## Classification:

Does $\vec{x}$ belong to class 1 or class 2?

| | | |
|---|---|---|
| patient | sick | well |
| email | spam | not spam |
| image | cat | dog |

Algorithmic decision making. Fraught w/ consequences
loans, recidivism, etc.

How to algorithmically classify? Most commonly find a function equal to +1 on class 1
-1 on class 2.

Example: If $\vec{w}^T \vec{x} \geq$ threshold:
  return 1
else:
  return -1

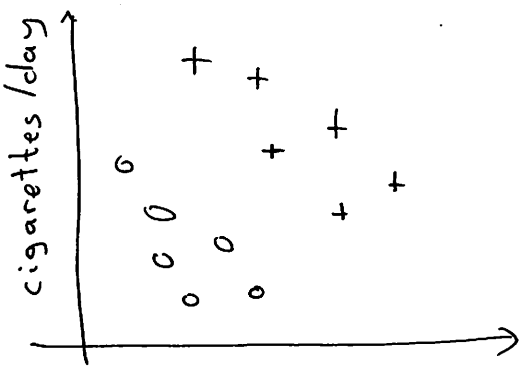
This is Linear Classification

# Linear classification

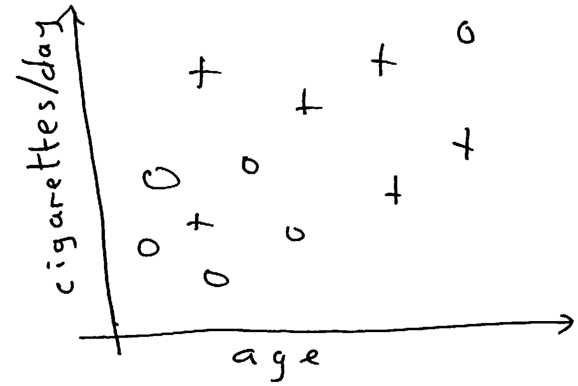
Assume $d=2$

$+$ class 1 (cancer)

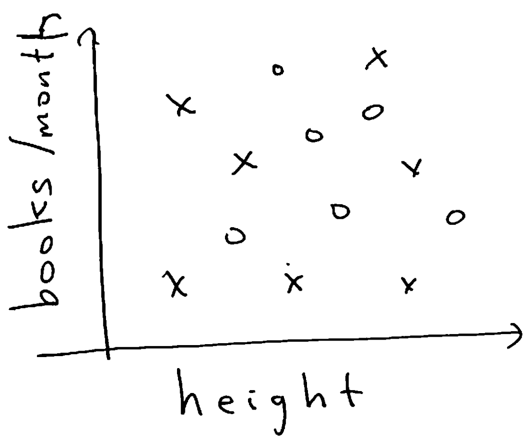
$\circ$ class 2 (not cancer)




$\boxed{\text{ideal}}$




$\boxed{\text{reality}}$




Use lines to separate data

use ~~???~~ signed distance to make predictions

**Training:** use data to fit the line (how?)

## Decision Boundary: Consider the classification rule

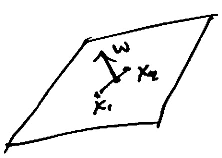
$$f(x) = \vec{w}^T \vec{x} + b \qquad \text{prediction}(x) = \begin{cases} 1 & f(x) \geq 0 \\ -1 & f(x) < 0 \end{cases}$$

This rule separates the space in half, and the boundary is defined by a hyperplane

Boundary $\mathcal{B} = \{ \vec{x} : f(\vec{x}) = 0 \} = \{ \vec{x} : \vec{w}^T \vec{x} + b = 0 \}$

The boundary is normal to $\vec{w}$: If $\vec{x}_1, \vec{x}_2 \in \mathcal{B}$, $\vec{w}^T(\vec{x}_1 - \vec{x}_2) = (\vec{w}^T \vec{x}_1 + b) - (\vec{w}^T \vec{x}_2 + b) = 0$.

normal means $\vec{w}$ is orthogonal to rays in $\mathcal{B}$

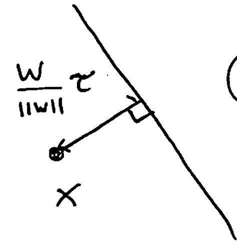



b is the bias. It sets the threshold

$$\text{prediction}(x) = \begin{cases} \text{class } 1 & \text{if } \vec{w}^T x \geq -b \\ \text{class } 2 & \text{if } \vec{w}^T x < -b \end{cases}$$

Distance from boundary to $\vec{x}$ is

$$\frac{|\vec{w}^T \vec{x} + b|}{\|w\|}$$




$\vec{w}$ normal to

PF:

$$\vec{w}^T \left( \vec{x} + \tau \frac{w}{\|w\|} \right) + b = 0 \quad \text{if}$$

$$\tau = \frac{-(w^T x + b)}{\|w\|} \qquad ///$$

Let $\vec{x}_1, \ldots \vec{x}_n$ be a seperable data set
Note if $y_i = 1$ for class 1, $-1$ for class 2

$$\frac{|\vec{w}^T \vec{x}_i + b|}{\|w\|} = \frac{y_i(\vec{w}^T \vec{x}_i + b)}{\|w\|}$$

(seperable means there is a $(\vec{w}, b)$ s.t. $\vec{w}^T x_i + b \geq 0$ for $x_i$ in class 1, $\vec{w}^T x_i + b < 0$ in class 2).

Margin: $\min\limits_i \dfrac{y_i(\vec{w}^T \vec{x}_i + b)}{\|w\|}$

Large margin $\Rightarrow$ data far from boundary $\Rightarrow$ small changes in data do not change the decisions.

Can we find a hyperplane with large margin?

If data is linearly separable

$$\underset{\vec{w}}{\text{maximize}} \quad \min_i \frac{y_i(\vec{w}^T\vec{x}_i + b)}{\|\vec{w}\|}$$

$$\underset{\vec{w}}{\text{maximize}} \quad \frac{\left\{\min_i \; y_i(\vec{w}^T\vec{x}_i + b)\right\}}{\|\vec{w}\|}$$

~~maximize~~

$$\underset{\vec{w}}{\text{minimize}} \quad \frac{\|\vec{w}\|}{\left\{\min_i \; y_i(\vec{w}^T\vec{x}_i + b)\right\}}$$

$$\boxed{\begin{aligned} &\underset{\vec{w}}{\text{minimize}} \quad \|\vec{w}\|^2 \\ &\text{subject to} \quad y_i(\vec{w}^T\vec{x}_i + b) \geq 1 \end{aligned}}$$

Max-margin (can be solved w/ quadratic programming solver.)

If data is not linearly separable, penalize for errors:

$$\text{minimize} \quad \|\vec{w}\|^2 + C\sum_{i=1}^{n} \epsilon_i$$

$$\text{subject to} \quad y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \epsilon_i$$

$$\epsilon_i \geq 0$$

$C$ is a regularization parameter.

This is the <u>support</u> <u>vector</u> <u>machine</u>.

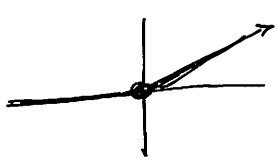
Equivalent form:

$$\boxed{\text{minimize} \quad C\sum_{i=1}^{n} \max\left(1 - y_i(\vec{w}^T\vec{x}_i + b), 0\right) + \|\vec{w}\|^2}$$

(to see this, if $y_i(\vec{w}^T\vec{x}_i + b) \geq 1$, $\epsilon_i = 0$.
if $y_i(\vec{w}^T\vec{x}_i + b) > 1$, $\epsilon_i = 1 - y_i(\vec{w}^T\vec{x}_i + b)$)

---

$$h(z) = \max(z, 0) \qquad h'(z) = \begin{cases} 1 & z > 0 \\ 0 & \text{o.w.} \end{cases}$$



$$\nabla_{\vec{w}}\left(\max(1 - y_i(\vec{w}^T\vec{x}_i + b)\right) = \begin{cases} -y_i\vec{x}_i & \text{if } \epsilon_i > 0 \\ 0 & \text{o.w.} \end{cases}$$

SVM cost is <u>convex</u> (why?)