SEPARATING HYPERPLANES

Vapnik's Maxim

There is a mantra in machine learning attributed to Vladimir Vapnik, a pioneer of ML:

"Don't solve a harder problem than you have to"

Plug-in methods require estimation of (conditional) densities or mass functions, which can be more difficult than estimating a decision boundary

\[
\begin{align*}
&1 \\
&\frac{1}{2} \\
&0
\end{align*}
\]

\(\eta(x)\) is quite complicated but the decision region is simple and \(\eta\) is smooth near \(\frac{1}{2}\).

In these notes we'll look at a method for linear classification that estimates the classifier more directly.

Hyperplanes

A hyperplane is a subset of \(\mathbb{R}^d\) of the form

\[
H = \{x : \mathbf{w}^\top x + b = 0\}
\]
for some \( \mathbf{w} \in \mathbb{R}^d \), \( b \in \mathbb{R} \). When \( d = 3 \), we have this picture:

In general, a hyperplane is an affine subspace of dimension \( d - 1 \).

The vector \( \mathbf{w} \) is orthogonal to the hyperplane. If \( \mathbf{w} \) is a vector that lies parallel to the hyperplane, we can write \( \mathbf{w} = \mathbf{x} - \mathbf{x}' \) for two points \( \mathbf{x}, \mathbf{x}' \) on the hyperplane. Thus

\[
\mathbf{w}^T \mathbf{w} = \mathbf{w}^T (\mathbf{x} - \mathbf{x}') = -b - (-b) = 0.
\]
We call w a normal vector.

Given a hyperplane $H = \{ x : w^T x + b = 0 \}$ and a point $z \notin H$, what is the distance of $z$ to $H$?

We can write $z$ as

$$z = z_0 + r \cdot \frac{w}{\|w\|}$$

for unique $z_0 \in H$ and $r \in \mathbb{R}$ (note $r$ may be negative).
Then
\[ w^Tz + b = w^Tz_0 + w^T(r \frac{w}{\|w\|}) + b \]
\[ = r \cdot \|w\| \quad \begin{bmatrix} w^Tz_0 + b = 0 \end{bmatrix} \]

Hence
\[ |r| = \frac{|w^Tz + b|}{\|w\|} \]

**Separating Hyperplanes**

Let \((x_1, y_1), \ldots, (x_n, y_n)\) be training data for a binary classification problem, and assume \(y_i \in \{-1, 1\}\.

We say the training data are **linearly separable** if there exist \(w \in \mathbb{R}^d\), \(b \in \mathbb{R}\) such that

\[ y_i (w^T x_i + b) > 0 \quad \forall i = 1, \ldots, n. \]

In this case we refer to
\[ H = \{x : w^T x + b = 0\} \]

as a **separating hyperplane**.

Are all separating hyperplanes equally good?
The Maximum Margin Hyperplane

The margin $p$ of a separating hyperplane is the distance from the hyperplane to the nearest training point:

$$p(w, b) := \min_{i=1, \ldots, n} \frac{|w^T x_i + b|}{\lVert w \rVert}$$

The maximum margin or optimal separating hyperplane is the solution of

$$\max_{w, b} p(w, b)$$

s.t. $y_i (w^T x_i + b) > 0 \quad \forall i$
A separating hyperplane is said to be in canonical form if \( w \) and \( b \) are such that

\[
y_i (w^T x_i + b) = 1 \quad \forall i
\]

\[
y_i (w^T x_i + b) = 1 \quad \text{for some } i
\]

Every separating hyperplane can be expressed in canonical form. Suppose \( H = \{ x : w_i^T x + b_i = 0 \} \) is a separating hyperplane (not necessarily in canonical form). Let

\[
m := \min_{i=1, \ldots, n} |w_i^T x_i + b_i|
\]

and define

\[
w_2 = \frac{w_i}{m} \quad \text{and} \quad b_2 = \frac{b_i}{m}.
\]
\[ w_2 = \frac{w_1}{w}, \quad b_2 = -\frac{b}{w}. \]

Then \( w_2, b_2 \) express \( H \) in canonical form.

This allows us to write the max-margin separating hyperplane as the solution of

\[
\max_{w,b} \frac{|w^T x_i + b|}{\|w\|} \\
\text{s.t.} \quad y_i (w^T x_i + b) \geq 1 \quad \forall i \\
y_i (w^T x_i + b) = 1 \quad \text{for some } i.
\]

How can this be simplified?

\[
\max_{w,b} \frac{1}{\|w\|} \\
\text{s.t.} \quad y_i (w^T x_i + b) \geq 1 \quad \forall i \\
y_i (w^T x_i + b) = 1 \quad \text{for some } i.
\]

Can we simplify further? Yes, we can drop the second constraint because it will automatically be satisfied by the optimizer. So we finally have

\[
\max_{w,b} \frac{1}{\|w\|} \\
\text{s.t.} \quad y_i (w^T x_i + b) \geq 1 \quad \forall i
\]
\[ \min_{w, b} \frac{1}{2} \|w\|^2 \]
\[ \text{s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall i = 1, \ldots, n \]

This is an example of a constrained optimization problem, and in particular it is called a quadratic program (quadratic objective, linear constraints).

**Optimal Soft-Margin Hyperplane**

To accommodate nonseparable data, we modify the above QP by introducing slack variables \( \xi_1, \ldots, \xi_n \geq 0 \). This leads to the so-called optimal soft-margin hyperplane, where
\[ \xi = (\xi_1, \ldots, \xi_n)^T \]

\[ \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i \]
\[ \text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \]
\[ \xi_i \geq 0, \quad \forall i \]
$3i \geq 0, \quad \forall c$

**Remarks**

- This is another QP
- $C > 0$ is a user-specified tuning parameter. Later we’ll discuss how to select it.
- If $x_i$ is misclassified, then $\xi_i \geq 1$. Therefore
  \[
  \frac{1}{n} \sum \xi_i \geq \frac{1}{n} \sum 1_{y_i \neq \text{sign}(w^T x_i + b)} \\
  = \text{training error}
  \]
- $C$ controls the influence of outliers

The optimal soft-margin hyperplane is a special case of the more general support vector machine
which we'll study later.

- We'll discuss algorithms to solve the QP in the future.