Recall:

- A hyperplane \( H \) in \( V = \mathbb{R}^n \) is a subspace of \( V \) of dimension \( n - 1 \) (i.e., a subspace of codimension 1).
- Each hyperplane is determined by a nonzero vector \( \beta \in \mathbb{R}^n \) via
  \[
  H = \{ x \in \mathbb{R}^n : \beta^T x = 0 \} = \text{span}(\beta)^\perp.
  \]
Recall:

- A hyperplane $H$ in $V = \mathbb{R}^n$ is a subspace of $V$ of dimension $n - 1$ (i.e., a subspace of codimension 1).
- Each hyperplane is determined by a nonzero vector $\beta \in \mathbb{R}^n$ via
  \[ H = \{ x \in \mathbb{R}^n : \beta^T x = 0 \} = \text{span}(\beta)^\perp. \]

- An affine hyperplane $H$ in $\mathbb{R}^n$ is a subset of the form
  \[ H = \{ x \in \mathbb{R}^n : \beta_0 + \beta^T x = 0 \} \]

where $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^n$. 
Recall:

- A hyperplane $H$ in $V = \mathbb{R}^n$ is a subspace of $V$ of dimension $n - 1$ (i.e., a subspace of codimension 1).
- Each hyperplane is determined by a nonzero vector $\beta \in \mathbb{R}^n$ via
  \[ H = \{ x \in \mathbb{R}^n : \beta^T x = 0 \} = \text{span}(\beta)^\perp. \]
- An affine hyperplane $H$ in $\mathbb{R}^n$ is a subset of the form
  \[ H = \{ x \in \mathbb{R}^n : \beta_0 + \beta^T x = 0 \} \]
  where $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^n$.

We often use the term “hyperplane” for “affine hyperplane”.

\[ \text{Diagram showing a hyperplane in } \mathbb{R}^3 \text{ with the normal vector } \beta \text{ and point } x_0. \]
Hyperplanes (cont.)

Let

\[ H = \{ x \in \mathbb{R}^n : \beta_0 + \beta^T x = 0 \} . \]
Let

\[ H = \{ x \in \mathbb{R}^n : \beta_0 + \beta^T x = 0 \}. \]

Note that for \( x_0, x_1 \in H \),

\[ \beta^T (x_0 - x_1) = 0. \]

Thus \( \beta \) is perpendicular to \( H \). It follows that for \( x \in \mathbb{R}^n \),

\[ d(x, H) = \frac{\beta^T (x - x_0)}{\| \beta \|} = \frac{\beta_0 + \beta^T x}{\| \beta \|}. \]
Separating hyperplane

Suppose we have binary data with labels \{+1, -1\}. We want to separate data using an (affine) hyperplane.

ESL, Figure 4.14. (Orange = least-squares)
Suppose we have binary data with labels \( \{+1, -1\} \). We want to separate data using an (affine) hyperplane.

Classify using \( G(x) = \text{sgn}(x^T \beta + \beta_0) \).
Separating hyperplane

Suppose we have binary data with labels \{+1, −1\}. We want to separate data using an (affine) hyperplane.

Classify using \( G(x) = \text{sgn}(x^T \beta + \beta_0) \).

- Separating hyperplane may not be unique.
- Separating hyperplane may not exist (i.e., data may not be separable).
Uniqueness problem: when the data is separable, choose the hyperplane to maximize the “margin” (the “no man’s land”).
**Uniqueness problem:** when the data is separable, choose the hyperplane to maximize the “margin” (the “no man’s land”).

Data: \((y_i, x_i) \in \{+1, -1\} \times \mathbb{R}^p\) \((i = 1, \ldots, n)\).

Suppose \(\beta_0 + \beta^T x\) is a separating hyperplane with \(\|\beta\| = 1\).
Uniqueness problem: when the data is separable, choose the hyperplane to maximize the “margin” (the “no man’s land”).

Data: \((y_i, x_i) \in \{+1, -1\} \times \mathbb{R}^p \quad (i = 1, \ldots, n)\).

Suppose \(\beta_0 + \beta^T x\) is a separating hyperplane with \(\|\beta\| = 1\).

Note that:

\[
y_i(x_i^T \beta + \beta_0) > 0 \implies \text{Correct classification}
\]

\[
y_i(x_i^T \beta + \beta_0) < 0 \implies \text{Incorrect classification}
\]
**Uniqueness problem:** when the data is separable, choose the hyperplane to maximize the “margin” (the “no man’s land”).

Data: \((y_i, x_i) \in \{+1, -1\} \times \mathbb{R}^p\) \((i = 1, \ldots, n)\).

Suppose \(\beta_0 + \beta^T x\) is a separating hyperplane with \(\|\beta\| = 1\).

Note that:

\[ y_i(x_i^T \beta + \beta_0) > 0 \Rightarrow \text{Correct classification} \]
\[ y_i(x_i^T \beta + \beta_0) < 0 \Rightarrow \text{Incorrect classification} \]

Also, \(\|y_i(x_i^T \beta + \beta_0)\| = \text{distance between } x \text{ and hyperplane (since } \|\beta\| = 1\)\).
Thus, if the data is separable, we can solve

$$\max_{\beta_0, \beta \in \mathbb{R}^p, \|\beta\| = 1} M$$

subject to $$y_i (x_i^T \beta + \beta_0) \geq M \quad (i = 1, \ldots, n).$$
Thus, if the data is separable, we can solve

\[
\max_{\beta_0, \beta \in \mathbb{R}^p, \|\beta\|=1} M
\]

subject to \( y_i (x_i^T \beta + \beta_0) \geq M \quad (i = 1, \ldots, n) \).

We will transform the problem into a usual form used in convex optimization.
Thus, if the data is separable, we can solve

$$\max_{\beta_0, \beta \in \mathbb{R}^p, \|\beta\|=1} M$$

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

We will transform the problem into a usual form used in convex optimization.

We can remove $\|\beta\| = 1$ by replacing the constraint by

$$\frac{1}{\|\beta\|} y_i(x_i^T \beta + \beta_0) \geq M,$$

or equivalently,

$$y_i(x_i^T \beta + \beta_0) \geq M \|\beta\|.$$
Thus, if the data is separable, we can solve

$$\max_{\beta_0, \beta \in \mathbb{R}^p, ||\beta|| = 1} M$$

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

We will transform the problem into a usual form used in convex optimization.

We can remove $||\beta|| = 1$ by replacing the constraint by

$$\frac{1}{||\beta||} y_i(x_i^T \beta + \beta_0) \geq M, \quad \text{or equivalently,} \quad y_i(x_i^T \beta + \beta_0) \geq M ||\beta||.$$

We can always rescale $(\beta, \beta_0)$ so that $||\beta|| = 1/M$. Our problem is therefore equivalent to

$$\min_{\beta_0, \beta \in \mathbb{R}^p} \frac{1}{2} ||\beta||^2$$

subject to $y_i(x_i^T \beta + \beta_0) \geq 1 \quad (i = 1, \ldots, n)$. 
Thus, if the data is separable, we can solve

$$\max_{\beta_0, \beta \in \mathbb{R}^p, \|\beta\|=1} \quad M$$

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

We will transform the problem into a usual form used in convex optimization.

We can remove $\|\beta\| = 1$ by replacing the constraint by

$$\frac{1}{\|\beta\|} y_i(x_i^T \beta + \beta_0) \geq M,$$

or equivalently,

$$y_i(x_i^T \beta + \beta_0) \geq M \|\beta\|.$$ 

We can always rescale $(\beta, \beta_0)$ so that $\|\beta\| = 1/M$. Our problem is therefore equivalent to

$$\min_{\beta_0, \beta \in \mathbb{R}^p} \quad \frac{1}{2} \|\beta\|^2$$

subject to $y_i(x_i^T \beta + \beta_0) \geq 1$ \quad (i = 1, \ldots, n).

We now recognize the problem as a convex optimization problem with a quadratic objective, and linear inequality constraints.
Support vector machines

- The previous problem works well when the data is *separable*. What happens if there is no way to find a margin?
Support vector machines

- The previous problem works well when the data is *separable*. What happens if there is no way to find a margin?
- We allow some points to be on the wrong side of the margin, but keep control on the error.
Support vector machines

- The previous problem works well when the data is *separable*. What happens if there is no way to find a margin?
- We allow some points to be on the wrong side of the margin, but keep control on the error. We replace \( y_i(x_i^T \beta + \beta_0) \geq M \) by
  \[
y_i(x_i^T \beta + \beta_0) \geq M(1 - \xi_i), \quad \xi_i \geq 0,
\]
and add the constraint
  \[
  \sum_{i=1}^{n} \xi_i \leq C \quad \text{for some fixed constant } C > 0.
  \]
The previous problem works well when the data is *separable*. What happens if there is no way to find a margin? We allow some points to be on the wrong side of the margin, but keep control on the error. We replace $y_i (x_i^T \beta + \beta_0) \geq M$ by

$$y_i (x_i^T \beta + \beta_0) \geq M(1 - \xi_i), \quad \xi_i \geq 0,$$

and add the constraint

$$\sum_{i=1}^{n} \xi_i \leq C \quad \text{for some fixed constant } C > 0.$$

The problem becomes:

$$\max_{\beta_0, \beta \in \mathbb{R}^p, \|\beta\| = 1} \quad M$$

subject to $y_i (x_i^T \beta + \beta_0) \geq M(1 - \xi_i)$

$$\xi_i \geq 0, \quad \sum_{i=1}^{n} \xi_i \leq C.$$
As before, we can transform the problem into its "normal" form:

\[
\begin{align*}
&\min_{\beta_0, \beta} \frac{1}{2} \|\beta\|^2 \\
&\text{subject to } y_i (x_i^T \beta + \beta_0) \geq 1 - \xi_i \\
&\xi_i \geq 0, \quad \sum_{i=1}^{n} \xi_i \leq C.
\end{align*}
\]

Problem can be solved using standard optimization packages.