6.8: Support Vector Machines

Simone Collier

University of Toronto

Overview

1. Maximal Margin Classifier
2. Support Vector Classifier
3. Support Vector Machines

Introduction

The support vector machine (SVM) is an approach for classification in a binary setting. We will cover:

- The maximal margin classifier
- The support vector classifier
- The support vector machine

The first two methods are specific cases of the support vector machine, but they can be very useful given the right scenario. We will start by introducing the concept of a hyperplane which is what these methods all rely on.

Hyperplane

In a p dimensional space, a hyperplane is a flat $p-1$ dimensional subspace.

- In two dimensions, a hyperplane is a flat one dimensional subsapce (a line) defined by

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

  for parameters β_0, and β_1. That is, any $X = (X_1, X_2)$ that satisfies that equation is on the hyperplane.

- In three dimensions a hyperplane is a flat two dimensional subspace (a plane) defined by

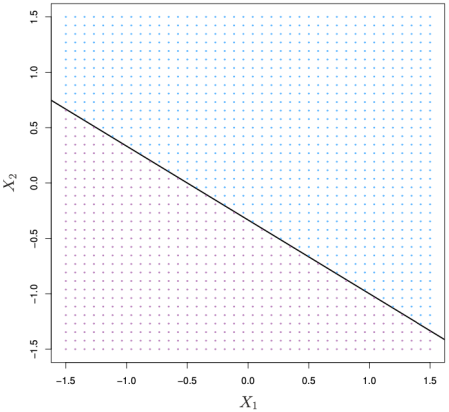
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

- In p dimensions it is a $p-1$ dimensional flat subspace defined by

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

Hyperplane

A hyperplane divides a p dimensional space into two.



- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p > 0$ implies that X is not on the hyperplane and is instead on one side of it.
- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p < 0$ implies that X on the other side of the hyperplane.
- The hyperplane in the figure is $1 + 2X_1 + 3X_2 = 0$

Classification Using a Separating Hyperplane

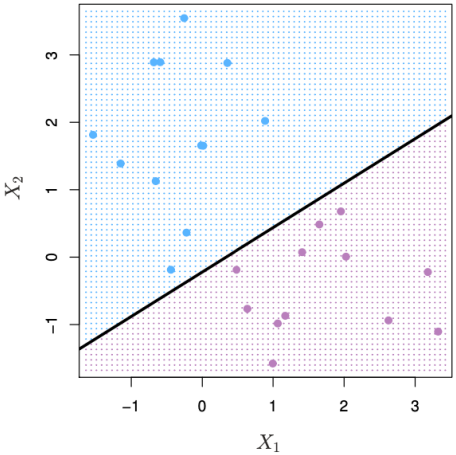
We have the following data:

- n training observations $x_1, \ldots, x_n$, each of which are p dimensional vectors $x_i = (x_{i1}, \ldots, x_{ip})$
- Qualitative response $y_1, \ldots, y_n \in \{-1, 1\}$ where -1 represents one class and 1 represents the other.
- Test observation $x^* = (x_1^*, \ldots, x_p^*)$

The goal is to develop a classifier from the training data that will correctly classify the test observation.

Classification Using a Separating Hyperplane

Suppose that is it possible to construct a hyperplane that separates the training observations according to their class. An example in 2 dimensions:



- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$ implies that X is part of the blue class.
- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$ implies that X is part of the purple class.

Classification Using a Separating Hyperplane

In our problem, we are trying to separate the $y_i = 1$ class from the $y_i = -1$ class which could also just be thought of as the blue class and the purple class.

Our separating hyperplane is constructed based on the properties:

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

If a separating hyperplane exists then we can use it to classify test observations by what side of the hyperplane they are on.

- Classify x^* based on the sign of $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$.
- If $f(x^*)$ is far from zero then x^* is far from the hyperplane and we can be more confident in our classification than if x^* were close to the hyperplane.

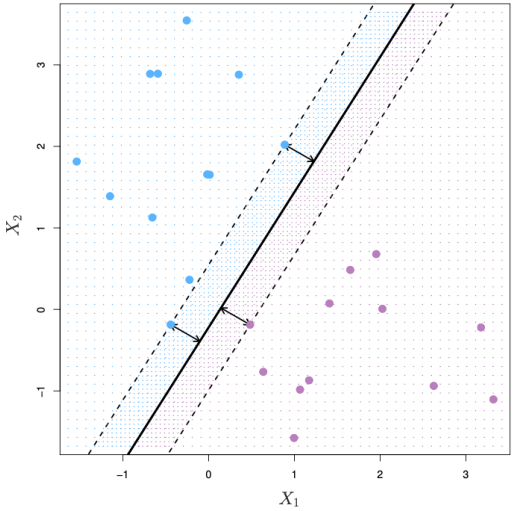
Maximal Margin Classifier

The Maximal Margin Classifier

If our data can be separated by hyperplane then there will be infinitely many separating hyperplanes. The maximal margin hyperplane is the separating hyperplane that is farthest from the training observations.

- Compute the perpendicular distance from each training observation to a given separating hyperplane (known as the **margin**).
- The maximal margin hyperplane is the hyperplane that maximizes the margin.
- Classify a test observation based on which side of the maximal margin hyperplane it is on.

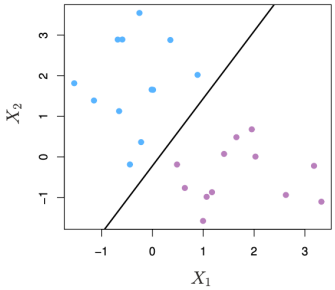
The Maximal Margin Classifier

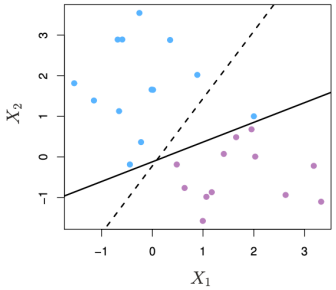


- The maximal margin classifier is the solid line.
- The margin is the distance from the solid line to the dashed lines.
- The three observations that are on the dashed lines are equidistant from the hyperplane and are called **support vectors**.
- The maximal margin classifier only depends directly on the support vectors.

The Maximal Margin Classifier

A separating hyperplane classifier will necessarily perfectly classify the training observations. This can lead to sensitivity to some observations and overfitting.





- Left: Maximal margin hyperplane separates two classes.
- Right: An additional blue training observation is added to the training set causing the hyperplane to shift dramatically.

Exercises: The Maximal Margin Classifier

Open the Support Vector Machines Exercises R Markdown file.

- Go over the "Maximal Margin Classifier" section together as a class.

Support Vector Classifier

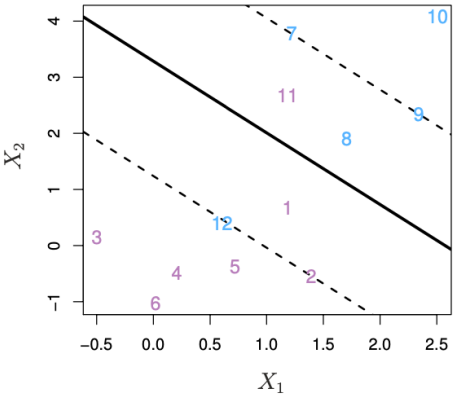
Support Vector Classifier

The support vector classifier is an extension of the maximal margin classifier that uses a **soft margin** which does not perfectly separate the two classes.

- Greater robustness to individual observations.
- Better classification of most of the training observations.
- Can accommodate data sets that are not perfectly separable by a hyperplane.

The idea is that misclassifying a few training observations could help to better classify the remaining observations.

Support Vector Classifier



- The support vector classifier hyperplane is the solid line and the margins are the dashed lines.
- The observations below the hyperplane are classified as purple and those above are classified as blue.
- Observations 1 and 8 are intentionally on the wrong side of the margin.
- Observations 11 and 12 are intentionally on the wrong side of the hyperplane and the wrong side of the margin.

Support Vector Classifier

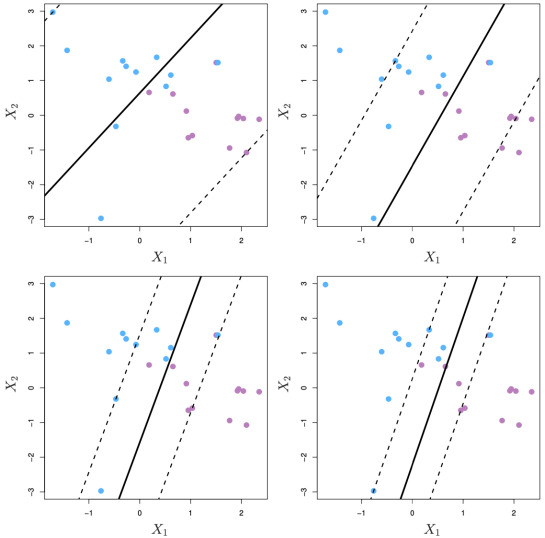
The support vector classifier is constructed using several parameters, the most important being the tuning parameter C which determines the number of and severity of violations to the margin and hyperplane that we will allow.

- No more than C observations can be on the wrong side of the hyperplane.
- $C = 0$ implies that no violations will be allowed so this gives the maximal margin hyperplane.
- As C increases the margin will widen.
- C is chosen using cross-validation.
- C controls the bias-variance trade-off of the model.
  - if C is small, then the classifier is highly fit to the data which yields low bias, high variance.
  -] if C is larger, then the classifier is fit less hard to the data which yields reduces variance and increases bias.

Support Vector Classifier

- The support vector classifier aims to make M, the width of the margin, as large as possible while staying within the budget of margin violations C.
- Observations that lie directly on the margin or on the wrong side of it are **support vectors**.
- The support vector classifier is only depends directly on the support vectors.
- As C increases, so does the number of support vectors. This means there are more observations that determine the hyperplane (hence lower variance).

Support Vector Classifier



A support vector classifier fit to the same data set with four different different values of C.

Exercises: Support Vector Classifier

Open the Support Vector Machines Exercises R Markdown file.

- Go through the "Support Vector Classifier" section together as a class.
- When a question is reached, allow 10 minutes for the students to work on it.
- Questions should be completed at home if time does not allow.
- Go over the rest of the "Support Vector Classifier" section together as a class.

Support Vector Machines

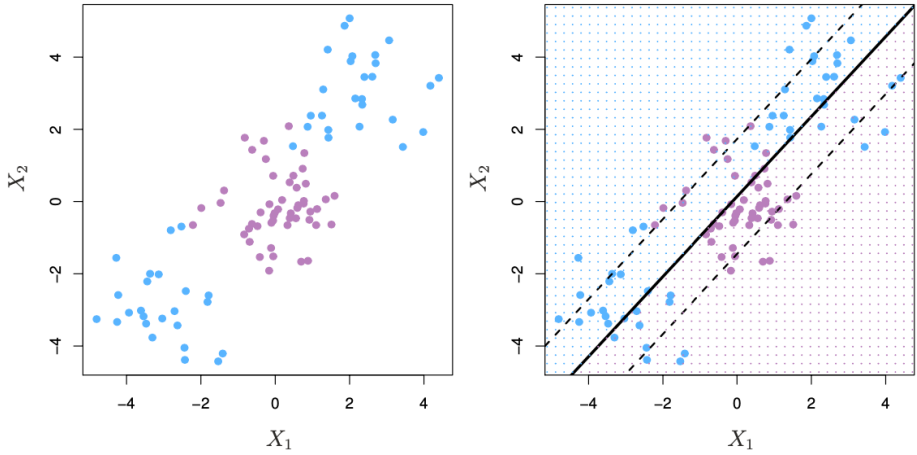
Support Vector Machines

The support vector machine (SVM) is an extension of the support vector classifier that enlarges the feature space in order to accommodate a non-linear boundary between classes.

- Uses **kernals** which are functions that quantify similarities between two observations.
- The support vector classifier happens to be fit with kernals as well, namely a polynomial kernal of degree 1 (linear).
- SVMs use non-linear kernals such as high degree polynomial kernals or radial kernals in order to get a more flexible boundary.
- The technical details of SVMs are out of scope for this course.

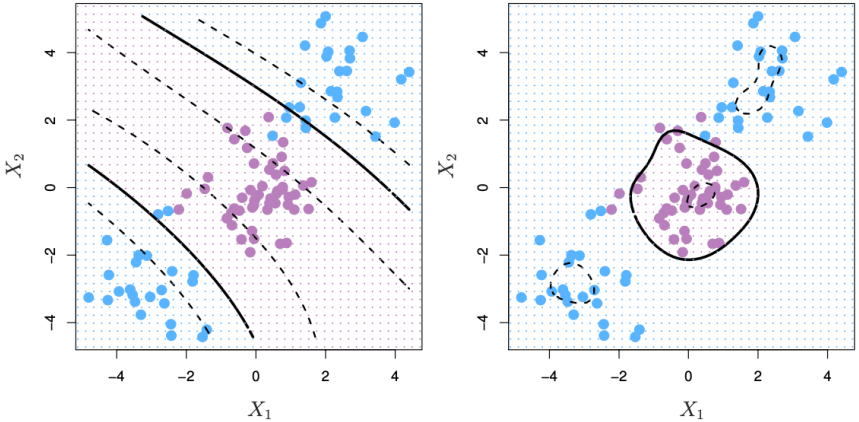
Support Vector Machines

A data set with two classes is fit with a support vector classifier. The linear boundary does not perform well.



Support Vector Machines

The data set is now fit with two different support vector machines.



- Left: SVM with a polynomial kernal of degree 3.
- Right: SVM with a radial kernal.

These SVMs capture the decision boundary much better than the support vector classifier.

SVMs with More than Two Classes

Extending the concept of separating hyperplanes to $K > 2$ classes is actually quite tricky. The two main approaches for this are briefly described.

- One-versus-one classification
  1. We construct SVMs to compare each combination of two classes.
  2. We classify a test observation to one of two classes using each of the SVM classifiers.
  3. The test observation is finally assigned to the class to which is was most frequently classified by the SVMs.

- One-versus-all classification
  1. We construct K SVMs which each compare one of the classes to the rest of the $K - 1$ classes.
  2. We assign the observation to the class for which the observation is the farthest away from the hyperplane (on the correct side).

Exercises: Support Vector Machine

Open the Support Vector Machines Exercises R Markdown file.

- Go through the "Support Vector Machine" section together as a class.
- 10 minutes to complete the questions at the end of the section.
- Questions should be completed at home if time does not allow.

References

Chapter 9 of the ISLR2 book:

James, Gareth, et al. "Support Vector Machines." An Introduction to Statistical Learning: with Applications in R, 2nd ed., Springer, 2021.