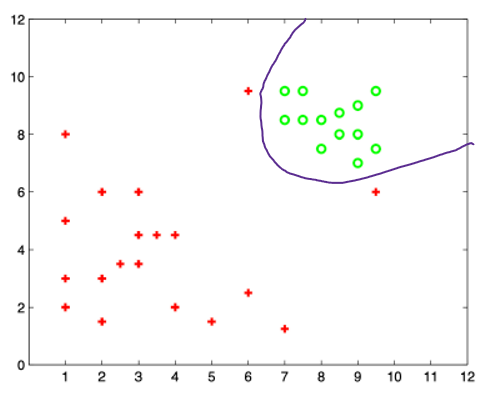
**Question 3: Theoretical Questions**

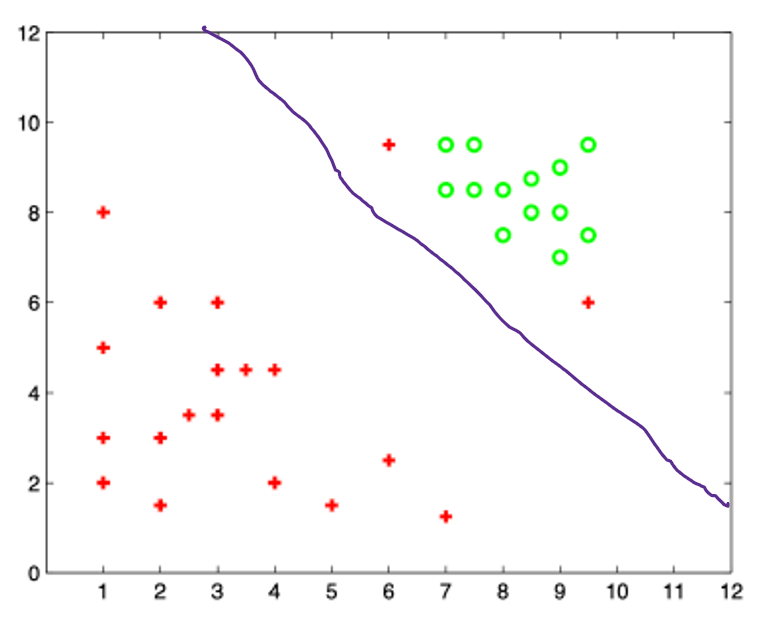
1. We have a dataset with k features and N samples and N>> k then:
2. In the case of discrete features (2 Category only) , the highest no of leaves would be 2k as the each feature can be used only one time from each path when we are taking root to the leaf thus making the highest no of features to be 2k.
3. In the event of continuous features, we can use the sample values many times. So the maximum no of leaf nodes could be equal to number of samples which in this case would be N samples.
4. Email classification
5. This plan may not work because the probability is only of the total emails containing the keywords. This is because the dataset considered is only for total emails containing specific set of keywords. The probability formula should look something like this :

Here the features are the words in the email and so the probability should be calculated of the each distinct word in the emails. After training the data set buy counting number of times each word appears in the email we can build the classifier and then test it on the email.

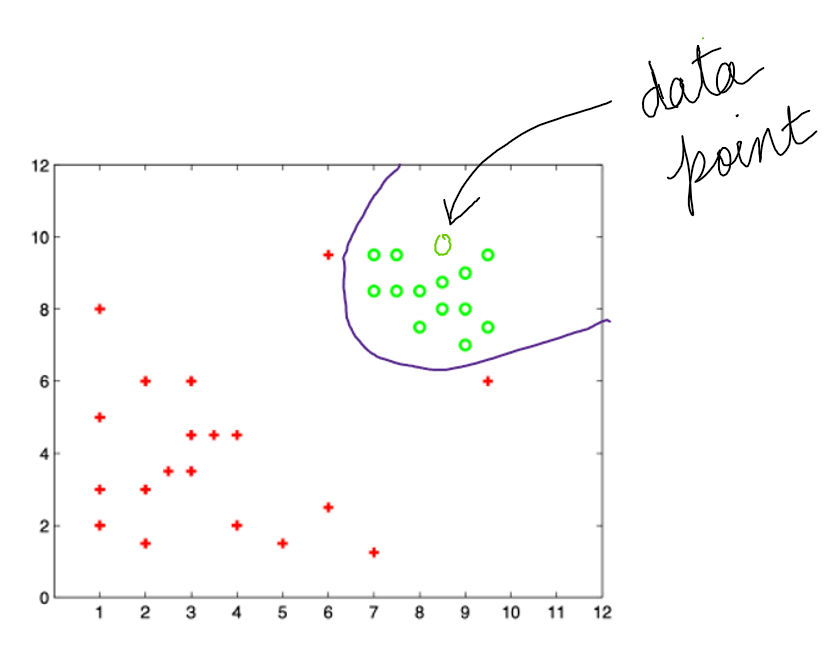
1. The dataset should consider all the emails while training the dataset and not only with those specific keywords. There should be a text column that contains the entire message in the email and another label column that contains the outcomes of emails as spam or legitimate while training the dataset.
2. Using nave Bayer’s the equations to get probability are :
3. It caters to our plan as it would correctly train and then classify the dataset by finding the correct no of occurrences of the spam words in the email.
4. SVM with Quadratic kernel : As the Kernel is Quadratic lines will not be straight
5. High C that is approaching infinity means that the cost function is very high that means that the decision boundary will try to reduce the misclassification as much as possible that is it will try to perfectly separate the data as shown as below if is possible to do so.



1. C is close to 0 means that the cost function is very low so cost for misclassifying (or in other words penalty) is low that means that decision boundary will misclassify few points as shown below :



1. As C close to 0 sort of marginalize the error between the major region of points, we think that would work best to classify, that is assuming if we do not rely on one specific data point a lot.
2. For very larges of C the datapoint made below will certainly not change the decision boundary made. This is because it is within the original correctly classified region.



1. Now as question demands to change the decision boundary then, a point should be drawn as shown below as it would certainly force the decision boundary to change for C approaching infinity. Reason : As this point is added in the region which was original classified as something else so adding an opposite point that is green in that region will force the boundary to shift in order to correctly classify this new data point.

