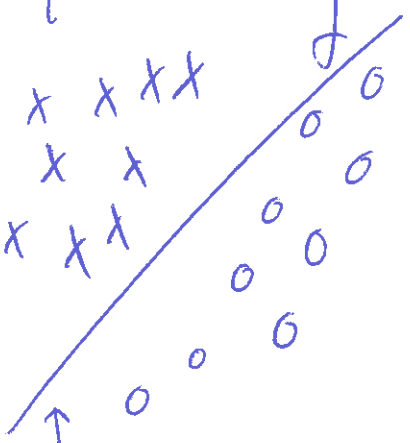
Lecture # 14

- KNN classifier basically does no training... we only use the data when it is time to label, and even then we only look locally.

- An alternative (and a preferable one, generally), is to use all the data ahead of time to first do some work (training), that will make the subsequent labelling of points easier.

- Let's try to "split" data by learning a hyperplane that separates two classes:



↑ Can we efficiently learn this, or a similar hyperplane?

Set-up: Let $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i \in \{-1, 1\}$ (classes)
$x_i \in \mathbb{R}^d$ (data points)

- A hyperplane is a set of $x \in \mathbb{R}^d$ satisfying the equation

$w^T x = -b$, some $w \in \mathbb{R}^d$, $b \in \mathbb{R}$.

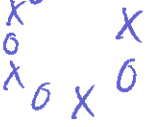
- In our sketch, w^T enforces the "direction" of the dividing plane and b is the "offset" from the origin, not unlike slope & y-intercept when discussing the equation of a line.

- We want $\{x_i\}$ to be "discriminated" by w^T, b, in the following sense:

$$w^T x_i + b = \begin{cases} +, & y_i = 1 \\ -, & y_i = -1 \end{cases}$$

- That is, points are on one side of the hyperplane if they have label 1, on the other side if they have label -1.

Remark: Not hard to construct datasets in which this is impossible, i.e. no separating hyperplanes:



Such examples can be dealt with using non-linear boundary fitting with kernel support vector machines.

- Linear algebra set-up: Let $v_i := (\underbrace{y_i x_i}_{\mathbb{R}^d}, \underbrace{y_i}_{\mathbb{R}^1}) \in \mathbb{R}^{d+1}$

(from C.B. slides)

$$u = (w, b) \in \mathbb{R}^{d+1}$$

$\mathbb{R}^d$ $\mathbb{R}^d$

Choose u such that $v_i \cdot u > 0$, $i = 1, \ldots, n$.

- This makes sense because $v_i \cdot u$

$$= (y_i x_i, y_i) \cdot (w, b)$$

$$= y_i (w^T x_i + b),$$

So want $y_i, w^T x_i + b$ to have the same sign.

- In essence, we have n inequalities that need to hold simultaneously. This is easily addressed using linear programming.

- A standard formulation to linear programming problems is:

minimize/maximize $c^T x$

s.t. $Ax \leq b$

or $\geq b$

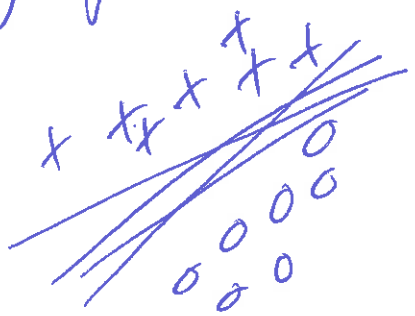
for some matrix A.

- Note that we don't even need to minimize/maximize here... we just need $y_i (w^T x_i + b)$

$= v_i \cdot u > 0$ for every i, which is analogous to finding u s.t. $\mathbb{R}^{n \times (d+1)} \left\{ \begin{pmatrix} - v_1 - \\ \vdots \\ - v_n - \end{pmatrix} u > (0, 0, \ldots 0) \right.$

- If we require u is bounded (for example, $u_i \in [-1,1]$, $i=1,\dots,d+1$), this can be solved in a variety of ways; take a course on optimization for details. MATLAB has "linprog" which is a useful blackbox.

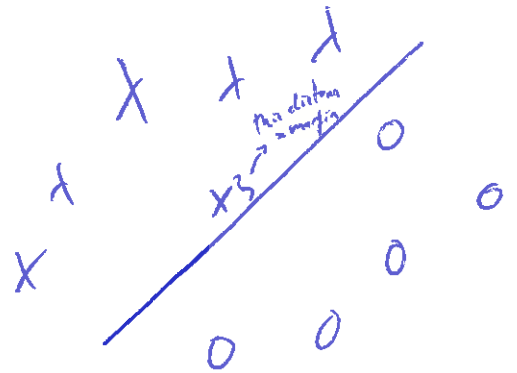
Remark: Except in very pathological cases in which some points are on the plane, if a separating hyperplane is slightly perturbed, it remains a separating hyperplane:



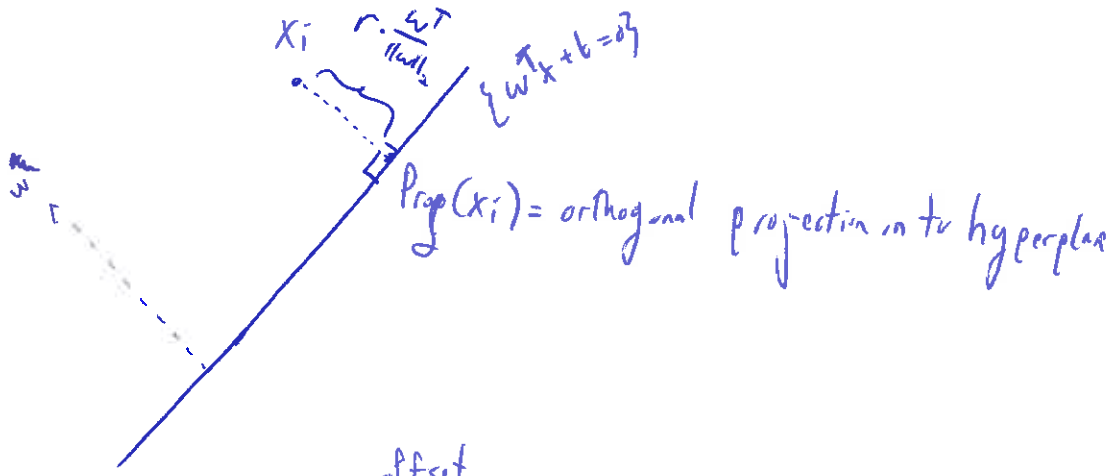
Which is "best"?

- Need an additional criterion to select "best" hyperplane. Pick the one that keeps points farthest away from the plane → "maximize the margin."

- The margin of a hyperplane $\{w^T x + b = 0\}$ separating points $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ is the minimum distance between the hyperplane and any point:



Q: How to compute the margin? Well, how to compute the distance between a point and a hyperplane?



Set $x_i = \underbrace{Proj(x_i)}_{\text{on hyperplane}} + \underbrace{r \cdot \frac{w}{\|w\|_2}}_{\text{offset}}$; r is the "signed distance" of x_i to the hyperplane.

Since $Proj(x_i)$ is on the hyperplane, we can compute:

$$w^T x_i + b = w^T \left(Proj(x_i) + r \frac{w}{\|w\|_2} \right) + b$$

$$= w^T Proj(x_i) + b + r \cdot \frac{w^T w}{\|w\|}$$

$w^T Proj(x_i) + b = 0$

$$= r \cdot \|w\|_2$$

$$\Rightarrow r = \frac{w^T x_i + b}{\|w\|_2}$$

- Note that if $w \mapsto \varepsilon w$
  $b \mapsto \varepsilon b$,

r stays the same. This gives us a choice in how to represent the hyperplane.

- Let $w^T x + b$ be a separating hyperplane. It is canonical if $|w^T x_i + b| = 1$ for all x_i achieving the margin; these $\{x_i\}$ are called support vectors.

- So, the problem of finding a separating hyperplane that maximizes the margin is the problem of

maximize margin $\Leftrightarrow$ minimize $\|w\|$

$$\text{minimize } \|w\|_2$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 \quad , i = 1, \ldots, n$$

(separating)

Next time: What to do in the (common) case there are no separating hyperplanes?