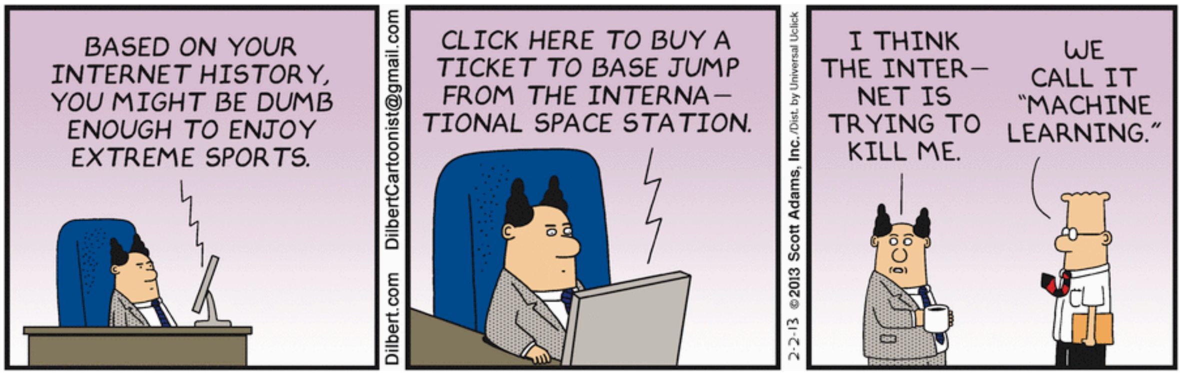
Chapter 9: Support Vector Machines

The *support vector machine* is an approach for classification that was developed in the computer science community in the 1990s and has grown in popularity.

The support vector machine is a generalization of a simple and intuitive classifier called the *maximal margin classifier*.

Support vector machines are intended for binary classification, but there are extensions for more than two classes.


BASED ON YOUR INTERNET HISTORY, YOU MIGHT BE DUMB ENOUGH TO ENJOY EXTREME SPORTS.
Dilbert.com DilbertCartoonist@gmail.com
CLICK HERE TO BUY A TICKET TO BASE JUMP FROM THE INTERNATIONAL SPACE STATION.
2-2-13 ©2013 Scott Adams, Inc./Dist. by Universal Uclick
I THINK THE INTERNET IS TRYING TO KILL ME.
WE CALL IT "MACHINE LEARNING."

Credit: https://dilbert.com/strip/2013-02-02

1 Maximal Margin Classifier

In p-dimensional space, a *hyperplane* is a flat affine subspace of dimension $p-1$.

The mathematical definition of a hyperplane is quite simple,

This can be easily extended to the p-dimensional setting.

We can think of a hyperplane as dividing p-dimensional space into two halves.

1.1 Classificaton Using a Separating Hyperplane

Suppose that we have a $n \times p$ data matrix $\boldsymbol{X}$ that consists of n training observations in p-dimensional space.

and that these observations fall into two classes.

We also have a test observation.

Our Goal:

Suppose it is possible to construct a hyperplane that separates tthe training observations perfectly according to their class labels.

Then a separating hyperplane has the property that

If a separating hyperplane exists, we can use it to construct a very natural classifier:

That is, we classify the test observation x^* based on the sign of $f(x^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$.

We can also use the magnitude of $f(x^*)$.

1.2 Maximal Margin Classifier

If our data cab ve perfectly separated using a hyperplane, then there will exist an infinite number of such hyperplanes.

A natural choice for which hyperplane to use is the *maximal margin hyperplane* (aka the *optimal separating hyperplane*), which is the hyperplane that is farthest from the training observations.

We can then classify a test observation based on which side of the maximal margin hyperplane it lies – this is the *maximal margin classifier*.

We now need to consider the task of constructing the maximal margin hyperplane based on a set of n training observations and associated class labels.

The maximal margin hyperplane is the solution to the optimization problem

This problem can be solved efficiently, but the details are outside the scope of this course.

What happens when no separating hyperplane exists?

2 Support Vector Classifiers

It’s not always possible to separate training observations by a hyperplane. In fact, even if we can use a hyperplane to perfectly separate our training observations, it may not be desirable.

We might be willing to consider a classifier based on a hyperplane that does *not perfectly* separate the two classes in the interest of

-
-

The *support vector classifier* does this by finding the largest possible margin between classes, but allowing some points to be on the “wrong” side of the margin, or even on the “wrong” side of the hyperplane.

The support vector classifier xlassifies a test observation depending on which side of the hyperplane it lies. The hyperplane is chosen to correctly separate **most** of the training observations.

Once we have solved this optimization problem, we classify x^* as before by determining which side of the hyperplane it lies.

ϵ_i

C

The optimization problem has a very interesting property.

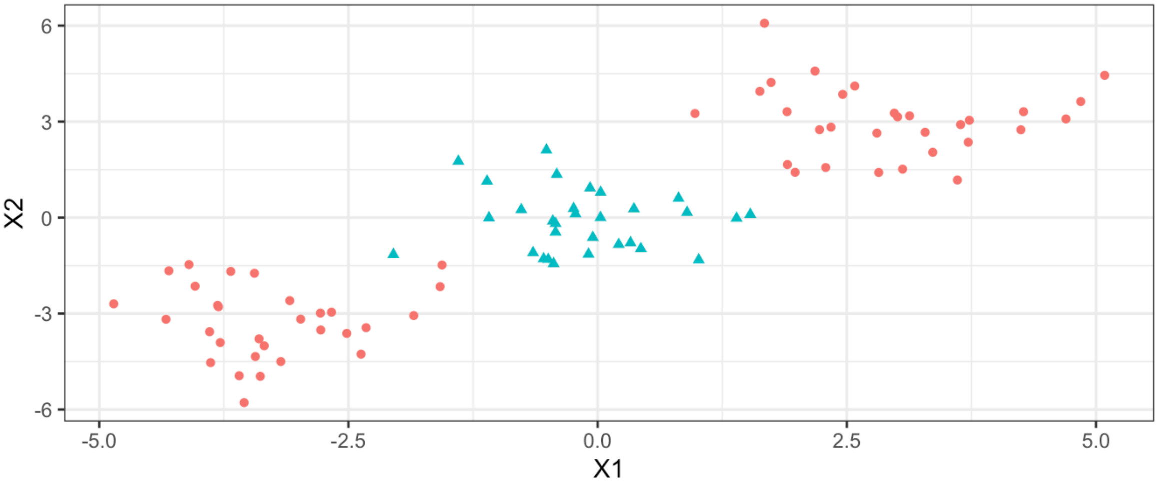
Observations that lie directly on the margin or on the wrong side of the margin are called *support vectors*.

The fact that only support vectors affect the classifier is in line with our assertion that C controls the bias-variance tradeoff.

Because the support vector classifier's decision rule is based only on a potentially small subset of the training observations means that it is robust to the behavior of observations far away from the hyperplane.

3 Support Vector Machines

The support vector classifier is a natural approach for classification in the two-class setting...



We've seen ways to handle non-linear classification boundaries before.

In the case of the support vector classifier, we could address the problem of possible non-linear boundaries between classes by enlarging the feature space.

Then our optimization problem would become

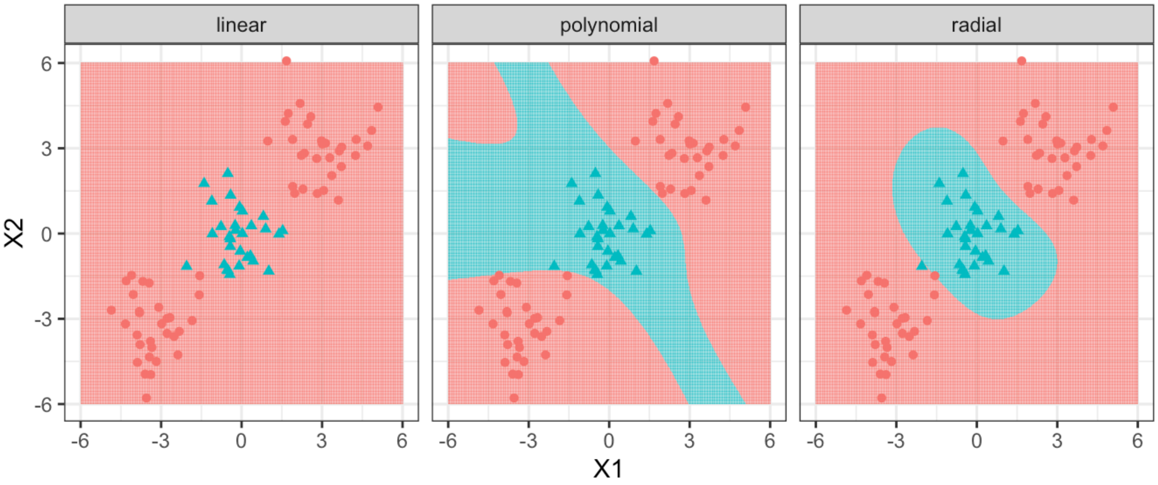
The *support vector machine* allows us to enlarge the feature space used by the support classifier in a way that leads to efficient computation.

It turns out that the solution to the support vector classification optimization problem involves only *inner products* of the observations (instead of the observations themselves).

It can be shown that

-
-
-

Now suppose every time the inner product shows up in the SVM representation above, we replaced it with a generalization.


linear
polynomial
radial
X2
X1
6
3
0
-3
-6
-6
-3
0
3
6

4 SVMs with More than Two Classes

So far we have been limited to the case of binary classification. How can we exted SVMs to the more general case with some arbitrary number of classes?

Suppose we would like to perform classification using SVMs and there are $K > 2$ classes.

One-Versus-One

One-Versus-All