# STAT562 Lecture 14 Support Vector Machines

Beidi Qiang

SIUE

# Introduction

Support vector machine (SVM) is an approach for classification that was developed in the 1990s and that has grown in popularity since then. SVMs have been shown to perform well in a variety of settings, and are often considered one of the best classifiers. We will introduce SVM in the following settings:

- Classes are separable by a linear boundary (hyperplane) in binary setting
- Classes are not separable in binary setting
- Extension of SVM to nonlinear class boundaries in binary setting
- Extensions of SVM to more than two classes.

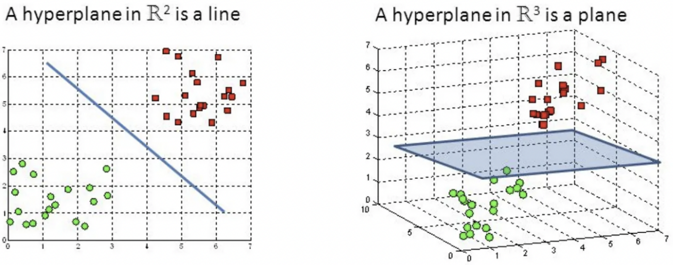
# Hyperplane

In a p-dimensional space, hyperplane is a flat (p-1)-dimensional subspace , defined as

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

- In two dimensions, a hyperplane is a line. In three dimensions, a hyperplane is a plane.
- If $X = (X_1 \cdots X_p)^T$ satisfy $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0$, we say $X$ lies on the hyperplane.
- $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p > 0$, or $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p < 0$, tells us that $X$ lies to one or the other side of the hyperplane.

# 2D and 3D Hyperplane


A hyperplane in $\mathbb{R}^2$ is a line

A hyperplane in $\mathbb{R}^3$ is a plane

Hyperplanes in 2D and 3D feature space

# Classification Using a Separating Hyperplane

Our goal is to develop a classifier based on the training data that will correctly classify the test observation using the predictors.

- Assuming in binary case, the observations fall into two classes, labeled as $y = -1$ and $y = 1$.
- Suppose that it is possible to construct a hyperplane that separates the training observations perfectly according to their class labels, i.e.,
  $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0$, for all $i$ when $y_i = -1$
  $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0$, for all $i$ when $y_i = 1$
- Equivalently, $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0$ for all observation in the training.
- A test observation is assigned a class depending on which side of the hyperplane it is located.

# Separating Hyperplane

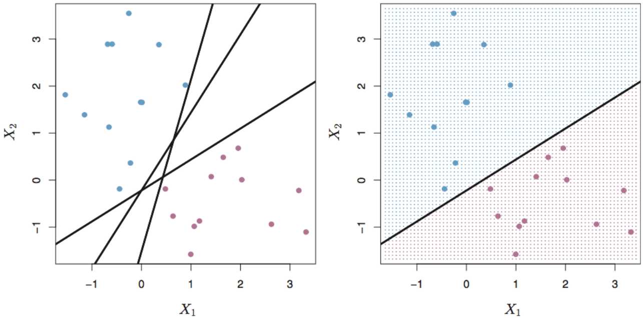


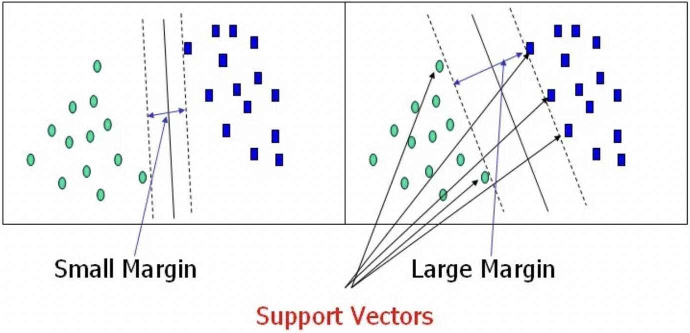
**FIGURE 9.2.** Left: *There are two classes of observations, shown in blue and in purple, each of which has measurements on two variables. Three separating hyperplanes, out of many possible, are shown in black.* Right: *A separating hyperplane is shown in black. The blue and purple grid indicates the decision rule made by a classifier based on this separating hyperplane: a test observation that falls in the blue portion of the grid will be assigned to the blue class, and a test observation that falls into the purple portion of the grid will be assigned to the purple class.*

# Maximal Margin Classifier

If the training data can be perfectly separated using a hyperplane, then there will in fact exist an infinite number of such hyperplanes. We must have a reasonable way to decide on one. A natural choice is the "maximal margin hyperplane".

- We can compute the distance from each training observation to a hyperplane. The smallest such distance is known as the margin.
- The maximal margin hyperplane is the separating hyperplane for which the margin is largest
- The maximal margin hyperplane is the separating hyperplane that is farthest from the training observations.

# Maximal Margin



# Support Vectors

- Training observations that are on the margin are called "support vectors". They support the maximal margin hyperplane.
- If the support vectors moved slightly then the maximal margin hyperplane would move as well.
- A small movement to any of the other observations would not affect the separating hyperplane.
- the maximal margin hyperplane depends directly on only a small subset of the observations, i.e. the support vectors.

# Construction of the Maximal Margin Classifier

The maximal margin hyperplane is the solution to the optimization problem:

- Chooses $\beta_0, \cdots, \beta_p$ to maximize M, subject to

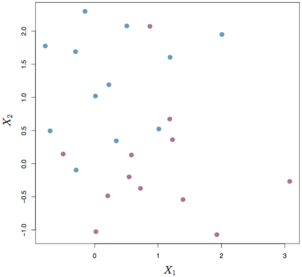
$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M \text{ for all i.}$$

- Every observation need to be not only on the correct side of the hyperplane but also on the correct side of the margin.
- With the constraints, one can show the perpendicular distance from the *i*th observation to the hyperplane is given by $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$.
- The optimization problem is usually done numerically.

# Non-separable Case

In many cases no separating hyperplane exists.



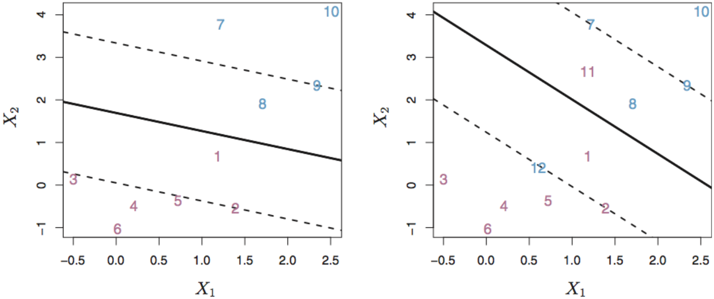
- In this case, the optimization problem on the previous slide has no solution with $M > 0$.
- We will extend the concept of a separating hyperplane in order to develop a hyperplane that almost separates the classes, using a so-called "soft margin".

# Soft margin classifier

We might need to consider a classifier based on a hyperplane that does not perfectly separate the two classes.

- In the case of non-separable
- Or it could be worthwhile to misclassify a few training observations in order to do a better job in classifying the remaining observations.
- We allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane.
- The margin is "soft" because it can be violated by some of the training observations.

# Soft Margin



# Construction of soft margin SVM

The soft margin support vector classifier is the solution to the optimization problem:

- Chooses $\beta_0, \cdots, \beta_p$ to maximize M, subject to

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \text{ for } \epsilon_i \geq 0$$

$$\sum_{j=1}^{p} \beta_j^2 = 1, \text{and } \sum_{i=1}^{n} \epsilon_i \leq C$$

- $C$ is a nonnegative tuning parameter.
- $\epsilon_i$ are slack variables that allow individual observations to be on the wrong side.
- We classify a test observation as before, by simply determining on which side of the hyperplane it lies.

# The tuning parameter C

- $C$ determines the number and severity of the violations to the margin that we will tolerate.
- Note if $\epsilon_i = 0$, the observation is on the correct side of the margin. If $\epsilon_i > 0$, the observation violated the margin. Further, If $\epsilon_i > 1$, the observation is on the wrong side of the hyperplane(misclassified).
- If $C = 0$ then there is no budget for violations,
- For $C > 0$, there can be no more than C observations misclassified.
- In general, a larger $C$ will allow for a wider margin.
- $C$ is generally chosen via cross-validation.

# Remarks

- Observations that lie directly on the margin, or on the wrong side of the margin for their class, are known as support vectors.
- Only observations that either lie on the margin or that violate the margin (the support vectors) will affect the hyperplane.
- When the tuning parameter C is large, then we allow many observations violate the margin (more support vectors). In this case, many observations are involved in determining the hyperplane. We have low variance but high bias.
- If C is small, then there will be fewer support vectors and hence the resulting classifier will have low bias but high variance.

# Example: Simulated Data

We will use e1071 package in R to train SVM models. (See attached R code) .

The parameterization used in e1071 is slightly different from the one introduced in the text book. The optimization in e1071 goes like

- Chooses $\beta_0, \cdots, \beta_p$ to minimize

$$1/M + C\sum_{i=1}^{n} \epsilon_i$$

$$\text{subject to } \frac{1}{M} y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq (1 - \epsilon_i)$$

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

- C is called the "cost", which controls the trade-off between the slack variable penalty and the margin.