Support Vector Machines

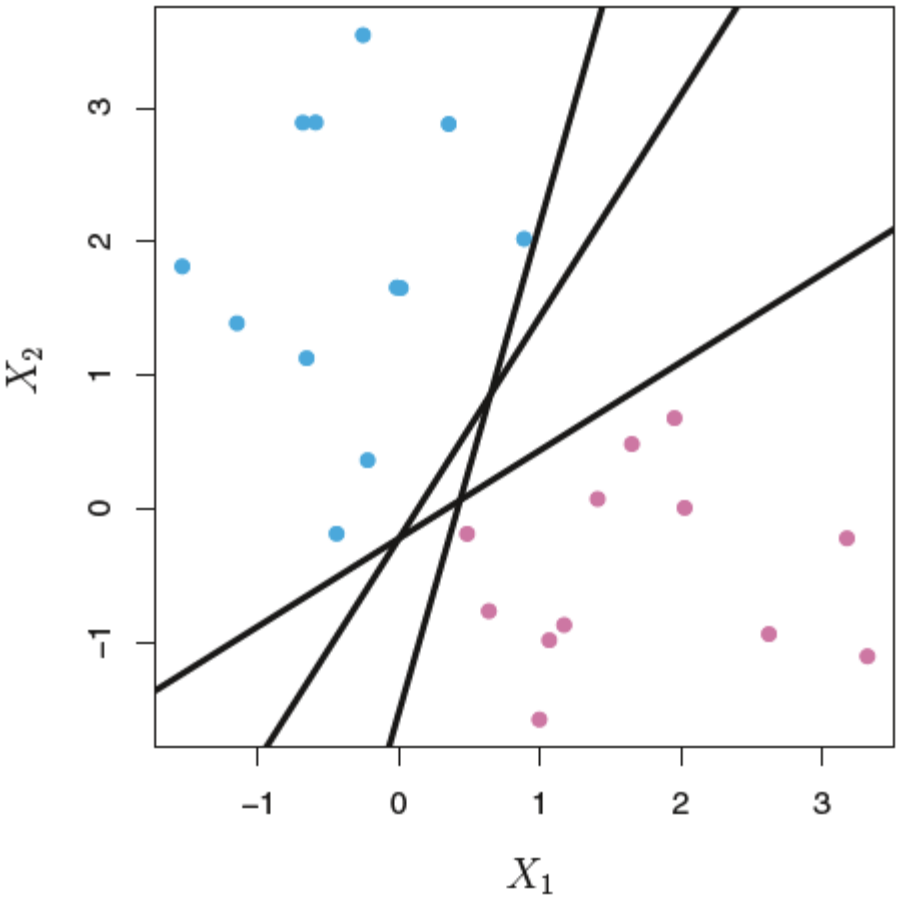
Stephen Jun Villejo

School of Statistics

Maximal Margin Classifier

Illustration

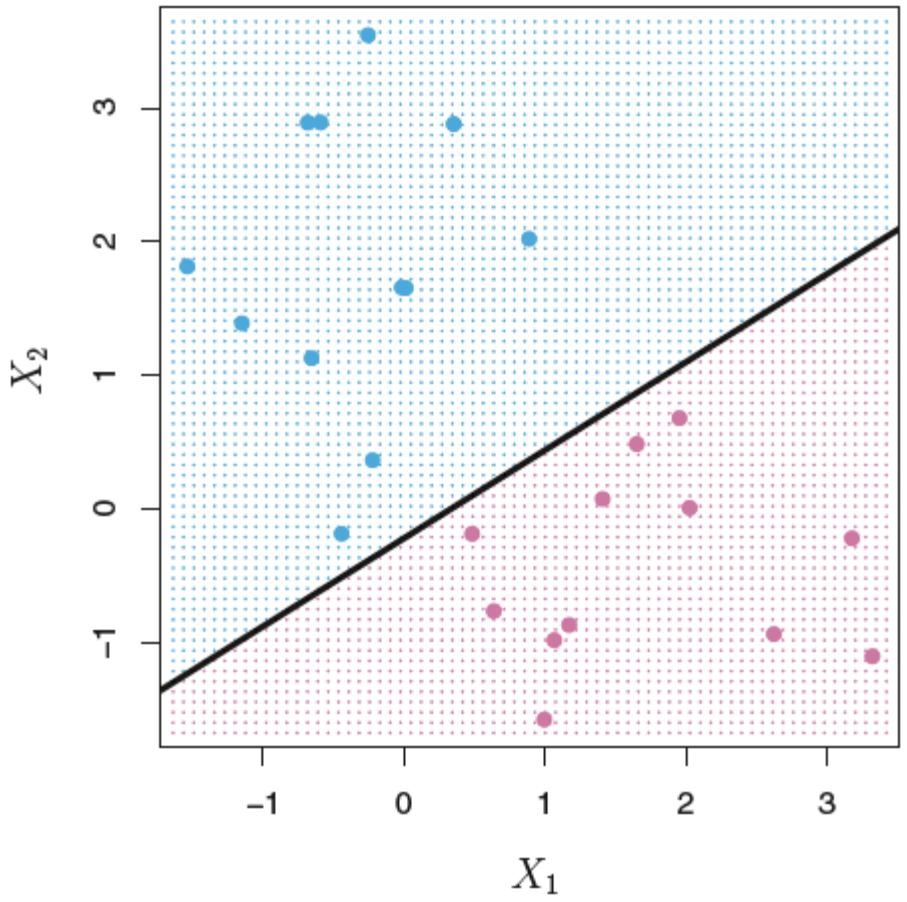
- Figure on the right shows two classes of observations (in blue and in purple). Three separating hyperplanes, out of many possible, are shown in black.



Maximal Margin Classifier

Illustration

- A separating hyperplane is shown in black. The blue and purple grid indicates the decision rule made by a classifier based on the separating hyperplane.



Maximal Margin Classifier

What is a hyperplane?

- In two dimensions (p=2), a hyperplane is a line. It is defined by the equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

- In three dimensions (p=3), a hyperplane is a plane.
- In $p > 3$ dimensions, it can be hard to visualize a hyperplane.

Maximal Margin Classifier

What is a hyperplane?

- In the p-dimensional setting we have the hyperplane

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

- When $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0$, the observation lies to one side of the hyperplane.
- When $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0$, the observation lies on another side of the hyperplane.

Maximal Margin Classifier

Classification using a Separating Hyperplane

- A hyperplane that perfectly classifies observations is called a *separating hyperplane*.
- Here, an observation is classified depending on which side of the hyperplane it is located.
- It has a property that

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} > 0 \text{ if } y_i = 1,$$

and

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} < 0 \text{ if } y_i = -1.$$

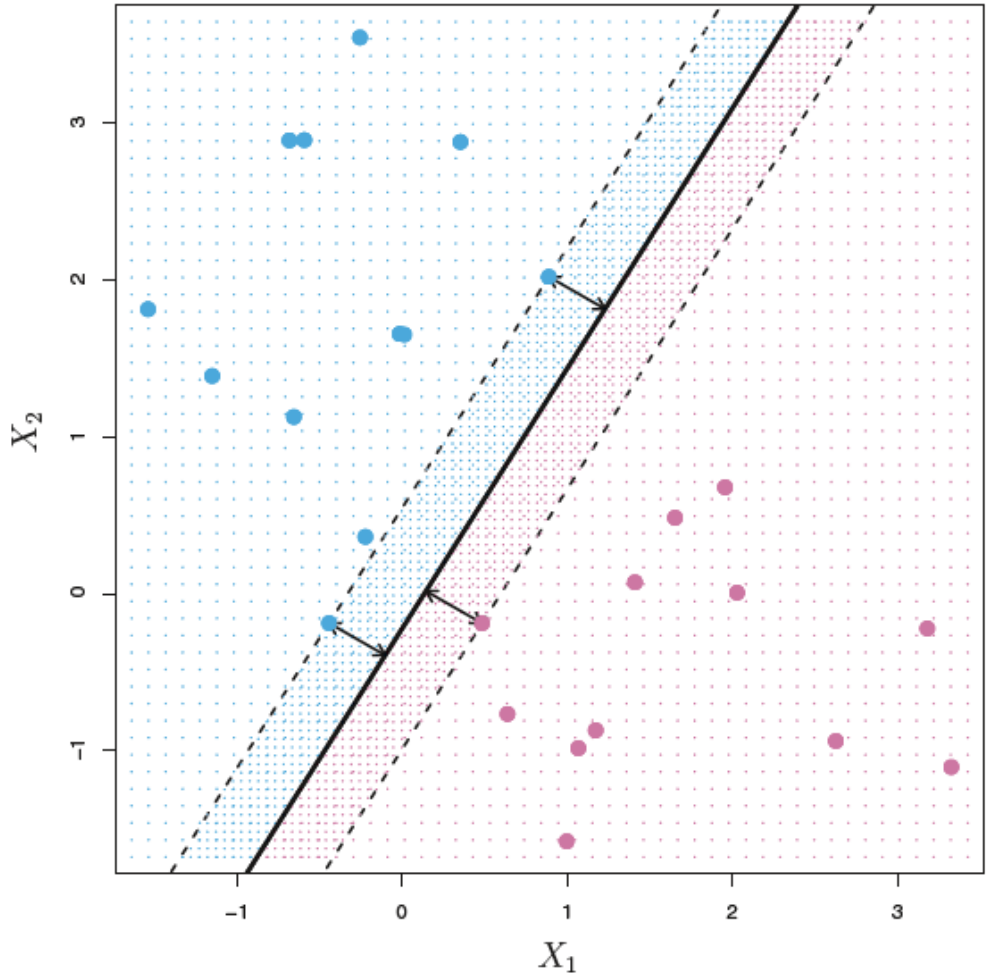
Maximal Margin Classifier

- If our data can be perfectly separated using a hyperplane, there will exist an infinite number of such hyperplanes
- The **maximal margin hyperplane** is the separating hyperplane that is farthest from the training observations.
- The smallest distance from each training observation to a hyperplane is called the margin. The maximal margin hyperplane is the separating hyperplane for which the margin is largest.

Maximal Margin Classifier

Illustration

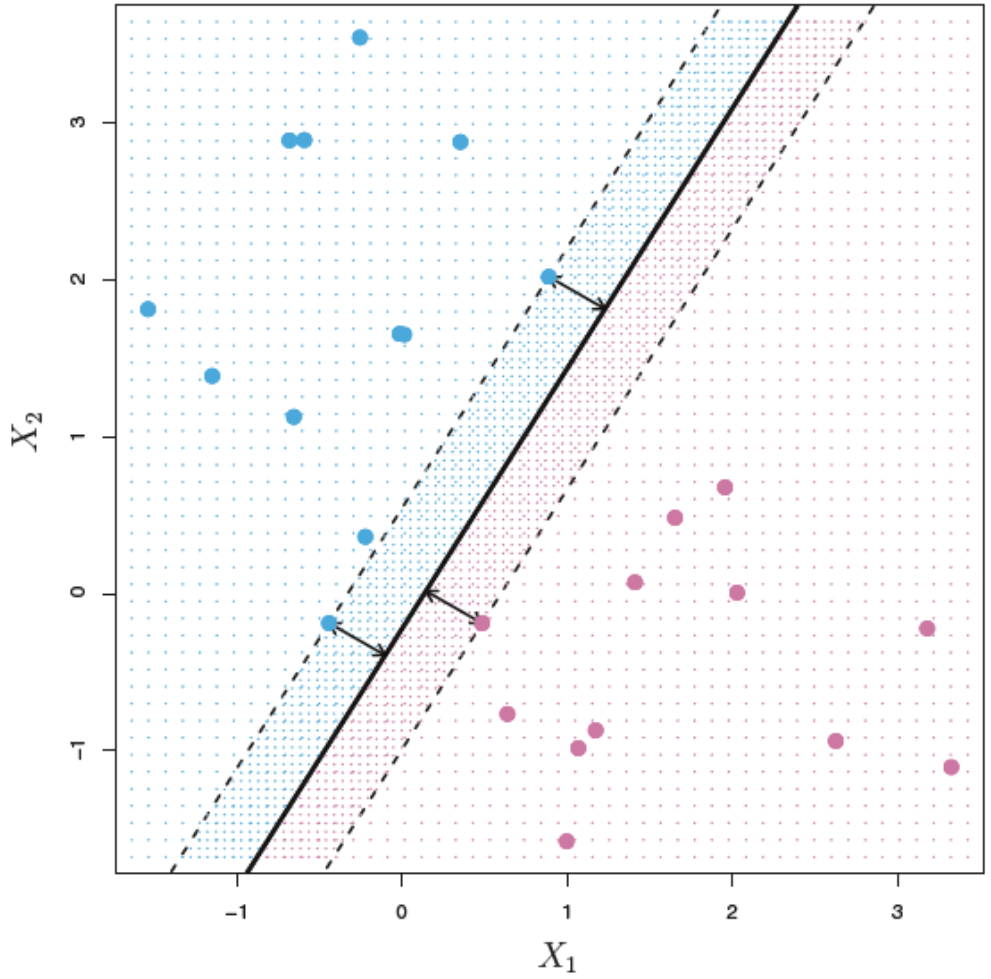
- The maximal margin hyperplane is shown as a solid line.
- The margin is the distance from the solid line to either of the dashed lines.
- The two blue points and the purple point that lie on the dashed lines are the **support vectors**.
- The purple and blue grid indicates the decision rule made by a classifier.



Maximal Margin Classifier

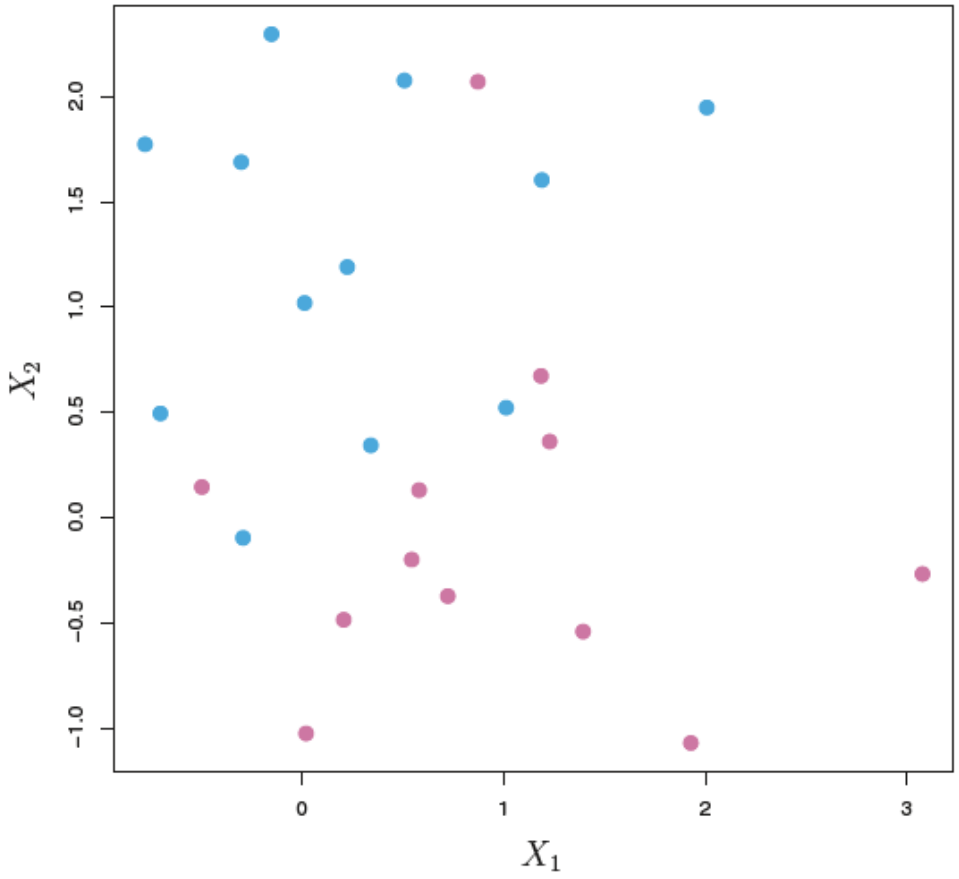
Illustration

- The maximal margine hyperplane depends only on a small subset of the observations called the **support vectors**.

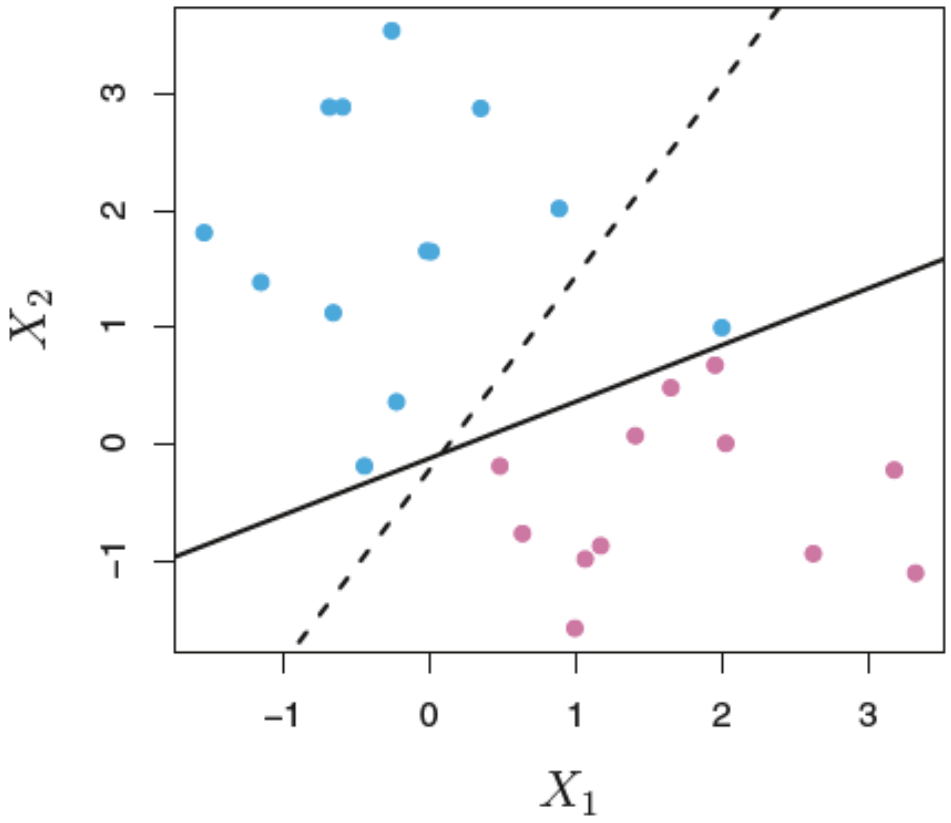


Support Vector Classifiers

- Observations belonging to two classes are not necessarily separable by a hyperplane.



Support Vector Classifiers



- Also there are instance, when a classifier based on a separating hyperplane is not desirable.
- The maximal margin hyperplane is extremely sensitive to a change in a single observation. This suggests overfitting to the data.

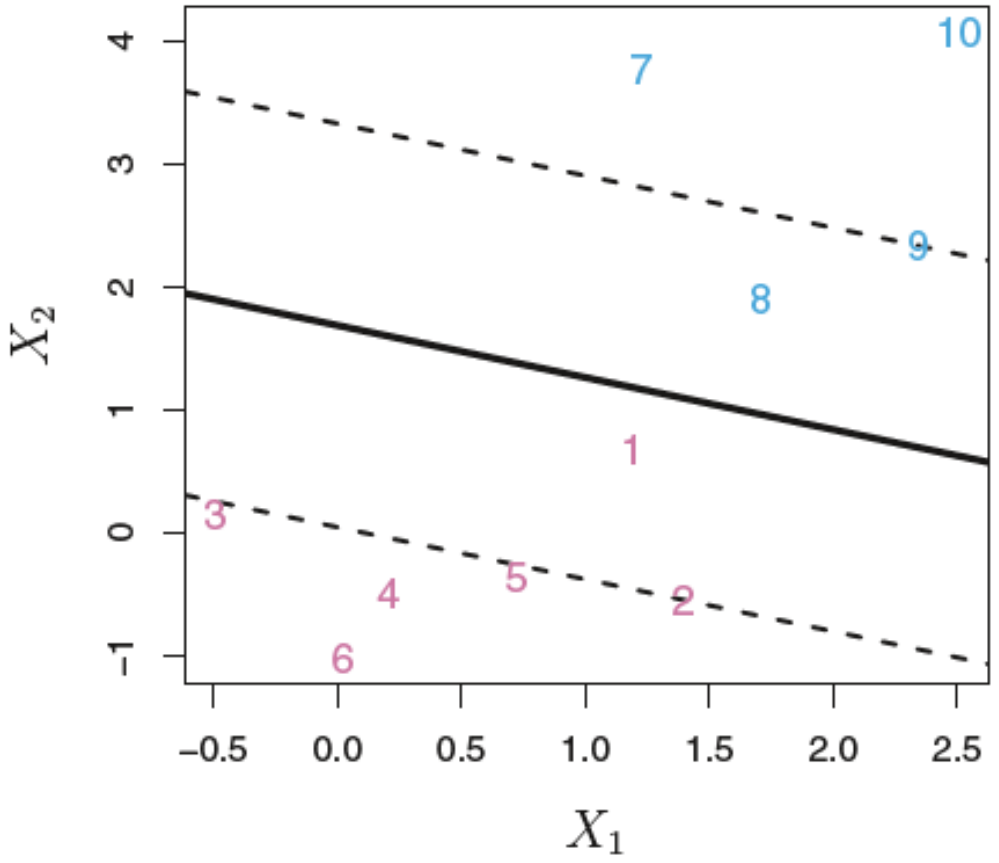
Support Vector Classifiers

- A *support vector classifier* is a classifier that does not perfectly separate the two classes in the interest of greater robustness to individual observation and better classification of most of the training observations.
- In a support vector classifier, we allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane.

Support Vector Classifiers

Illustration

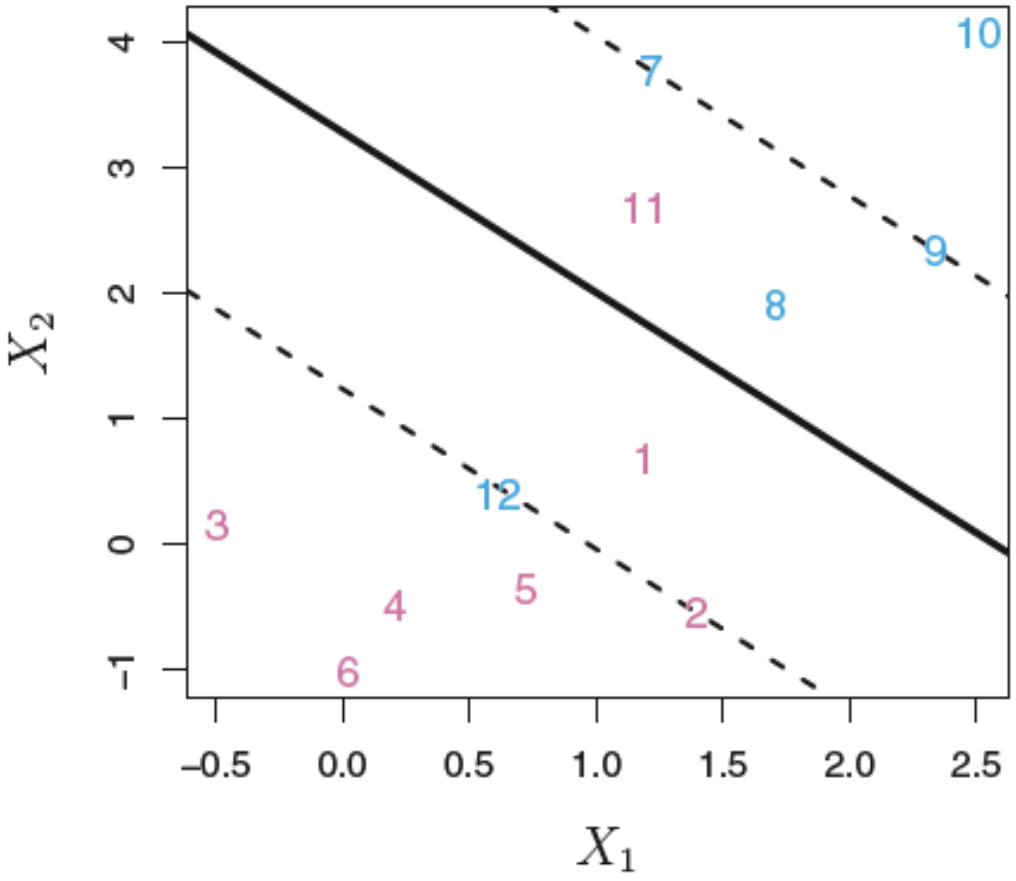
- Observations 3, 4, 5, and 6 are on the correct side of the margin.
- Observation 1 is on the wrong side of the margin.
- Observation 8 is also on the wrong side of the margin.
- No observations are on the wrong side of the hyperplane.



Support Vector Classifiers

Illustration

- Two observations are on the wrong side of the hyperplane.



Support Vector Classifiers

Tuning Parameter

- An important parameter involved in the estimation of the classifier is the tuning parameter.
- The tuning parameter, C, serves as a *budget* for the amount that the margin can be violated by the observations.
- If C=0, then there is no budget for violations to the margin, which yields the maximal margin hyperplane.
- As the budget C increase, we become more tolerant of violations to the margin.

Support Vector Classifiers

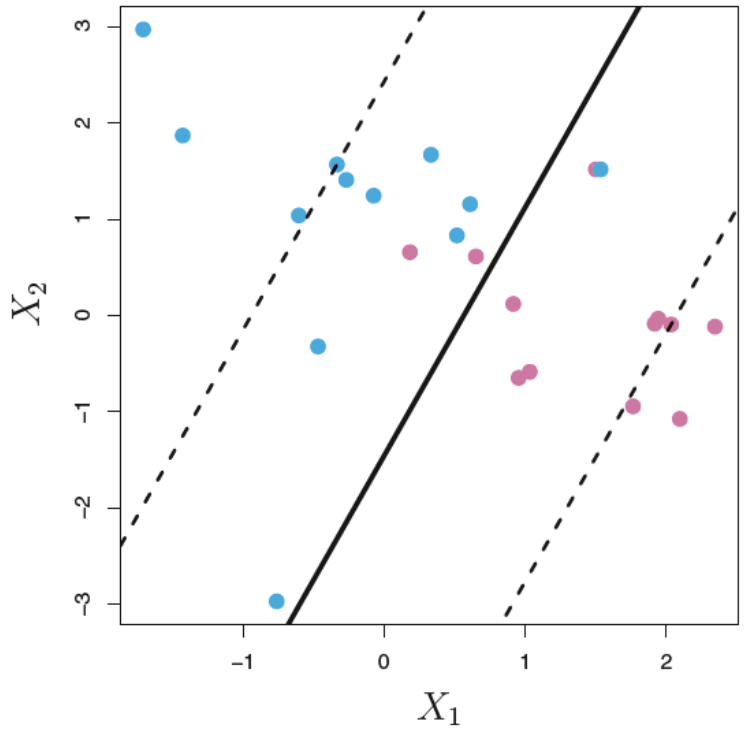
Tuning Parameter

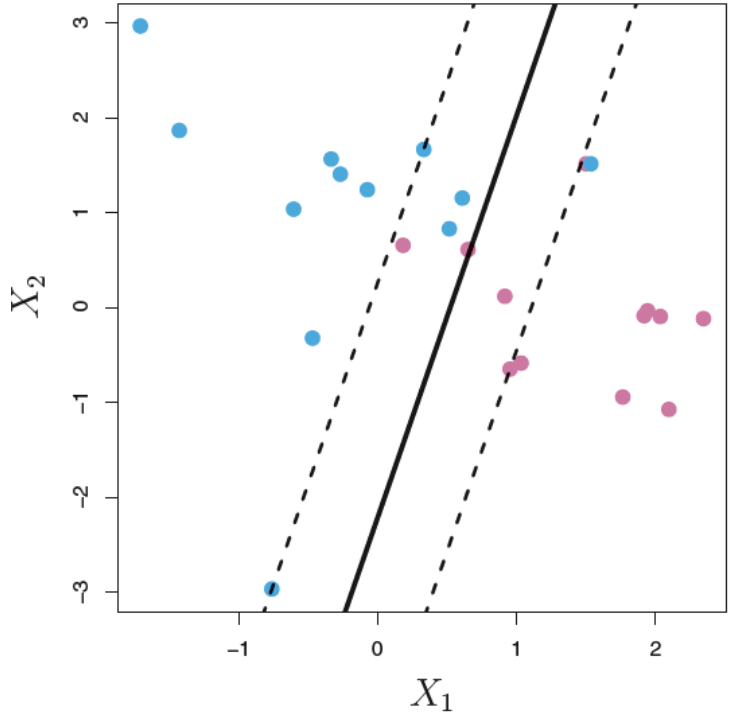
- When C is small, we seek narrow margins; this amounts to a classifier that is highly fit to the data, with low bias, but high variance.
- When C is large, that margin is wide, allowing more violations; this amounts to fitting the data less hard, allowing more bias but lower variance.

Support Vector Classifiers

Tuning Parameter

- As C decreases, the tolerance for observations being on the wrong side of the margin decreases.





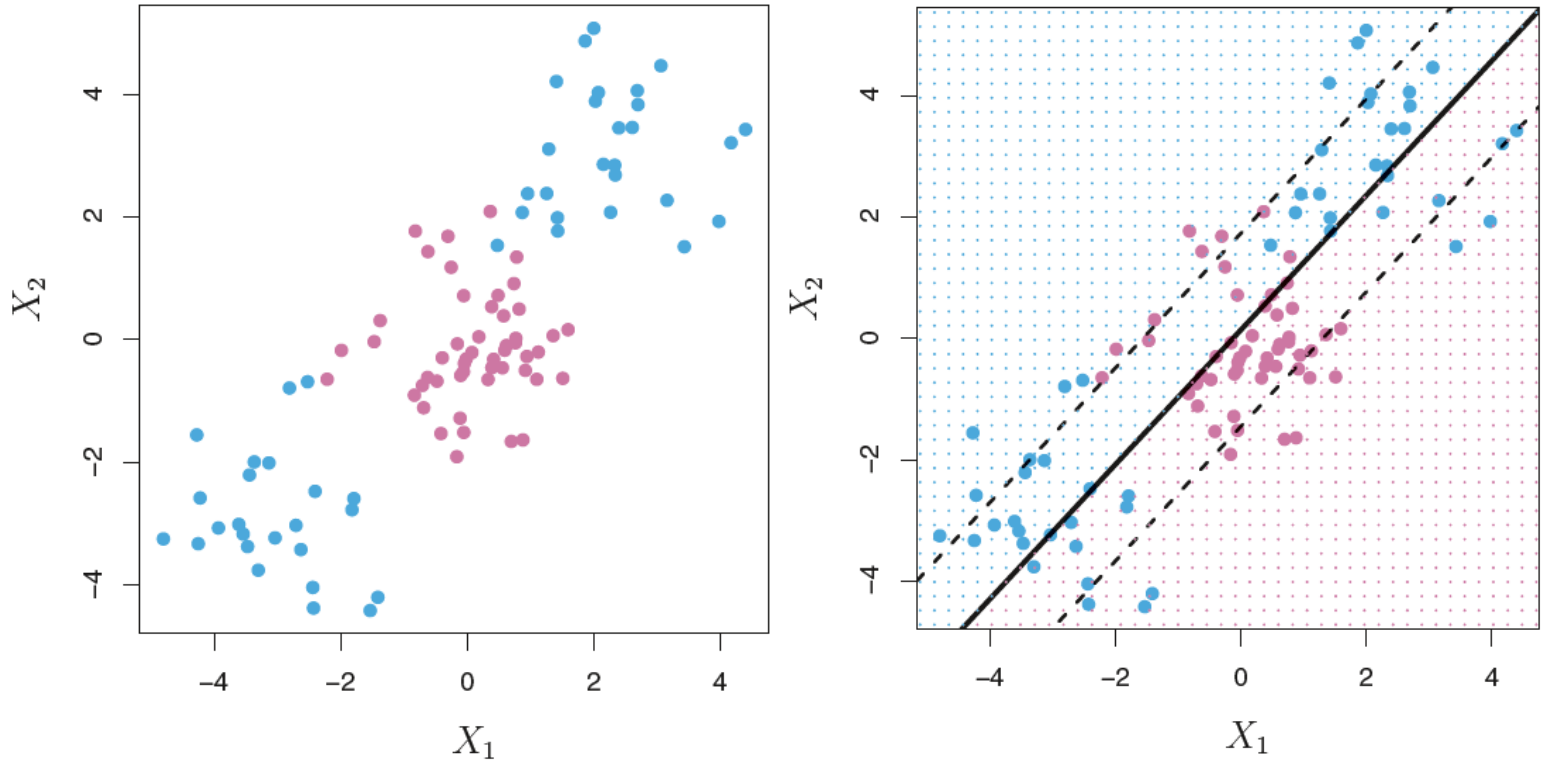
Support Vector Classifiers

Tuning Parameter

- The support vectors of a *support vector classifier* are the observations that lie on the margin or that violate the margin.
- When C is large, there are many support vectors.
- When C is small, there are few support vectors, so the resulting classifier have low bias at high variance.

Support Vector Machines

- We are sometimes faced with non-linear class boundaries. Any linear classifier will perform poorly here.



Support Vector Machines

- *Support vector machine* (SVM) is an extension of the support vector classifier that results from enlarging the feature space using **kernels**.
- The feature space is enlarged in order to accommodate a non-linear boundary between the classes.
- The kernel approach is simply an efficient computational approach for enacting this idea.
- One example of a kernel is the *polynomial kernel* of degree d, given by

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d.$$

Support Vector Machines

- Another kernel is the *radial kernel* given by

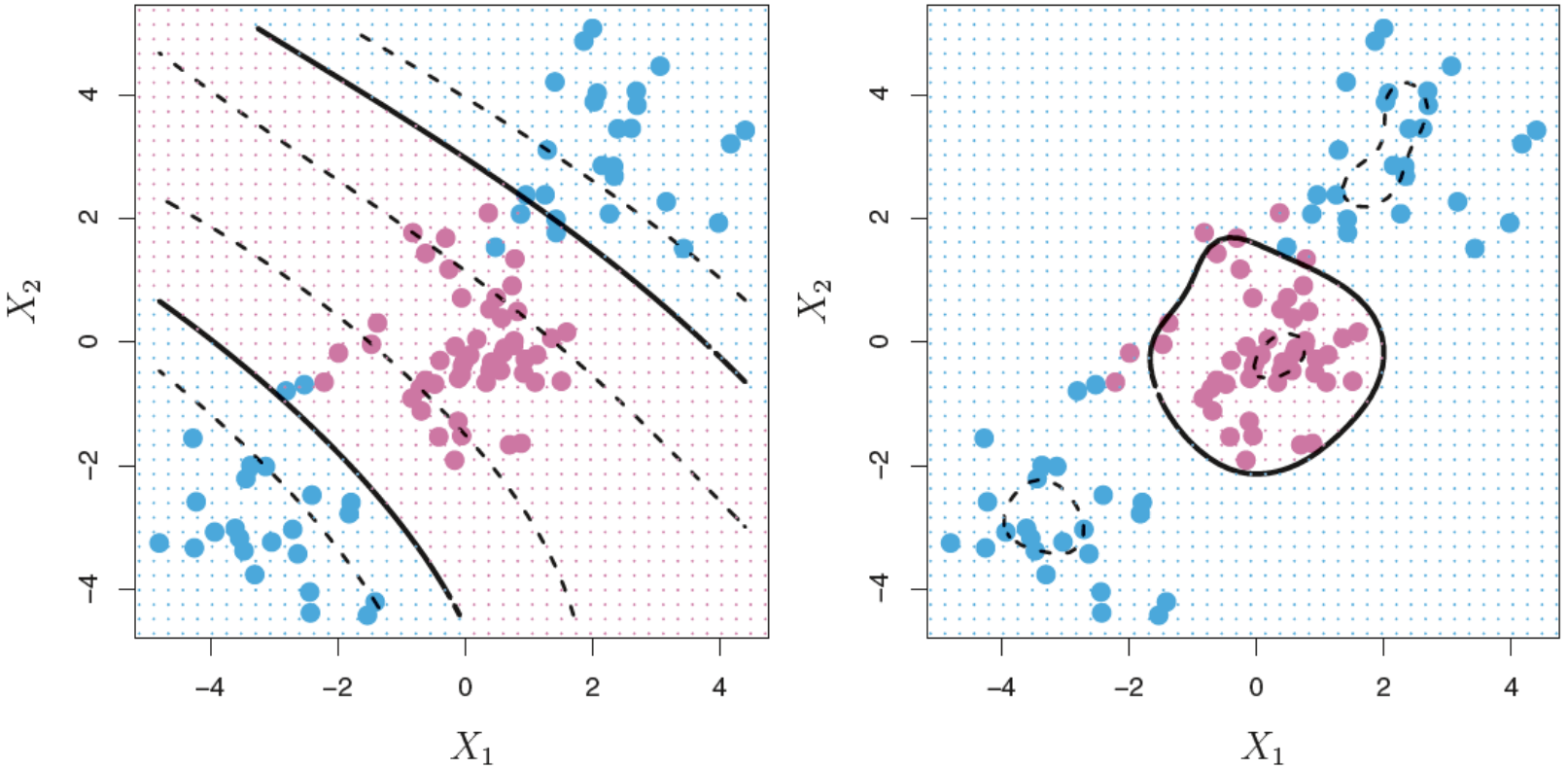
$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

- The advantage of using a kernel is computational.

Support Vector Machines

Illustration

Figure on the left shows a SVM with a polynomial kernel of degree 3. On the right is an SVM with a radial kernel.



Support Vector Machines

SVMs with more than two classes

- The standard SVM strategy for a multiclass classification problem (over K classes) has been to reduce it to a series of binary problems.
- Two strategies: **One-versus-One**, **One-versus-All**

Support Vector Machines

One-versus-One Classification

- This approach constructs SVMs for each pair of classes.
- An observation is classified by tallying the numbers of times the observation is assigned to each of the K classes.
- The final classification is performed by assigning the test observation to the class to which it was most frequently assigned in the pairwise classifications.

Support Vector Machines

One-versus-All Classification

- With K classes, we fit K SVMs, each time comparing one of the K classes to the remaining K – 1 classes.
- We assign an observation to the class for which the value of the classifier is the largest, as this amounts to a high level of confidence that the observation belongs to the kth class rather than to any of the other classes.