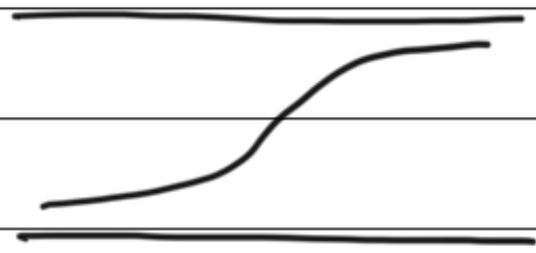
Consider logistic regression

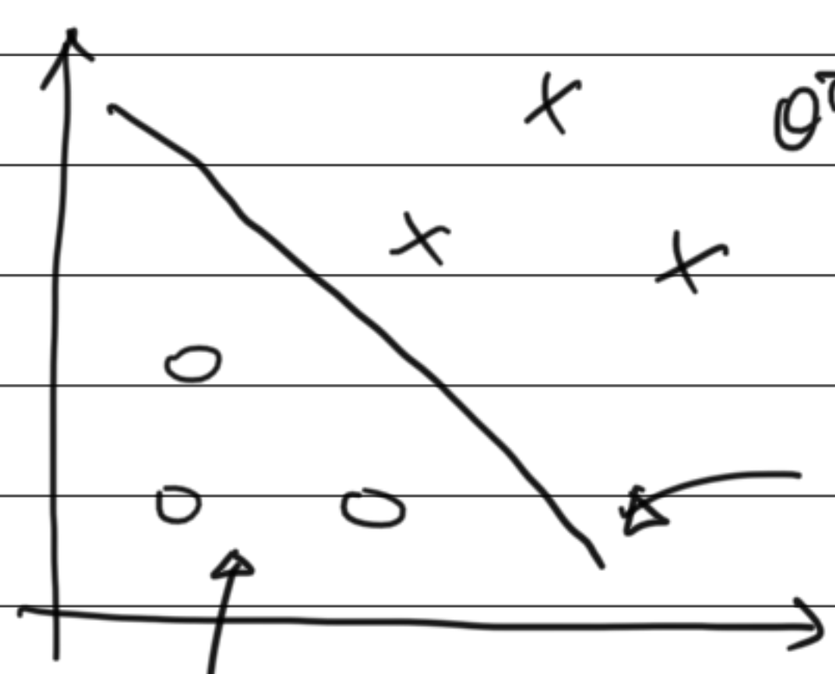
$P(y=1 \mid x;\theta)$ is modeled by $h_\theta(x) = g(\theta^T x)$



if $\theta^T x \gg 0$, $y=1$
if $\theta^T x \ll 0$, $y=0$
] very confident prediction

Given a training set.
if we can find θ so that $\theta^T x^{(i)} \gg 0$
whenever $y^{(i)} = 1$
$\theta^T x^{(i)} \ll 0$ whenever $y^{(i)} = 0$

⇒ This seems to be a nice goal to aim for, and we'll soon formalize this idea using the notion of functional margin



$\theta^T x > 0$: positive training example

decision boundary ($\theta^T x = 0$)
= separating hyperplane

$\theta^T x < 0$: negative training example

* If a point is far from the separating hyperplane, we may be significantly more confident in our predictions

- Notation

We will be considering a linear classifier for a boundary classification problem with labels y and features x

$$y \in \{+1, -1\}$$

$$h_\theta(x) = g(\theta^T x) \longrightarrow h_{w,b}(x) = g(w^T x + b)$$

b takes the role of what was previously θ_0, and w takes the role of $[\theta_1 \cdots \theta_n]^T$

- Functional margin

Given a training example $(x^{(i)}, y^{(i)})$.
We define the functional margin of (w,b)

$$\Rightarrow \quad \hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$

if $y^{(i)} = 1$, then for the functional margin to be large (very confident), we need $w^T x + b$ to be a large positive number.

if $y^{(i)} = -1$, we need $w^T x + b$ to be a large negative number.

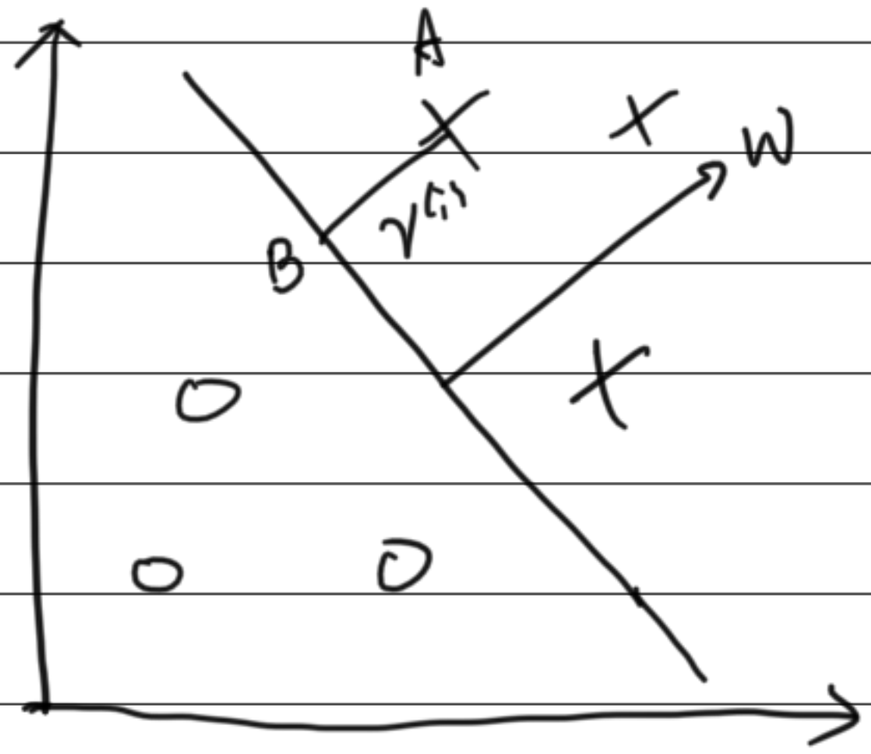
Moreover, if $y^{(i)}(w^T x + b) > 0$, then our prediction on this example is correct.

Given a training set $S = \{(x^{(i)}, y^{(i)}) ; i = 1, \dots, m\}$,
We define the functional margin of (w, b) with respect to S to be the smallest of the functional margins of the individual training examples.

$$\hat{\gamma} = \min_{i=1,\dots,m} \hat{\gamma}^{(i)}$$

* Replacing (w, b) with $(2w, 2b)$ results in multiplying our functional margin by a factor of 2.
⇒ It seems that by exploiting freedom to scale w and b. We can make the functional margin arbitrarily large without really changing anything meaningful.

- Geometric margin (Generalization of the functional margin)



The decision boundary corresponding (w, b) is shown, along with the vector w.
(Note that w is orthogonal to the separating hyperplane)

Consider the point A, which represents the input $x^{(i)}$ of some training example with label $y^{(i)} = 1$.
Its distance to the decision boundary, $\gamma^{(i)}$ is given by the line segment AB.

How can we find $\gamma^{(i)}$?

- $w/\|w\|$ is a unit-length vector pointing in the same direction as w
- A represents $x^{(i)}$
- B is given by $x^{(i)} - \gamma^{(i)} \cdot w/\|w\|$

This point lies on the decision boundary

All points x on the decision boundary satisfy $w^T x + b = 0$.

$$w^T\left(x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|}\right) + b = 0$$

$$w^T x^{(i)} + b = \gamma^{(i)} \underbrace{\frac{w^T w}{\|w\|}}$$

$$w^T x^{(i)} + b = \gamma^{(i)} \|w\|$$

$$\Rightarrow \gamma^{(i)} = \frac{w^T x^{(i)} + b}{\|w\|} = \left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}$$

We define the geometric margin of (w, b) with respect to a training example $(x^{(i)}, y^{(i)})$ to be

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}\right)$$

Note that, if $\|w\| = 1$, the functional margin equals the geometric margin

* Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \dots, n\}$, the geometric margin of (w, b) with respect to S to be the smallest of the geometric margins on the individual training examples.

$$\gamma = \min_{i=1,\dots,n} \gamma^{(i)}$$

- The optimal margin classifier
  : Find a decision boundary that maximize the geometric margin

$$\begin{cases} \max_{\gamma, w, b} \ \gamma \\ \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \dots, m \\ \|w\| = 1 \leftarrow \text{non-convex} \end{cases}$$

$\Rightarrow$ Transform the problem into a nicer one

$$\begin{cases} \max_{\hat{\gamma}, w, b} \ \dfrac{\hat{\gamma}}{\|w\|} \leftarrow \text{non-convex} \\ \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \dots, m \end{cases}$$

We can add an arbitrary scaling constraint on w and b without changing anything.

$\Rightarrow$ We will introduce the scaling constraint that the functional margin of w.b with respect to the training set must be 1:

$$\hat{\gamma} = 1$$

$$\begin{cases} \max_{w, b} \ \dfrac{1}{\|w\|} \\ \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 \end{cases}$$

Maximizing $\frac{\hat{\gamma}}{\|w\|} = \frac{1}{\|w\|}$ is the same thing as

minimizing $\|w\|^2$

$$\begin{cases} \min_{w,b} \|w\|^2 \\ \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 \end{cases}$$

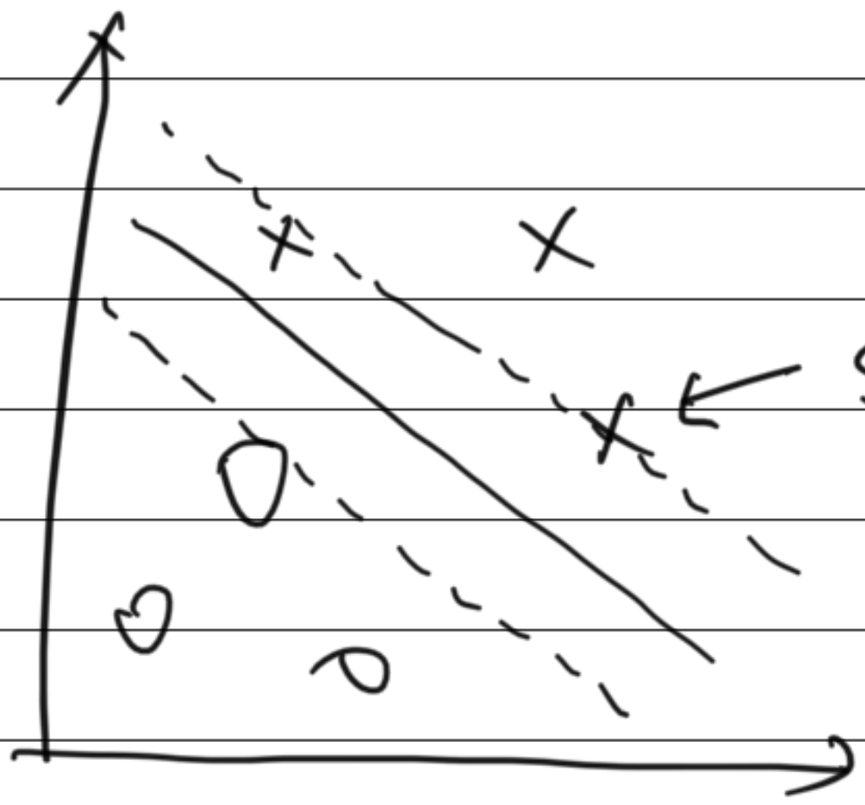
⇓

Nicer form:

$$\begin{cases} \min_{w,b} \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \cdots, m \end{cases}$$

This optimization problem can be solved using quadratic programming (QP).

↳ optimal margin classifier (SVM)



* SVM: Support vector 들로부터 거리(margin)은 최대화하여 도출.