Mathematical techniques in data science

Lecture 8: Support Vector Machines

March 6th, 2019

- Maximal Margin Classifier
- Support Vector Classifiers
- Support Vector Machines

- Friday (03/08): Homework 1 due
- Groups (for class projects) need to be formed by the end of the week

Hyperplane

- In a p-dimensional space, a hyperplane is an affine subspace of dimension p.
- In two dimensions, a hyperplane is defined by the equation

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

- In p dimensions:

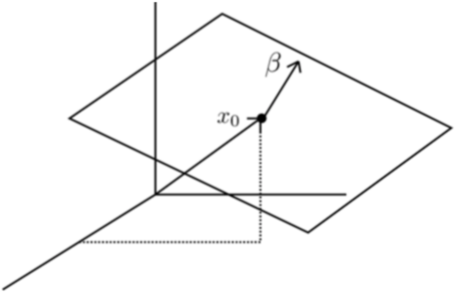
$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = 0$$

or alternatively

$$\beta_0 + \beta^T x = 0, \quad \text{where } \beta \in \mathbb{R}^p$$

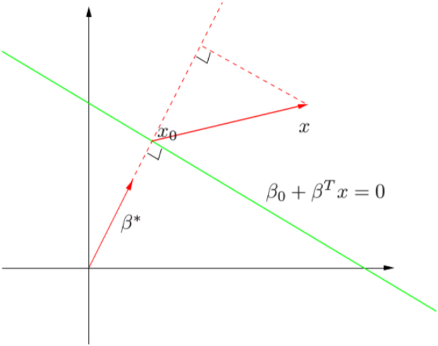
Hyperplane

$$H = \{x \in \mathbb{R}^p : \beta_0 + \beta^T x = 0\}$$



If $x_1, x_2 \in H$, then $\beta^T(x_1 - x_2) = 0 \rightarrow \beta$ is perpendicular to the hyperplane H

Hyperplane



If $x \in \mathbb{R}^p$, the distance from x to H can be computed by

$$d(x, H) = \frac{1}{\|\beta\|}|\beta^T(x - x_0)| = \frac{|\beta_0 + \beta^T x|}{\|\beta\|}$$

Hyperplane

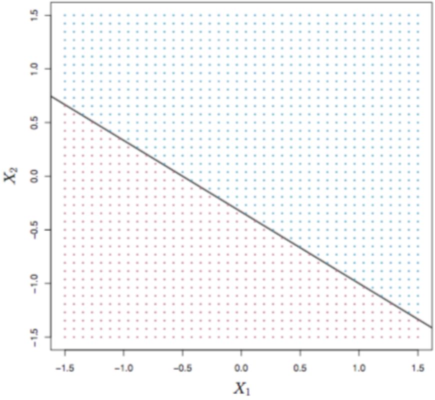
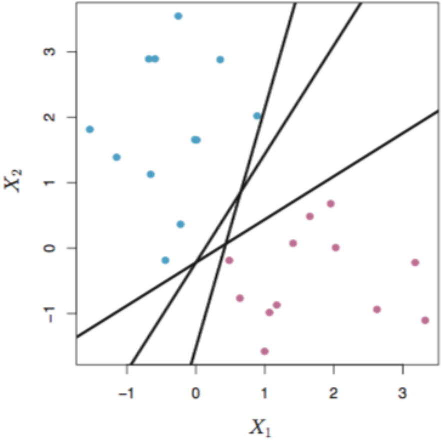


FIGURE 9.1. *The hyperplane* $1 + 2X_1 + 3X_2 = 0$ *is shown. The blue region is the set of points for which* $1 + 2X_1 + 3X_2 > 0$*, and the purple region is the set of points for which* $1 + 2X_1 + 3X_2 < 0$.

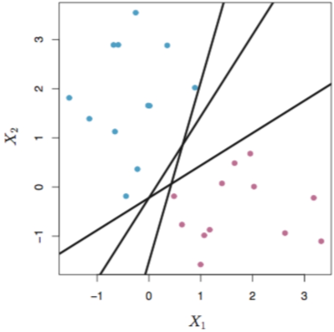
Separating hyperplane

Suppose we have data with label $\{-1, 1\}$, we want to separate the data using a hyperplane

$$y_i = \text{sign}(\beta_0 + \beta^T x_i)$$



Separating hyperplane



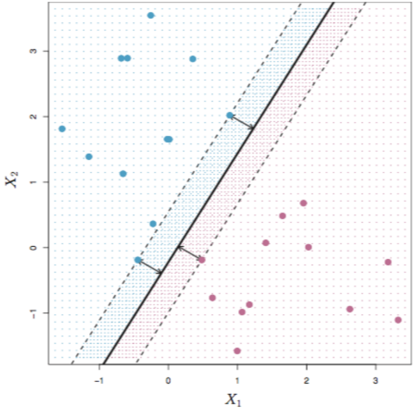
Problems:

- Separating hyperplane may not exist
- Assume that the data are perfectly separable by a hyperplane $\rightarrow$ then there might exist an infinite number of such hyperplanes

Maximal Margin Classifier

Maximal Margin Classifier

- Assume that the data are perfectly separable by a hyperplane
- The minimal distance from the data to the hyperplane is call the *margin*
- Maximal margin hyperplane: the separating hyperplane that is farthest from the training observations



Maximal Margin Classifier: formulation

- Given a set of n training observations $x_1, \ldots, x_n \in \mathbb{R}$ and associated class labels $y_i \in \{-1, 1\}$
- Maximal margin hyperplane:

$$\max_{\beta_0, \beta, M} M$$

$$\text{subject to } \|\beta\| = 1$$

$$\text{and } y_i(\beta_0 + \beta^T x_i) \geq M \quad \forall i = 1, \ldots, n.$$

Why?

- First, for every separating hyperplane, we want the classifier associated with the hyperplane to predict the labels correctly, or
$$y_i(\beta_0 + \beta^T x_i) \geq 0 \quad \forall i = 1, \ldots, n.$$
- Second, we want the distance from the points to the hyperplane to be greater than the margin
$$\frac{|\beta_0 + \beta^T x_i|}{\|\beta\|} \geq M$$
- If we constrain $\|\beta\| = 1$ then this becomes
$$y_i(\beta_0 + \beta^T x_i) \geq M \quad \forall i = 1, \ldots, n.$$
- The idea of MMC is to find the separating hyperplane that maximizes the margin

MMC: Alternative form

$$\max_{\beta_0, \beta, M} M$$

$$\text{subject to } \|\beta\| = 1$$

$$\text{and } y_i(\beta_0 + \beta^T x_i) \geq M \quad \forall i = 1, \ldots, n.$$

- If we remove the constraint $\|\beta\| = 1$ then the optimization problem becomes

$$\max_{\beta_0, \beta, M} M$$

$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq M\|\beta\| \quad \forall i = 1, \ldots, n.$$

MMC: Alternative form

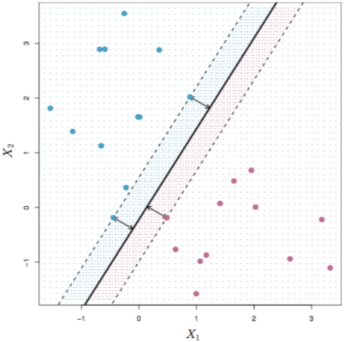
$$\max_{\beta_0,\beta,M} M$$
$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq M\|\beta\| \quad \forall i = 1, \ldots, n.$$

- If we rescale (β_0, β) such that $M\|\beta\| = 1$, then the optimization problem becomes

$$\min_{\beta_0,\beta} \|\beta\|^2$$
$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq 1 \quad \forall i = 1, \ldots, n.$$

- This is a convex optimization problem with a quadratic object and linear constraints

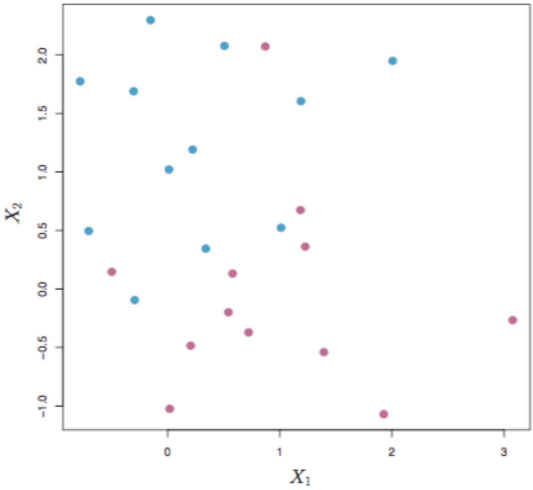
Remark: support vectors



In this figure, we see that three training observations are equidistant from the maximal margin hyperplane and lie along the dashed lines indicating the width of the margin.

Support Vector Classifiers

Realistically, data are not separable by hyperplanes



MMC is not robust to noises

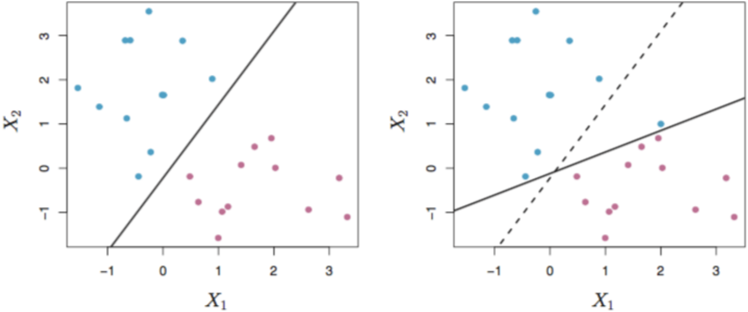


FIGURE 9.5. Left: *Two classes of observations are shown in blue and in purple, along with the maximal margin hyperplane.* Right: *An additional blue observation has been added, leading to a dramatic shift in the maximal margin hyperplane shown as a solid line. The dashed line indicates the maximal margin hyperplane that was obtained in the absence of this additional point.*

Support Vector Classifier

- Idea: willing to consider a classifier based on a hyperplane that does not perfectly separate the two classes
- Goals:
  - Greater robustness to individual observations
  - Better classification of most of the training observations

Support Vector Classifier

The hyperplane is chosen to correctly separate most of the training observations into the two classes, but may mis-classify a few observations

$$
\begin{aligned}
&\max_{\beta_0,\beta,M,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} M \\
&\text{subject to } \|\beta\| = 1 \\
&\qquad\qquad y_i(\beta_0 + \beta^T x_i) \geq M(1-\epsilon_i) \quad \forall i = 1,\ldots,n \\
&\qquad\qquad \epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C.
\end{aligned}
$$

$$
\max_{\beta_0,\beta,M,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} M
$$

$$
\text{subject to } \|\beta\| = 1
$$

$$
y_i(\beta_0 + \beta^T x_i) \geq M(1 - \epsilon_i) \quad \forall i = 1, \ldots, n
$$

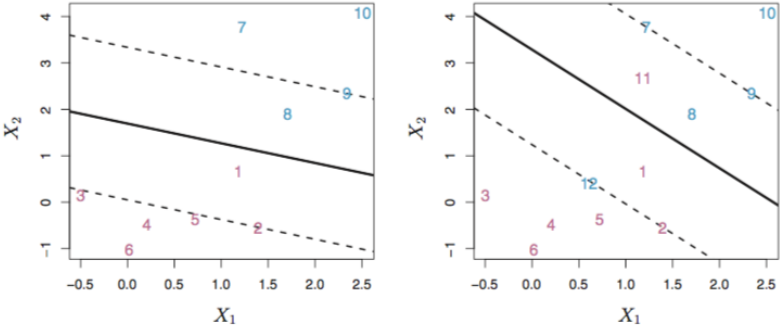
$$
\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C.
$$

- $\epsilon_1, \ldots, \epsilon_n$ are refereed to as *slack variables*
- C can be regarded as a budget for the amount that the margin can be violated by the n observations

Slack variables

- $\epsilon_1, \ldots, \epsilon_n$ are refereed to as *slack variables*
- If $\epsilon_i = 0$, the i^{th} observation is on the correct side of the margin
- If $\epsilon_i > 0$, the i^{th} observation is on the wrong side of the margin
- If $\epsilon_i > 1$, the i^{th} observation is on the wrong side of the separating hyperplane

Support Vector Classifier



Budget

- C can be regarded as a budget for the amount that the margin can be violated by the n observations
- If $C = 0$ then there is no budget for violations to the margin
  $\rightarrow \epsilon_i = 0$ for all i
  $\rightarrow$ maximal margin classifier
- Budget C increases $\rightarrow$ more tolerant of violations to the margin $\rightarrow$ margin will widen
- is a tunable parameter, usually chosen by cross-validation

SVC: alternative form

The hyperplane is chosen to correctly separate most of the training observations into the two classes, but may misclassify a few observations

$$\min_{\beta_0,\beta,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} \|\beta\|^2$$

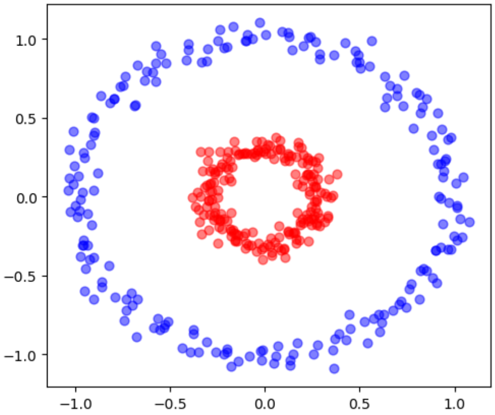
$$\text{subject to } y_i(\beta_0 + \beta^T x_i) \geq (1 - \epsilon_i) \quad \forall i = 1, \ldots, n$$

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

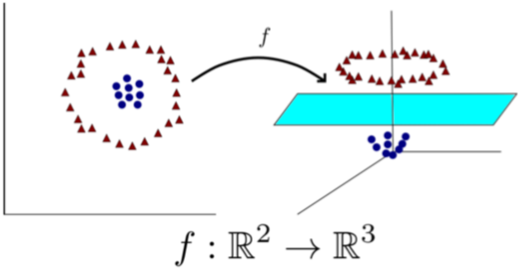
Can be solved using standard optimization packages.

Support Vector Machine

Realistically, the boundary may be non-linear



Idea: map the learning problem to a higher dimension



$$f(x, y) = (x, y, x^2 + y^2)$$

Idea: map the learning problem to a higher dimension

More rigorously,

$$f(x, y) = (x, y, x^2, y^2, xy)$$

A hyperplane on $\mathbb{R}^5$, modeled by the equation $\beta_0 + \beta^T x = 0$ will classify the points based on the sign of

$$\beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy$$

This corresponds to a quadratic boundary on the original space $\mathbb{R}^2$