SEPARATING
HYPERPLANES

LDA and logistic regression are "plug-in" methods for linear classification. They make assumptions about the distribution of the data, and reduce classification to estimation.

In these notes we'll discuss an approach to linear classification that

1. makes no distributional assumptions
2. does not require solving an intermediate (and potentially more difficult) problem.
Let \((x_1, y_1), \ldots, (x_n, y_n)\) be training data, \(x_i \in \mathbb{R}^d\), \(y_i \in \{-1, +1\}\)

**Definition**: we say the data are **linearly separable** if there exists \(w \in \mathbb{R}^d\), \(b \in \mathbb{R}\) such that

\[
y_i = \text{sign} \left\{ \mathbf{w}^T \mathbf{x}_i + b \right\}
\]

for \(i = 1, \ldots, n\). We refer to

\[
\{ \mathbf{x} : \mathbf{w}^T \mathbf{x} + b = 0 \}
\]

as a **boundary**.
Assume for now that the data are linearly separable. How can we find a separating hyperplane?

**Geometry**

Let $w, b$ define a hyperplane.

If $x, x'$ are points on the hyperplane, then

$$0 = (w^T x + b) - (w^T x' + b)$$

Hence $w$ is normal to all vectors that are parallel to the hyperplane.

\[
\text{d} = 3
\]
We call \( \frac{w}{\|w\|} \) the vector to the hyperplane. It is unique up to its ___.

**Question** Let \( z \in \mathbb{R}^d \). How far is \( z \) from \( \exists x \in \mathbb{R}^d : w^T x + b = 0 \) ?

**Answer** Write
\[
z = z_0 + r \cdot \frac{w}{\|w\|}
\]
where \( w^T z_0 + b = 0 \) and \( r \) may be negative.

Then
\[
w^T z + b =
\]
\[
= 
\]
\[
= 
\]
\[
\implies
\]
We refer to $r$ as the "signed distance" from $z$ to the hyperplane.

**Rosenblatt's Perceptron**

The perceptron learning algorithm seeks $w, b$ to minimize the total distance of misclassified points to the decision boundary.

How can we formulate this criterion mathematically?

Recall that $x_i$ is misclassified iff

$$y_i (w^T x_i + b) < 0.$$  

Let $I(w, b)$ be the indices $i$ such that $y_i (w^T x_i + b) < 0$. 

Then the total (unsigned) distance of the misclassified points to the hyperplane is

\[ \alpha =: D(\omega, b) \]

The perceptron learning algorithm attempts to minimize \( D(\omega, b) \) using __________.

The gradient of \( D \) is given by:

\[ \frac{\partial D}{\partial \omega} = \]

\[ \frac{\partial D}{\partial b} = \]
Instead of stepping in the negative gradient direction, we cycle through the data points and perform

\[ \text{If } i \in I(w, b) \]

\[ W \leftarrow W + \gamma y_i x_i \]
\[ b \leftarrow b + \gamma y_i \]

\text{End}

---

contribution to gradient from \( i \)th term

Here \( \gamma > 0 \) is the learning rate. Since \( w, b \) can be rescaled without changing the classifier, we may take \( \gamma = 1 \).
+ If the data are linearly separable, then a separating hyperplane is found after a finite number of steps.

- This finite number can be very large, depending on the gap between classes.

- The final solution depends on the

+ If the data are not linearly separable, the algorithm will never converge.

+ The perceptron algorithm adapts naturally to an online setting, and is the basis of many strategies for online learning.
The Maximum Margin Hyperplane

Rosenblatt's perceptron algorithm will find a separating hyperplane when one exists, but it does not prefer one separating hyperplane over another.

Are all separating hyperplanes equally good?
Definitions

1. The margin $\rho$ of a separating hyperplane is the distance from the hyperplane to the closest $x_i$.

2. The maximum margin or optimal separating hyperplane is the solution of

$$ (w^*, b^*) = \arg \max_{w, b} \, \rho(w, b) $$
Canonical Form

We may rescale any separating hyperplane so that it is in:

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

Exercise: Express the margin of a hyperplane in canonical form as a function of \( w \) and \( b \).
Express \( w^*, b^* \) as the solution of a constrained optimization problem.
A. density / function
B. separating hyperplane
C. \( w^T(x - x') \), orthogonal, parallel
D. normal, sign
E. \( w^Tz + b = w^T(z_0 + r \frac{w}{\|w\|}) + b \)
   \[ = w^Tz_0 + b + r \frac{w^Tw}{\|w\|} \]
   \[ = r \cdot \|w\| \]
   \[ \Rightarrow |r| = \frac{|w^Tz + b|}{\|w\|} \]
F. \[ \sum_{i \in I(w, b)} -y_i \frac{(w^T x_i + b)}{\|w\|} \leq \sum_{i \in I(w, b)} -y_i \frac{(w^T x_i + b)}{\|w\|} \]
   sequential gradient descent
\[ \frac{\partial D}{\partial w} = -\sum_{i \in I(w, b)} y_i x_i \]
\[ \frac{\partial D}{\partial b} = -\sum_{i \in I(w, b)} y_i \]
G. initialization
H. \( \rho(w, b) = \min_{i=1, \ldots, n} \frac{|w^T x_i + b|}{\|w\|} \)
I. canonical form
Solution

\[ p(w,b) = \min_{i=1,...,n} \frac{|w^T x_i + b|}{||w||} = \frac{1}{||w||} \]

The optimal separating hyperplane is therefore the solution of

\[ \min_{w,b} \frac{1}{2} ||w||^2 \]

s.t. \( y_i (w^T x_i + b) \geq 1 \), \( i = 1, ..., n \)

Terminology

\( \bigstar \) is an example of a ________

\( \bigstar \) Those \( x_i \) such that \( y_i (w^T x_i + b) = 1 \) are called ________.  

Optimal Soft-Margin Hyperplane

Real data is often not linearly separable. To accommodate nonseparable data, we modify the QP by introducing $\xi_1, \ldots, \xi_n \geq 0$.

This results in the optimal soft-margin hyperplane:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t. \( y_i (w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \)

$\xi_i \geq 0 \quad i = 1, \ldots, n.$

Remarks:

- This is another QP
- If $x_i$ is misclassified, then

Therefore

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \geq$$
C is a cost-complexity tradeoff parameter. It should be set using error estimation.

C also controls the influence of .

What happens when

\[ C \to 0 \]
\[ C \to \infty \]
J. quadratic program, support vectors
K. slack variables
L. \( x_i \) misclassified \( \Rightarrow \delta_i > 1 \)
\[
\frac{1}{n} \sum \delta_i \geq \text{training error}
\]
M. outliers