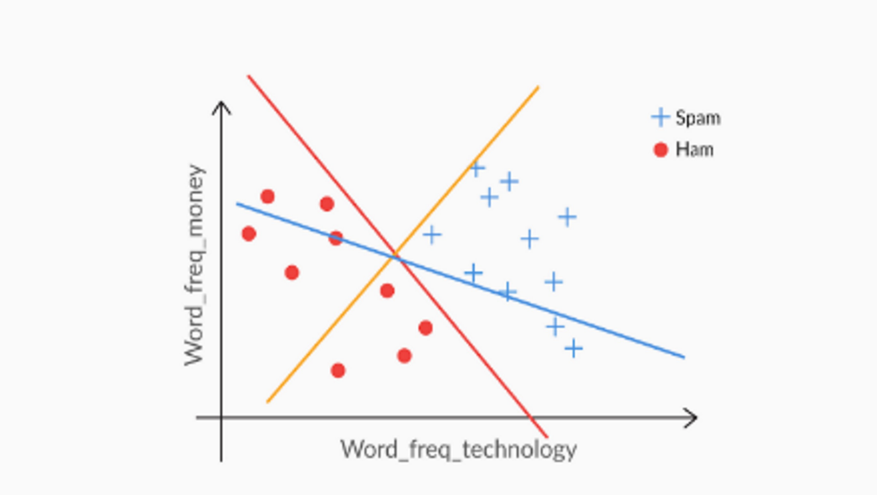
Q&A

15 February 2020 16:10

1.) Hyperplane
Let's say that a straight line, L, as shown in the plot below, is given by x2=w1x1+w0, and divides the points belonging to two classes, C1 and C2, in a 2D space. If the points (a,b) and (p,q) belong to C1 and C2 respectively, then:

(b−w1a−w0)∗(q−w1p−w0)<0
Feedback :
Since the two points belong to different classes, the terms (b−w1a−w0) and (q−w1p−w0) will have different signs for instance, if you observe the red points which lies below the line(hyperplane), give the negative value of expression x2−w1x1−w0. On the other hand, blue points which lies above the line(hyperplane), give the positive value of same expression x2−w1x1−w0. Thus, if you multiply both the expression's value, you will get a negative value.

2.) Hyperplane
In the vector space shown in the plot below, if there are various hyperplanes plotted, which hyperplane do you think is the right classifier?



Red
Feedback :
Yes. Only the hyperplane in red correctly classifies both the classes.

3.) Characterisation of line
If a point lies above the line, the value of **RHS** will be greater than zero (ax+by+c > 0) and if a point

lies below the line, the value of ax+by+c will be less than zero. But what will the **RHS** of the line equation (ax+by+c) be if a point (p,q) lies on the line?
ap+bq+c = 0
Feedback :
If the equation of the line is ax + by +c = 0, and if points p and q lie on the line they will satisfy this equation and hence.RHS will be 0

Consider a data set with two independent variables, say **X1** and **X2,** and one dependent binary variable, **Y**, with two classes, +1 or -1. The data set is plotted in the figure below.

Observations	X1	X2	Y
1	3	4	1
2	2	2	1
3	4	4	1
4	1	4	1
5	2	1	-1
6	4	3	-1
7	4	1	-1

Table-1: ***Data Set***

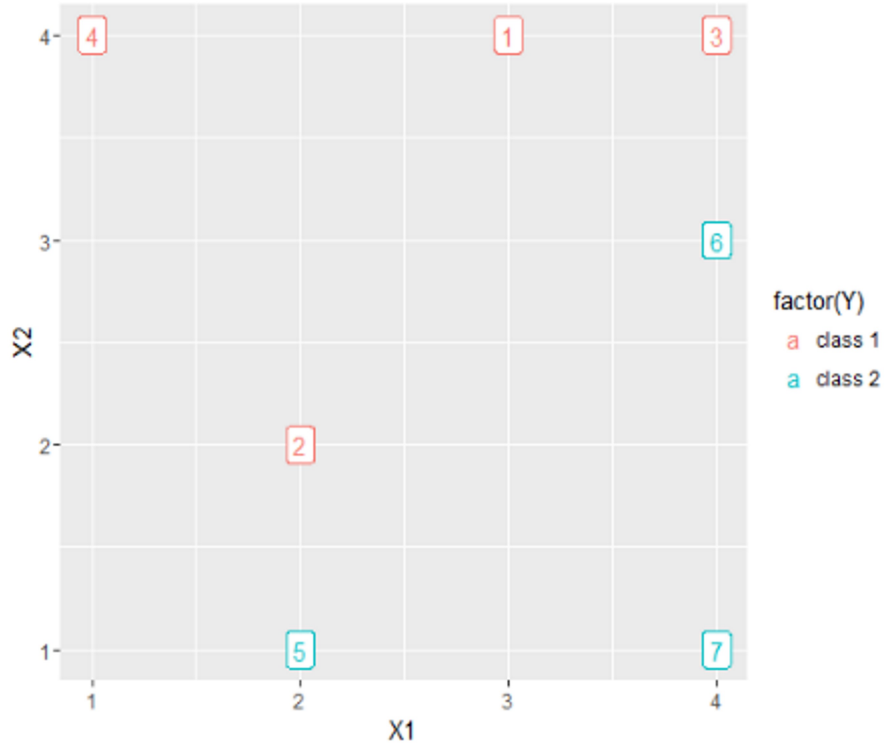


Figure-1*: Data Distribution*

4.)Hyperplane
Based on the information above, can the data set be separated by a **hyperplane,** i.e. a straight line?
Yes. It is possible to divide the classes by a straight line
Feedback :
As shown in the plot above, points 1, 2, 3 and 4 can be separated from 5, 6 and 7 by the hyperplane or line.

5.)Hyperplane

If the equation of the **optimal hyperplane** is represented by X2=W1.X1+W0, then (Hint: W1 is the slope of the line and W0 is the intercept, i.e. the point where it cuts the X2 axis) imagine the hyperplane and say whether the slope and intercept will be positive or negative.

W1 > 0, W0 < 0

Feedback :*'W1' is the slope of the line and 'W0' is the intercept (the point where it cuts the X2 axis). The slope will be positive since the hyperplane or line is tilted towards the right. The constant 'W0' is where it cuts the X2 axis, which is a point below X2 = 0.*

6.)Hyperplane

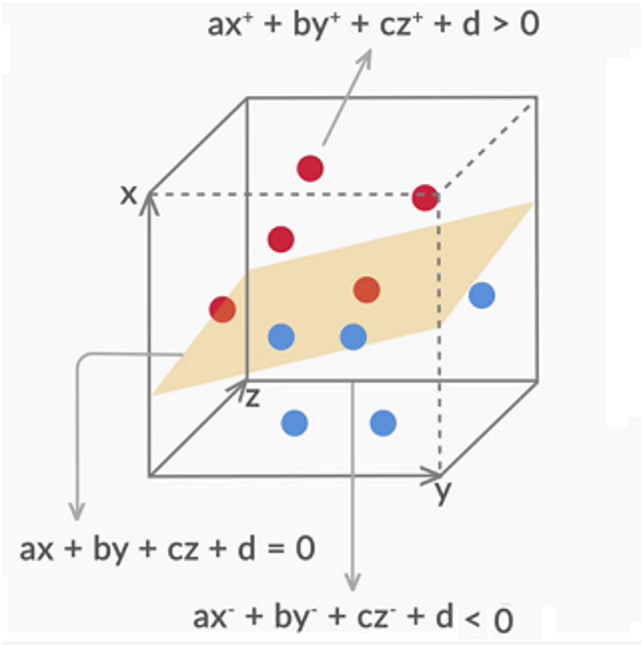
The Y labels for classes 1 and 2 are 1 and -1 respectively. If the optimal separating hyperplane is represented byX2=W1∗X1+W0, then the classification rule to classify a general point (p, q) is given by:

(W1.p−q+W0).Y>=0

Feedback :*The equation of the hyperplane is given as W1X1−X2+W0=0.If any new point lies below the line, the value of Y will be -1, and the value of the expression W1X1−X2+W0 is less than zero; this means the product will be positive. Similarly, if the point lies above the line, the value of Y will be +1, and the expression W1X1 - X2 + W0 is greater than zero. Thus, the product is positive. Also, in the case of the point on the line, the value of the expression W1X1−X2+W0 is equal to zero. Thus, the expression (W1.p−q+W0).Y gives a positive value or zero.*

7.)3D Hyperplane

What will the dimension of a hyperplane in a 3D space be?



2

Feedback :

The dimension of the hyperplane is 2. It can be calculated as [number of features - 1]. If you look at the 3D plot as well, you will see that the data can be easily separated by a plane.

8.) n-Dimensional Hyperplane

In an **n-dimensiona**l setting, with n features and p data points, what will the equation of the hyperplane be?

W0+W1x1+W2x2+...+Wn−1xn−1+Wnxn=0

Feedback :*If there are n features, the equation of the hyperplane will be the additive sum of xn features multiplied by Wn coefficients. The number of data points does not affect the hyperplane equation. For example, with two features, say x1 and x2 , the equation of the hyperplane will be a straight line — i.e. W0+W1x1+W2x2=0, where W0 is the intercept.*

9.)5-Dimensional Hyperplane

In a 5-dimensional setting, if the point **X = (X1, X2, X3, X4, X5)** lies exactly on the hyperplane, then
W0 + W1X1 + W2X2 + W3X3 + W4X4 + W5X5 = 0
Feedback :*The equation of the hyperplane is given by W0 + W1X1 + W2X2 + . . . + WnXn in an n-dimensional space. If a point is exactly on the hyperplane, it will be equal to zero.*

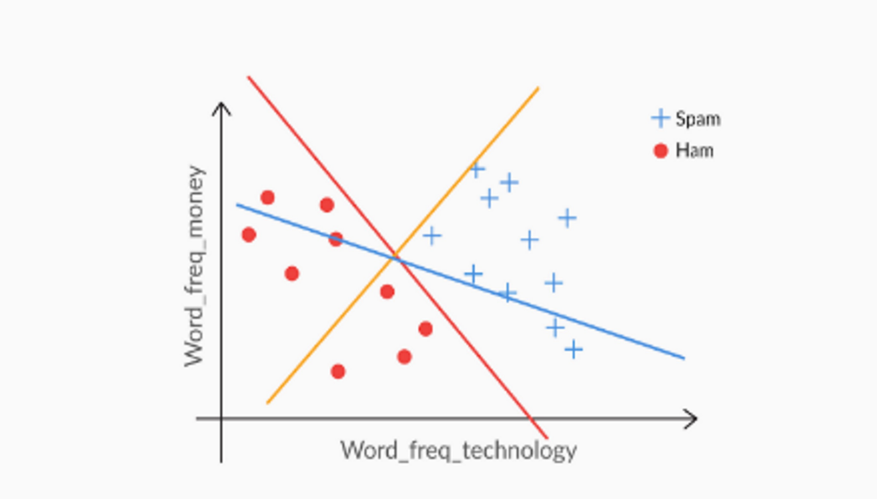
10.) Maximal Margin Classifier
How many separating hyperplanes are possible in the figure below?
Infinite
Feedback :
As you can see in the plot above, the two classes, spam and ham, that can be separated by many possible lines.

11.) Maximal Margin Classifier
The Maximal Margin Classifier ensures a margin of safety that the normal classifier (hyperplane) doesn't. But what is the advantage of this?
The model becomes less biased
Training errors are reduced
All of the above
Feedback :*The Maximal Margin Classifier divides the data set in such a way that it is equidistant from both the classes. Thus, it maintains an equal distance from both classes, making the model less biased to the training data. Also, training errors are reduced.*

12.) Maximal Margin Classifier Formulation
While defining the maximal margin classifier formulation, we are trying to:
Find the weights corresponding to the hyperplane having the maximum possible margin
Feedback :*Yes, a hyperplane is basically defined by its weights. We are trying to find the hyperplane, or the weights, such that the margin is maximum.*

13.) Scaling the weights of a hyperplane
Which of the following hyperplanes in 3-D are identical? Choose all that apply.
2x + 3y + z + 6 = 0
Feedback : If you multiply or divide all the weights by the same number, the resulting equation represents the same hyperplane.
Correct
x + 1.5y + 0.5z + 3 = 0
Feedback : If you multiply or divide all the weights by the same number, the resulting equation represents the same hyperplane.
Correct
4x + 6y + 2z + 12 = 0
Feedback : If you multiply or divide all the weights by the same number, the resulting equation represents the same hyperplane.
Correct

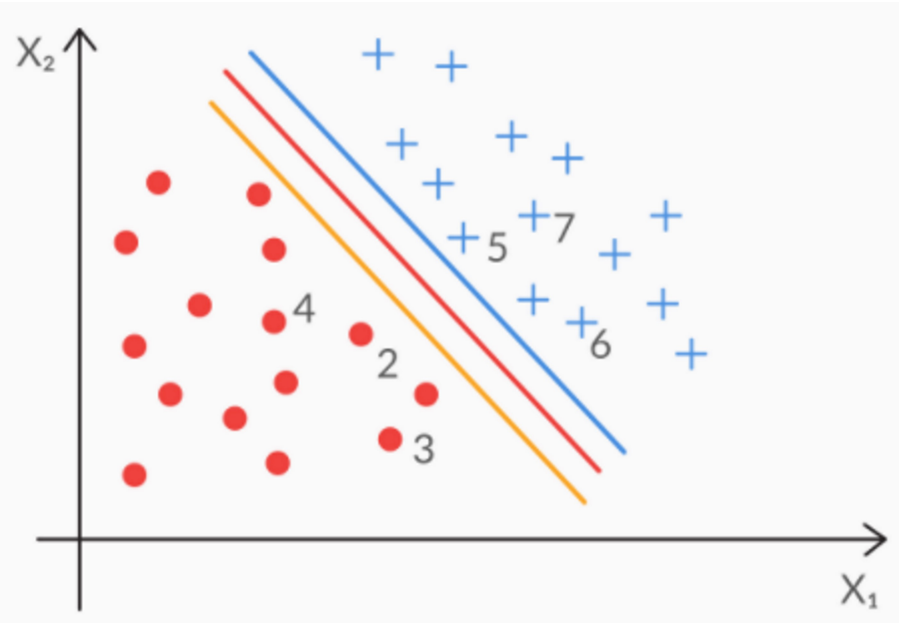
Maximal Margin Classifier



14.) In the figure below, which hyperplane makes the maximum margin hyperplane?
Red
Feedback :
Yes. The red margin is equidistant and maximal on both sides.

15.)Maximal Margin Classifier
Which are the points that are considered while constructing a margin?
Closest points to the hyperplane
Feedback : *A margin is calculated by considering only the points closest to the hyperplane.*

16.) Maximal Margin Classifier
Which points are used in constructing the maximal margin hyperplane in the figure?



Points 2 and 5
Feedback :
Since points 2 and 5 lie closer to the hyperplane than all the other points, they are the only points considered while constructing it.

17.)Maximal Margin Classifier
Among the multiple possible hyperplanes, why is one being called 'better' than the others?
The better one is a 'safe' distance away from both the classes and thus, will minimise the chances of incorrect identification.
Feedback : *The Maximal Margin Classifier is better than the others because it maintains an equal*

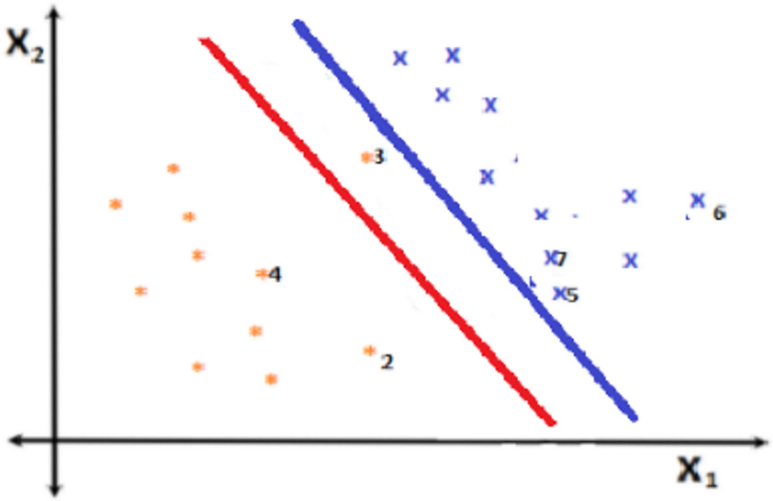
distance from both the classes; this performs better on the test set.

18.)Maximal Margin Classifier
State whether true or false. The maximal margin hyperplane is equidistant from both the classes, where 'distance' implies the distance of the closest point to the hyperplane.
True
Feedback :*Yes. The margin should be selected in such a way that it has the maximum distance from both the classes. For example, if you want to categorise spam and ham by a plane, then the plane should be drawn in such a way that it is equally set apart from both the classes, i.e. 'spam' and 'ham'.*

19.) Support Vector Classifier
Can you say that SVCs(Support Vector Classifiers) are relatively immune to **outliers**? If yes, then justify your answer.
Suggested Answer
Yes, because SVCs are formulated from the support vector points. It implies that the SVC (i.e hyperplane) will not be changed if we do not change the support vectors.

20.)Support Vector Classifier
The Maximal Margin Classifier has certain limitations and drawbacks. Due to this, we are now moving towards the Support Vector Classifier. Which of the following problems of the Maximal Margin Classifier are we trying to solve? Mark all the options that apply.
It can be extremely sensitive to individual observations. In other words, the model can drastically change if a few points are changed.
Feedback : This is shown by the addition of new training points, that changes the hyperplane drastically.
Correct
It cannot classify data that is linearly inseparable, i.e. if the classes cannot be divided by a straight line.
Feedback : The Maximal Margin Classifier does not misclassify any point. So it cannot be built on linearly inseparable data i.e intermingled data.

21.)Support Vector Classifier
Which of the following points is misclassified by the **Soft Margin Classifier** in the figure below?



Point 3
Feedback :
The points violating the margin are those that fall in the incorrect class. If you look at the image above, you will see that point 3 violates the red soft margin line.

22.) Support Vector Classifier

State whether true or false. All the data points are considered while constructing the support vector classifier.

False

Feedback : *Only the data points that lie closest to the hyperplane are useful for constructing the classifier.*

23.) Support Vector Classifier

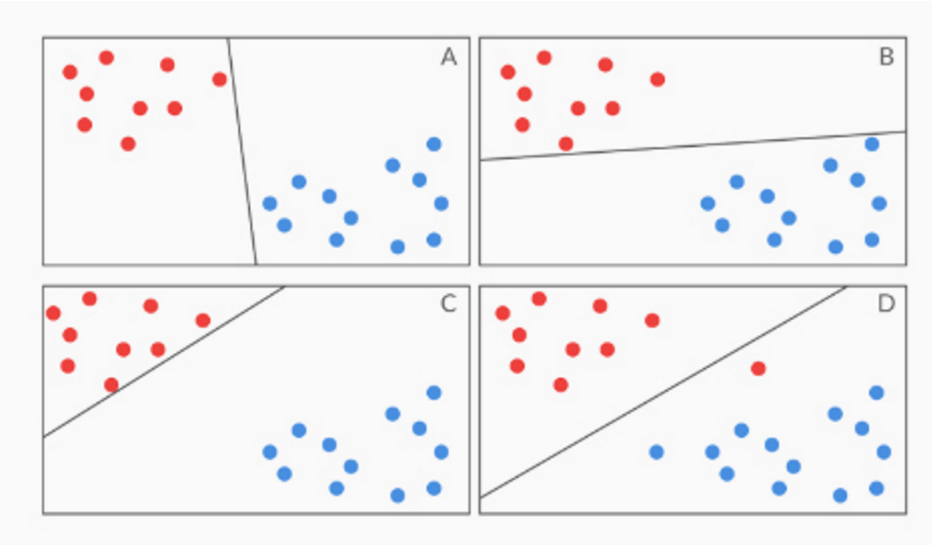
How do you differentiate between soft margin and hard margin?

A soft margin allows some points to misclassify; a hard margin ensures all points get classified correctly.

Feedback :*The soft margin is used in constructing the Soft Margin Classifier(Support Vector Classifier) which allows some points to be misclassified, whereas the hard margin ensures no points are misclassified.*

24.) Support Vector Classifier

In the figure below, which box represents the **Soft Margin Classifier**?



D

Feedback :

Yes. Plot D represents the Support Vector Classifier which allows some points to be misclassified.

25.) Cost of Misclassifications

A higher C (summation of all slack variables) leads to

High bias, low variance

Feedback :*When C is large, the slack variables can be large, i.e. the model allows a larger number of data points to be misclassified or violate the margin. In this case, the model is flexible, more generalisable, and less likely to overfit. In other words, it has a high bias. As you learnt in the model selection lectures, if you apply this model to unseen data, it can result in less variance.*

26.) Cost of Misclassifications

Assume that you have no idea what data will appear in the future. While building a model, it is a good idea to

Set C to moderate

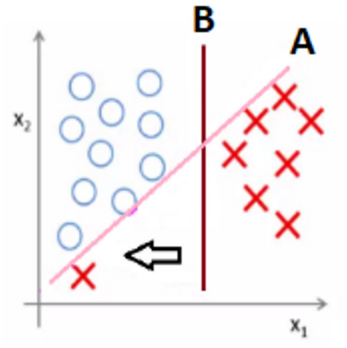
Feedback :*If the value of C is very low, then there will be no misclassifications; this may overfit the training data, and the model becomes less generalisable. Similarly, when the value of C is very high, then many points will be misclassified; this results in a bad model. So it is better to set the value of C to moderate.*

27.) Cost of Misclassifications

Which of the following steps should be taken when the SVM is overfitting?

Add more training data
Increase the value of C
All of the above
Feedback :*Yes, adding more training points and increasing value of C will reduce model overfitting problem. Because if you train the model on a large number of data, it will try to learn from the data rather fitting to the each data points. and if you increase the value of C, ultimately you are allowing your model to misclassify some data points. And, hence, the model will not be overfitted in these two cases.*

28.) Slack Variable
ε1>1 is valid for:



B
If you increase the value of epsilon(ε) from 0 to 1 the right side i.e M(1−ei) of the equation liX(Wi.Yi)> M(1−ei) will be always positive or equal to zero. It means that the data point is correctly classified by the given hyperplane. But if the value of epsilon(ε) is greater than 1, that means M(1−ei) has a negative value, in which case the data point falls on the wrong side of the hyperplane.

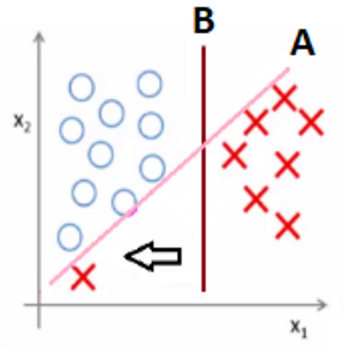
29.) Slack Variable
If the ith data point has epsilon(ei) = 0, then:
The point falls on the correct side of the margin, i.e. correctly classified and at a positive distance away from the margin
Feedback :*Yes, In this case, the value of liX(Y.Wi)>=M, which is exactly same as the maximal margin classifier. In this case, the data point falls on the correct side of the hyperplane.*

30.) Notion of Slack Variables
As you learnt, the summation of all the epsilon is equal to cost "C" So, what happens when C is large, say 100, and what happens when C is small, say 5?
Compare the two cases on the basis of:

- In which case would a higher number of points be misclassified?
- In which case is the model more likely to overfit?

Suggested Answer
If C is large, then the slack variables εi can take higher values. And you know that when slack(εi) > 1, the point is misclassified, between 0 and 1, it falls inside the margin, and when slack(εi) = 0, it is correctly classified. Thus, for higher C, more points will be allowed to get misclassified or fall inside the margin (compared to a lower C). On the other hand, a lower C implies that each slack(εi) will have to take a lower value, and thus not be allowed to stray on the other side of the margin or the hyperplane. This is a more strict condition.In other words, a lower C does not give the model freedom to misclassify even a few points, and thus the model tries to overfit the data.

31.) Support Vector Classifier
The two figures (A and B) below represent two values of C used to fit a hyperplane — one is **small** and the other is **large**. According to you, which **hyperplane** will be the best fit for the given data?



B

Feedback :

When C is large, the slack variables can be large, i.e. you allow a larger number of data points to be misclassified or to violate the margin. So you get a hyperplane where the margin is wide and misclassifications are allowed. And hence, it will work well on unseen data.

32.) Cost

Which of the hyperplanes shown below corresponds to a **higher C**?

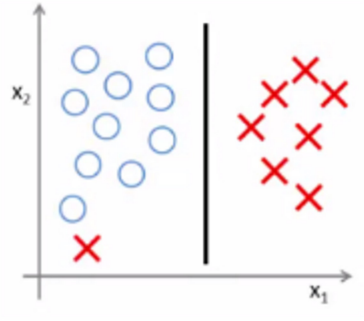


Figure 4: Hyperplane 1

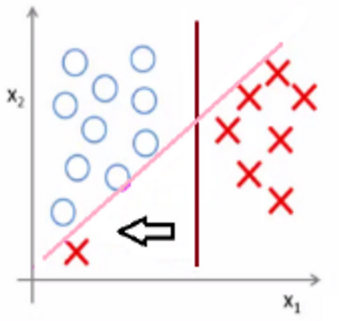


Figure 5: Hyperplane 2

Hyperplane 1

Feedback :

If C is large, the slack variables (epsilons(ϵ)) can be large, i.e. you allow a larger number of data points to be misclassified, that indicates wider margin

33.) Cost

Which of the following hyperplanes is more generalisable and stable?

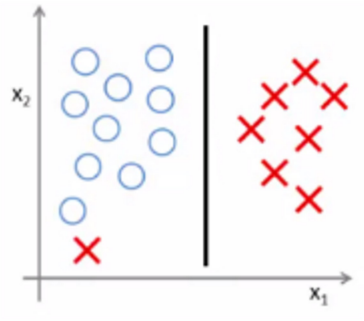


Figure 4: Hyperplane 1

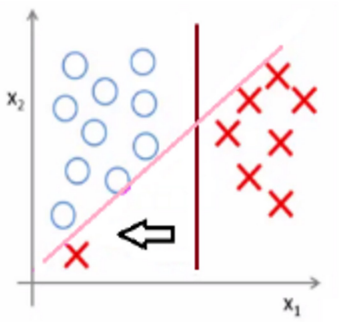


Figure 5: Hyperplane 2

Hyperplane 1

Feedback :

Yes. Even though hyperplane 1 allows a few misclassifications, it maximises the margin so the test points are classified correctly and the model is not biased.

34.) Cost
Overfitting refers to a case where a model tries to fit the available training data (often) without representing the underlying relationship between the input and output variables. Which of the two hyperplanes can be said to have overfit the training data?

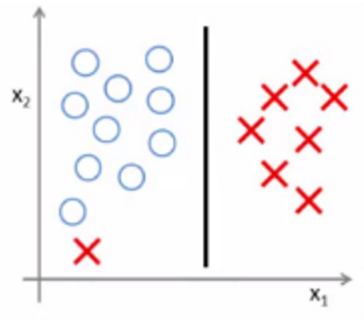


Figure 4: Hyperplane 1

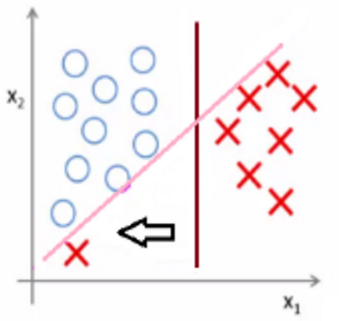


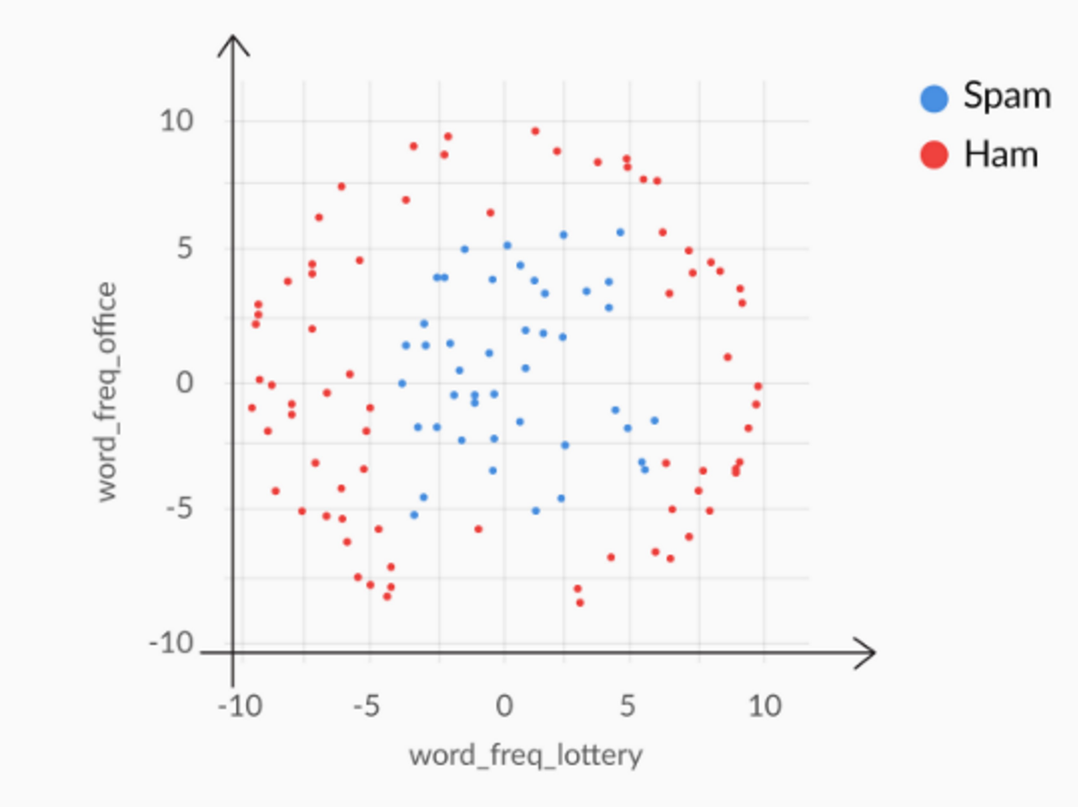
Figure 5: Hyperplane 2

Hyperplane 2
Feedback :
Hyperplane 2 is overfitting because it has classified all the training points correctly; this may lead to more chances of model variance.

35.)Data Preprocessing
We have checked that the data does not contain missing or incorrect values. Which of the following preprocessing steps is **the most crucial** before building an SVM model?
Rescaling the features (standardising or normalising) so they are all on a comparable scale
Feedback :
Yes, rescaling is an important step since some features may range from a extremely small range of numbers (say fractions) and others may be orders of magnitude higher (say 10k-100k). This may cause some features (in SVMs, they are 'dimensions') whose values are higher to dominate over others.

36.) Choice of Model Evaluation Metrics
You have two classes in your model - spam and ham. Let's say you are building the model for a client who is a logistics company and they'll use your spam classifier for maintaining the sanity of employees' inboxes.
Now, they've demanded that your model should never misclassify a genuine email as spam (imagine your appraisal email misclassified as spam), though it is okay if some spams seep into the inbox.
Assume that 'spam' is called the 'positive class', and ham is 'negative'. Which of the following metrics should your model maximise?
Specificity (True Negative Rate)
Feedback :
Yes, specificity is the fraction of negatives/hams correctly classified, a metric you ideally want to be 100%.

37.) Decision Surface
Is the data set given below linearly separable?



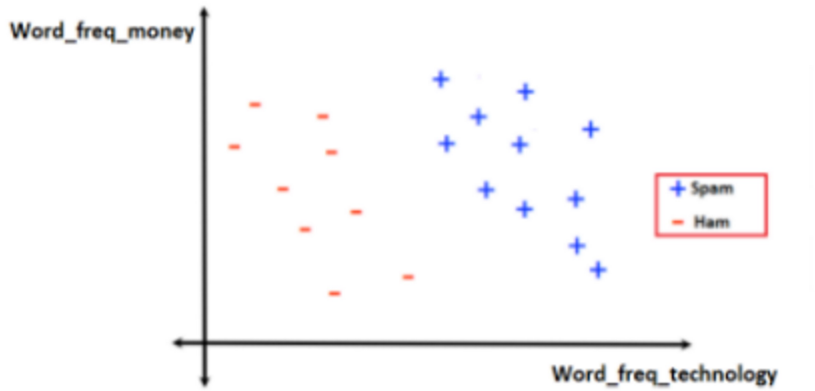
No

Feedback :

The plot is nonlinearly separable because you cannot draw a single straight line that separates the two classes distinctly.

38.) Decision Surface

Is the data set given below linearly separable?



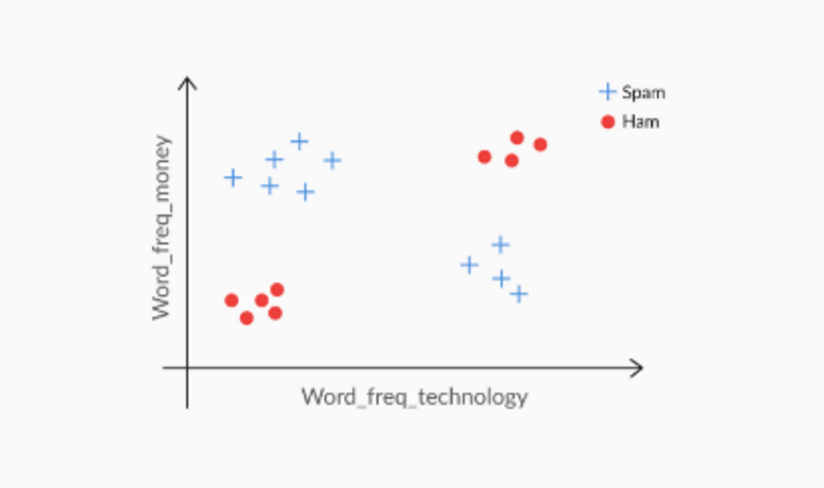
Yes

Feedback :

Yes. The plot is linearly separable because you can draw a single straight line that separates the two classes distinctly.

39.)Decision Surface

Is the data set given below linearly separable?



No
Feedback :
The plot is nonlinearly separable because you cannot draw a single straight line that separates the two classes distinctly.

40.) Mapping Nonlinear to Linear
Run the following code in Python. You can see that the relation between x and y is nonlinear by plotting them. Which value of M should be substituted in the equation p and q to make x and y linear?
(You can try a hit-and-trial from the options given below.)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
x=np.random.uniform(low=10, high=35, size=(200,)) + np.random.normal(loc=2, scale=0.5, size=200)
y = np.sqrt(400 - (x-20)**2)+ np.random.normal(loc=20, scale=0.5, size=200)
plt.scatter(x,y,color=['green'])
plt.figure(figsize=(20,20))
p = (x-M)**2
q = (y-M)**2
plt.scatter(p,q,color=['green'])
```

20
Feedback :
If you change the value of M from 5 to 20, you will see that p = (x-20)^2 and q = (y-20)^2. This makes the data linearly distributed.

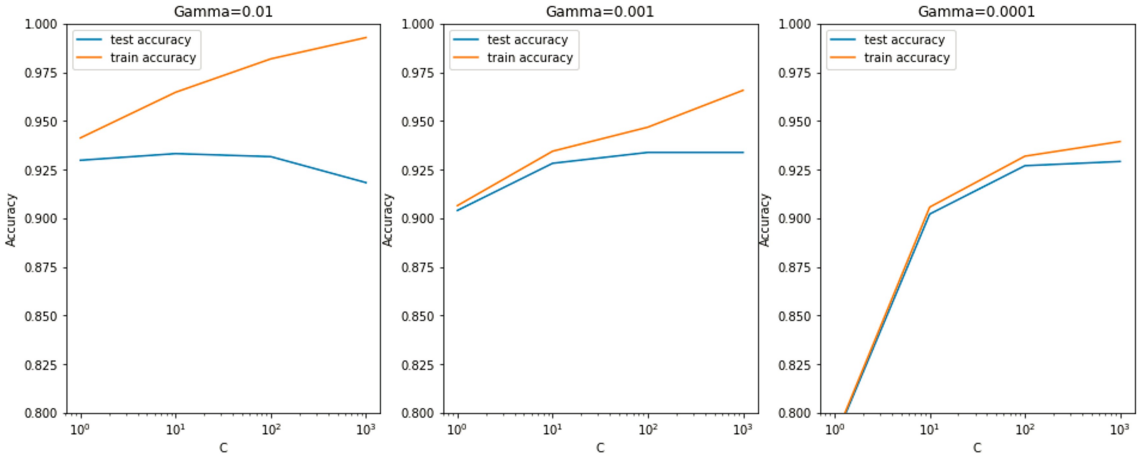
41.) Mapping Nonlinear to Linear
Why is the transformation of attributes required?
To make the relationship between the variables linear
Feedback :*SVM works well if the data is linearly separable. Hence, it is advantageous to make the data linearly separable.*

42.) Feature Transformation
Given a data point (x, y) = (2, 3), what will be the transformed data point in the feature space (x,y,xy,x2,y2,1)?
(2, 3, 6, 4, 9, 1)
Feedback :*At x=2 and y=3, the value of (x,y,xy,x2,y2,1) is equal to (2, 3, 6, 4, 9, 1)*

43.) Feature Transformation

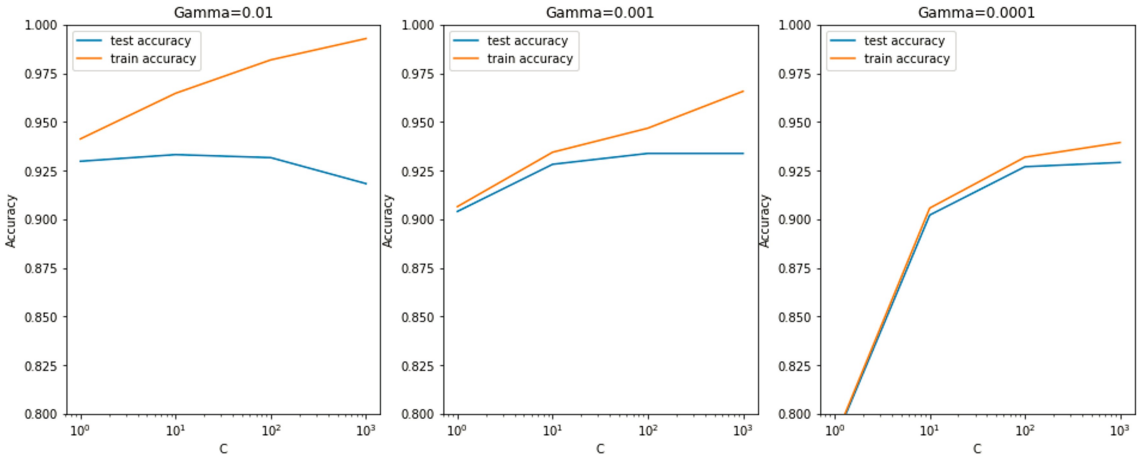
The transformation from the 2-D attribute space resulted in a linearly separable feature space which is 6-D, i.e. the number of dimensions have increased. In theory, this looks fine. But what practical problems do you think this can cause?
Suggested Answer
A higher number of features will increase the computational cost

44.)Hyperparameters C and Gamma
The image below shows how the training and test accuracies vary with C and gamma. Answer whether true or false based on the image - the model tends to overfit at higher values of gamma (keeping C constant).



True
Feedback :
Yes, at higher values of gamma, the performance (accuracy) on training data is much better than that on test data - a clear sign of overfitting.

45.) Hyperparameters C and Gamma
The image below shows how the training and test accuracies vary with C and gamma. Answer whether true or false - increasingly complex (non-linear) models result in a higher training accuracy, though the test accuracy does not increase significantly with increasing model complexity.



True
Feedback :
Yes, this is correct - as you increase gamma (i.e. make a more complex model), the training accuracy goes up, though the test accuracy does not improve with an equal amount.

Note: for Questions 46,47,48 Please note that the hyperparameter 'C', used in this comprehension refers to the 'C' used in the SVC() function, i.e. higher the C, more complex the model.

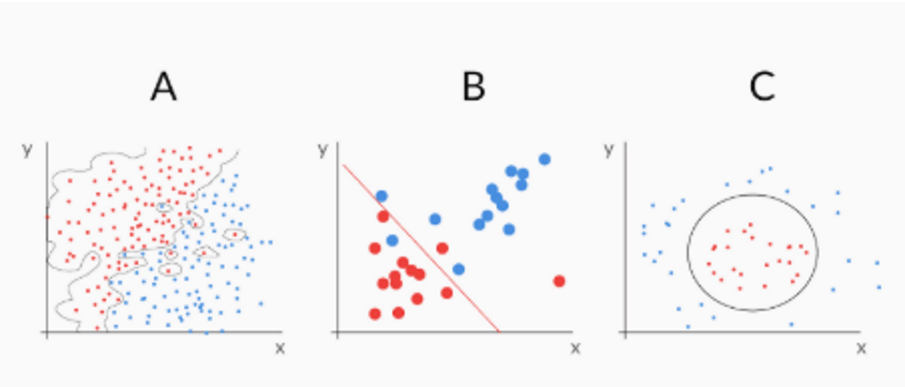
46.) Hyperparameter Gamma
What happens if the value of gamma is very high?
Nonlinearity increases in the decision surface
Feedback :
Gamma is used to increase the nonlinearity in the decision surface.

47.) Hyperparameters
Overfitting can be controlled by
C
Gamma
Kernels such as RBF, linear kernels, etc.
All of the above
Feedback :
'C', gamma, and the types of kernels, all have an effect on constructing the decision boundary.

48.) Hyperparameters
Which of the following cases are likely to overfit the model? More than one options may be correct. Also, the C here implies the C in the SVC() implementation in sklearn.
High value of C
Feedback :
A higher value of 'C' (tuning parameter i.e used in SVC() in Python lab) will not allow any points to be misclassified. The SVM model will be overfitting when there is no misclassification.
Correct
High value of gamma
Feedback :
A higher value of gamma will add more nonlinearity to the decision surface. The SVM model will be overfitting when no points are misclassified, and the nonlinearity is highly introduced than required.

49.) Kernels
Using the **interactive graphic**, create two models using the RBF kernel, model A and model B. Suppose that the model-A takes the maximum value of both hyper-parameters i.e maximum value of 'C' and 'gamma', whereas, model-B takes the minimum value of hyper-parameters.
Which model overfits the dataset? Note that the two classes are shown in red and black colours respectively.
Model-A
Feedback :
If you set the value of 'C' and gamma to maximum, the model tries to fit all the training data thus the model will be more complex and it overfits the data.

50.) Kernel Functions
Which of the following SVM decision surfaces seems overfitted?

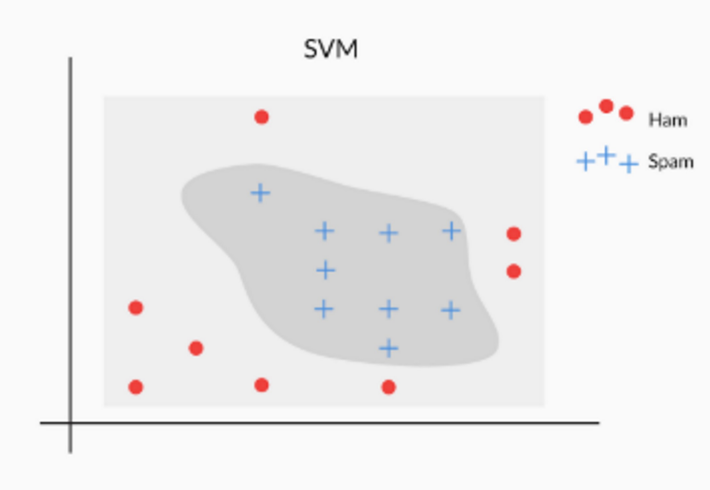


A
Feedback :

An overfitting model covers all the data points and becomes more biased towards the training data.

51.) Kernel Functions

Which type of kernel is shown in the figure below?



RBF

Feedback :

Yes. An RBF kernel adds a good amount of nonlinearity while classifying the data.

52.) RBF Kernel

What if you tune the hyperparameters for RBF kernels using cross-validation? Does the accuracy change or not? If yes, then what is the value of 'c' and gamma?

Yes, it will change. The best test score is 95.18 corresponding to C:1000 and gamma : 0.01

53.)