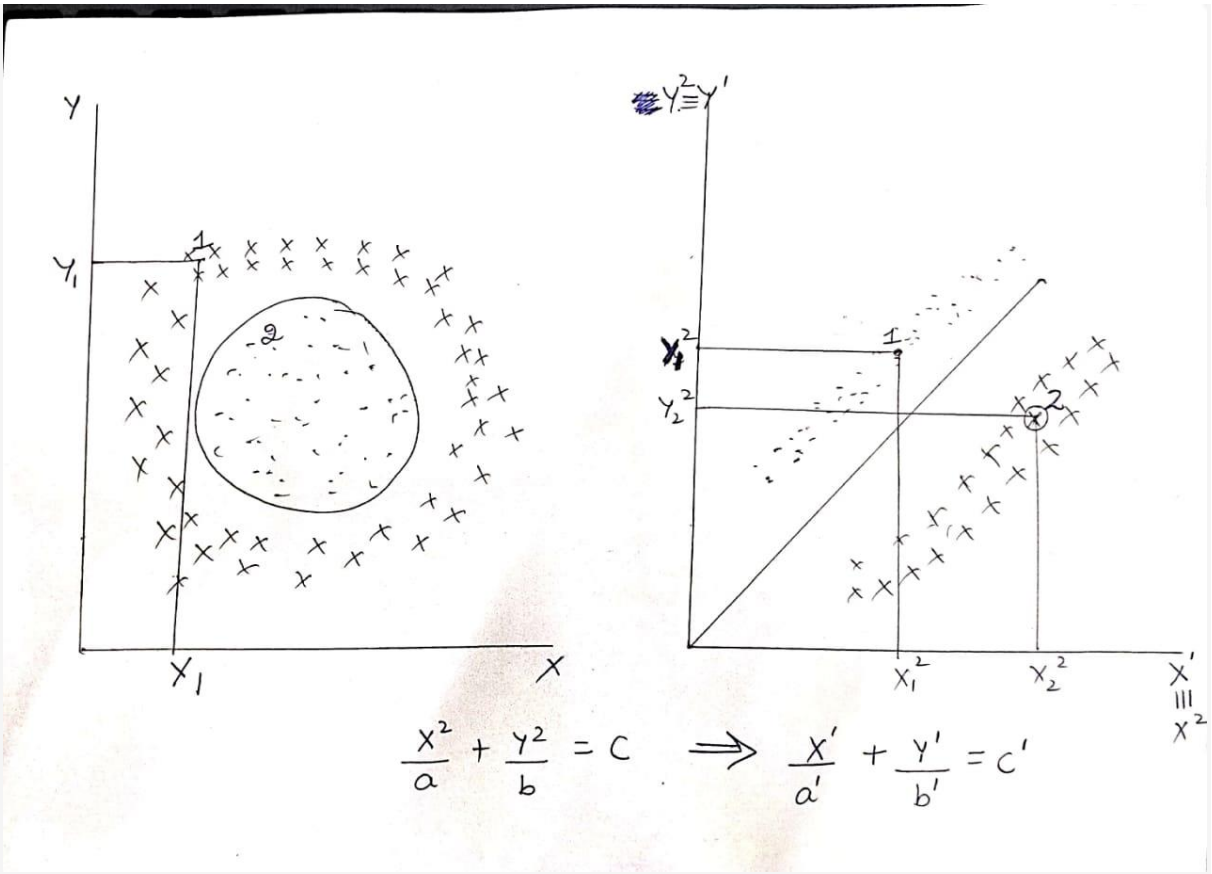
So we will use kernels here, which somehow enable the model to separate nonlinear data.

We will **transform nonlinear data sets to linear** ones. We can easily see in the figure that the boundary which will separate the two datasets is a circle or an ellipse.

Now we will do the transform by plotting this into another space and taking x1 as x1 ^2 and y1 as y1^2.

We will take two points 1 and 2 in the original given figure and mark them in the new space that is in figure 2 as shown below. Now after plotting the points in the new space we can separate them with a line because here the equation of ellipse changes to straight line as x-y dimension (original space) has been changed to x^2 – y^2 (new transformed space). We did a simple mapping here by mapping every (x, y) to (x^2, y^2).

Now it has become linear so we can run any linear methods now like linear regression, logistic regression, SVM etc. Here we will do by SVM by applying maximal margin classifier or soft margin classifier.



We can do all this procedure in python, python notebook attached with this can explain the python code for this.

Question 5

What do you mean by feature transformation?

Solution 5

We can **transform nonlinear data sets to linear** ones. This makes our work easier as calculation becomes simple and evaluation becomes eye catching.

We can do so by applying certain functions to the original attributes. The original space (X, Y) is called the original **attribute space,** and the transformed space (X', Y') is called the **feature space.**

These new features may not have the same interpretation as the original features, but they may have more discriminatory power in a different space than the original space.

To convert this data set into a linearly separable one, a simple transformation into a new feature space (X', Y') can be made.

Hence, the process of transforming the original attributes into a new feature space is called '**feature transformation'**

As the number of attributes increases, there is an **exponential increase** in the number of dimensions in the transformed feature space. Suppose you have four variables in your original data set, then considering only a polynomial transformation with **degree 2**, you end up making **15 features** in the new feature space

It is the problem of pre-processing a set of features (m) to create a new feature set (n) while retaining as much information as possible

Let me take an example and explain, consider the titanic survival prediction problem from Kaggle. Out of various features provided in the dataset consider the names of each passenger. We know that children and women are given more priority than the men from this we can say that the title in the name might be a useful feature, so from the name we could extract the titles and group them into Mr, Mrs, Miss, Master. Hence we were able to create a new and useful feature out of 'Name' feature.