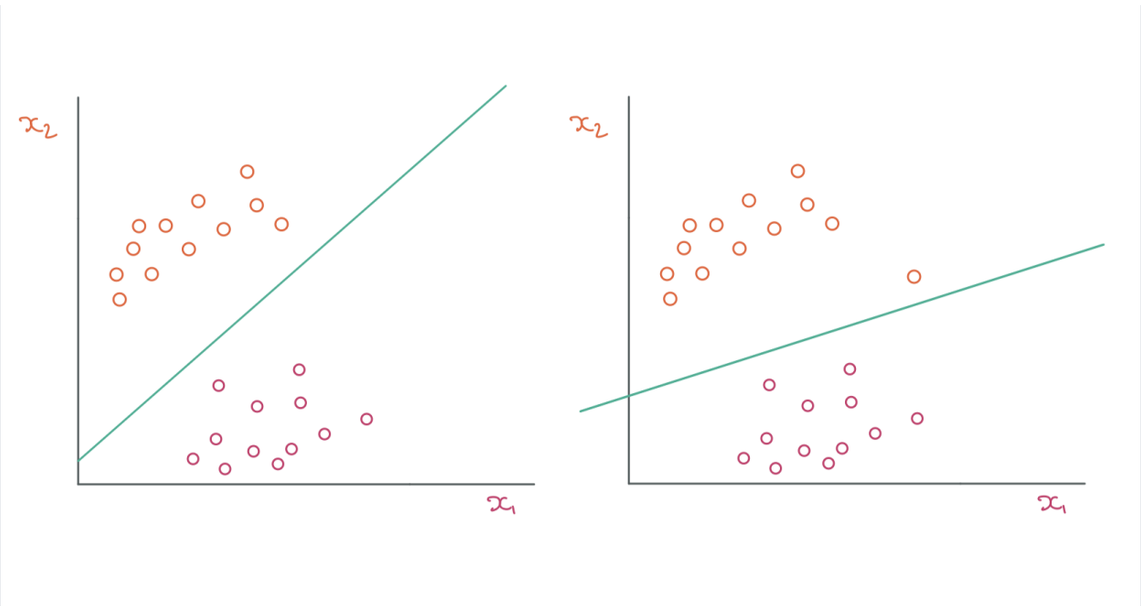
Support Vector Machines Part 2

In Support Vector Machines Part 1, we looked at the concept of maximal margin hyperplane to divide the training observations into separate classes. In this part, we will look at it's limitations and the idea of Support Vector Classifiers.

The problem with maximal margin hyperplane is that it can be very sensitive to individual training data points. Addition of one more training observation can shift the hyperplane too much. Lets look at the following figures.



In the left figure, we have the green hyperplane separating the data into two classes, now adding one more orange data point at a location shown in the left moves the green classifier to the new position which does not look very satisfactory for the following reasons

1. It's margin is small, so we have less confidence in its classification.
2. It seems to be following noise rather than the signal aka **Overfitting**.

The problem with such a classifier is that it might do very well on training data but will rate poorly on unseen test data. We would prefer a hyperplane that has a good training and testing accuracy over one that is excellent training but poor testing accuracy.

So we are fine with misclassifying some of the training observations if it performs better on the remaining ones. This is exactly what a **Support Vector Classifier**

does.

Support vector classifier enjoys a freedom to misclassify some observations by allowing them to cross the margin but also the hyperplane. The amount of freedom to do this is a hyperparameter that can be tuned.

Just like the maximal margin hyperplane, the Support vector classifier decides to classify the observation data based on which side of the hyperplane it lies. The selected hyperplane may misclassify some of the data. We can find the hyperplane by solving for the following four constraints

> maximize M

> $\sum_j^p \beta_j^2 = 1$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ...\beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Optimization problem

Where M is the width of our margin and i stands for ith observation. Let's talk about these equations next.

The equation $\sum_j^p \beta_j^2 = 1$ along with $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ...\beta_p x_{ip}) \geq M$

ensures that the perpendicular distance of the ith data point is greater than or equal to the Margin. Now when we multiply M with $(1 - \epsilon_i)$ we are allowing some freedom in where this data point can lie with respect to the hyperplane.

For $1 \geq \epsilon_i \geq 0$ the point lies between the Margin and the hyperplane and for $\epsilon_i > 1$, the data point lies on the wrong side of the hyperplane. And of course for $\epsilon_i = 0$, the point lies on the correct side of the margin. These ϵ_i are known as Slack variables.

Let's talk about C now. C is a tuning parameter that decides how many points and by how much can they be misclassified. A high value of C would allow more data to be misclassified. So C kind of decides our tolerance level for misclassification. If C = m, then we can have at most m data points misclassified on the other side of the hyperplane ($\epsilon = 1$)

since $\sum_{i=1}^{n} \epsilon_i \leq C$

Bias-variance trade-off

A low value of C means less tolerance for misclassification which means narrow margins which can overfit the data, hence low bias but high variance (overfitting). On the other hand, a large value of C will allow a higher number of misclassifications which means high bias but low variance. So C really controls the bias-variance trade-off.

Support vectors

One of the interesting feature of this optimization problem defined by the sets of equation above, is that the hyperplane separating the classes is affected by only those data points that either lie on the margin or on the wrong side of the margin. **Those observations that lie on the correct side of the margin will not affect the classifier.** Moving it around, as long as it does not cross the margin will not affect the position of the hyperplane.

These data points that lie either on the margin or on the wrong side of it are called s**upport vectors**. These support vectors determine the hyperplane.

If we increase the value of C, then that means we have a large number of data points either on the margin or on the wrong side of, thus large number of support vectors. Thus by controlling C, we control the number of observations that have an impact on the classifier.

References

1. An introduction to Statistical learning by Gareth James et al