LECTURE 9

BOGDAN ICHIM

Classification using a Separating Hyperplane

Assume we have N training observations $X_n = (x_{n1}, ..., x_{np}) \in \mathbb{R}^p$, $n = \overline{1, N}$, each with p predictors. For each observation we have a class label $y_n \in \{-1, 1\}$.

Suppose that it is possible to construct a hyperplane H_β that separates the training observations perfectly according to the labels. In other words:

$$\left.\begin{array}{l} y_n = 1 \Rightarrow x_n \in H_\beta^> \Leftrightarrow \beta(x_n) > 0 \Leftrightarrow \beta_0 + \sum_{i=1}^p \beta_i x_{ni} > 0 \\ y_n = -1 \Rightarrow x_n \in H_\beta^< \Leftrightarrow \beta(x_n) < 0 \Leftrightarrow \beta_0 + \sum_{i=1}^p \beta_i x_{ni} < 0 \end{array}\right\} \Leftrightarrow y_n \beta(x_n) > 0$$

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

For a test instance x we compute $\beta(x)$:

$\beta(x) > 0 \Rightarrow x \in H_\beta^> \Rightarrow$ we assign x to class 1;
$\beta(x) < 0 \Rightarrow x \in H_\beta^< \Rightarrow$ we assign x to class -1.

We may also use the **magnitude** of $\beta(x)$.

If $|\beta(x)| \gg 0$ this means that x lies far from the hyperplane H_β and we can be confident about our class assignment for x. On the other hand, if $\beta(x) \simeq 0$, then x is located near the hyperplane and so we are less certain about our class assignment for x.

The Maximal Margin Classifier

Remark 1. If the data can be separated perfectly using a hyperplane, then there exist *infinitely many hyperplanes* that can separate the data.

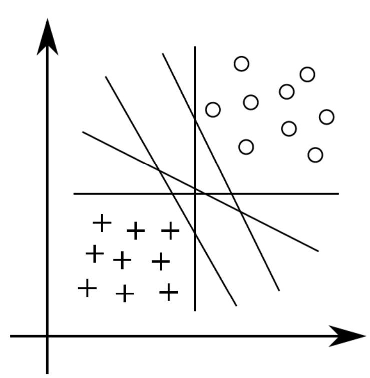

FIGURE 1. Hyperplanes separating the data

We have to decide which of the infinitely many possible separating hyperplanes is the optimal one.

The **maximal margin hyperplane** (or **optimal separating hyperplane**) is the separating hyperplane that is the farthest from the training observations, with respect to a certain metric. We compute the perpendicular distance from each training observation to a given separating hyperplane; the smallest such distance is *the minimal distance from the observations to the hyperplane* and it is called *the margin*.

The **maximal margin hyperplane** is the separating hyperplane for which the **margin is the largest**.

Support Vector Classifier

IF a separating hyperplane does exist, **THEN** the maximal margin classifier is a very natural way to perform classification. However, in many cases no separating hyperplane exists and so there is no maximal margin classifier.

A common solution to this is to allow the classifier to *misclassify* a certain number of points.

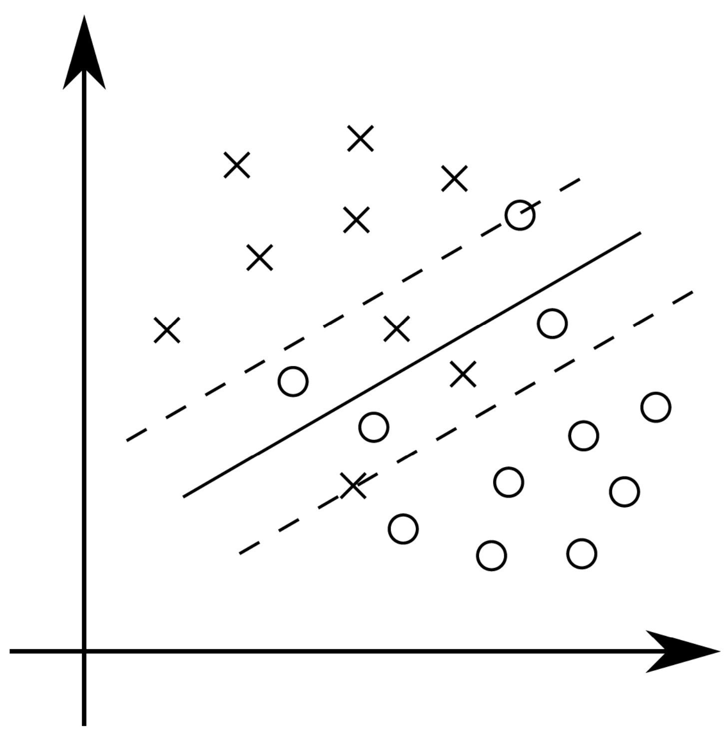

FIGURE 2. Misclassifing the points

The **support vector classifier** (**soft margin classifier**) classifies a test observation depending on which side of the hyperplane it lies. The hyperplane is chosen to correctly separate **most** of the training observations into two classes, but may misclassify **a few** observations.

It is the solution to the following optimization problem:

$$\max_{\beta_0,\dots,\beta_p,\epsilon_1,\dots,\epsilon_N} M \text{ subject to:}$$

$$\begin{cases} \sum_{j=1}^{p} \beta_j^2 = 1 \\ y_n(\beta_0 + \sum_{i=1}^{p} \beta_i x_{ni}) \geq M(1-\epsilon_n),\ n = \overline{1, N} \\ \epsilon_n \geq 0 \\ \sum_{n=1}^{N} \epsilon_n \leq C \end{cases}$$

- M = width of the margin. As for the maximal margin classifier, we want to make it as large as possible.
- C = tuning parameter.
  $C = 0 \Rightarrow \epsilon_n = 0 \Rightarrow$ we optimize the maximal margin classifier.
  Remark: This may not have a solution for $M > 0$!
- $\epsilon_1, \epsilon_2, \dots, \epsilon_N$ = slack variables.
  The slack variable ϵ_n tells us where the n^{th} observation is located relative to the hyperplane and to the margin.
  - $\epsilon_n = 0 \Rightarrow$ the n^{th} observation is on the correct side of the margin.
  - $\epsilon_n \in (0, 1] \Rightarrow$ the n^{th} observation is on the wrong side of the margin; the n^{th} observation has violated the margin.
  - $\epsilon_n > 1 \Rightarrow$ the n^{th} observation is on the wrong side of the hyperplane.

Remarks:

a) If an observation is on the wrong side of the hyperplane, then $\epsilon_n > 1 \Rightarrow$ if k observations are on the *wrong side of the hyperplane*, then:

$$\left.\begin{matrix} \sum_{i=1}^{N} \epsilon_n > k \\ \sum_{i=1}^{N} \epsilon_n < C \end{matrix}\right\} \Rightarrow \boxed{\text{k < C}}$$

Therefore, **C controls the number of observations on the wrong side of the hyperplane**.

b) As a consequence of the previous proposition, **C controls the bias-variance trade**:
   - C small $\Rightarrow$ the margin is narrow, thus we have low bias and high variance.
   - C large $\Rightarrow$ the margin is wide, thus we have high bias and low variance.

c) Slack variables measure the error in case of misclassification.

The optimization problem which defines the support vector classifier has the following property:

Only observations that lie on the margin or that violate the margin will affect the hyperplane!

Likewise, an observation that lies strictly on the correct side of the margin **does not** affect the support vector classifier.

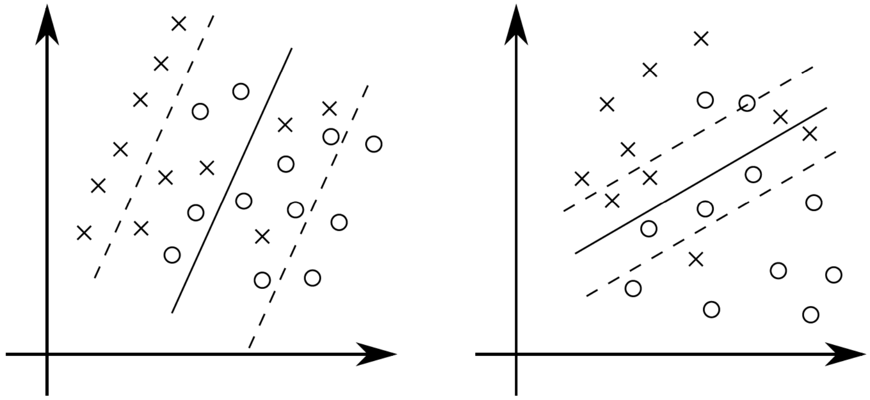

FIGURE 3. Left figure: large C. Right figure: small C

The observations that lie directly on the margin, or on the wrong side of the margin in respect to their class, are called *support vectors*.

Remark 2. The support vectors *do affect* the hyperplane.