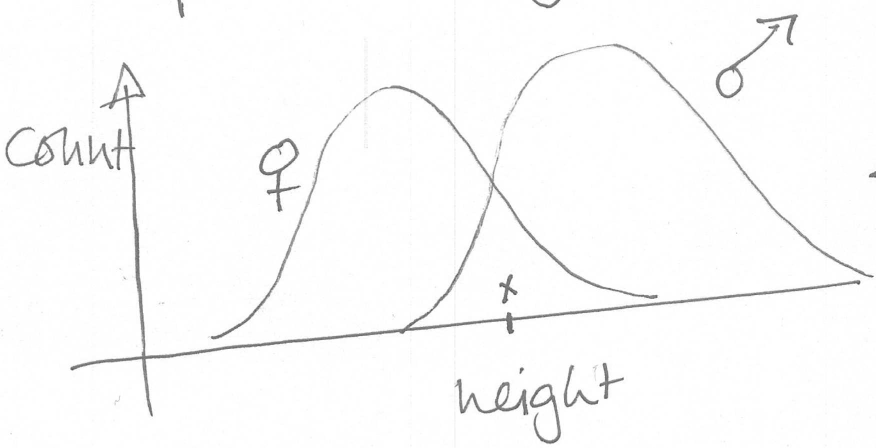
Classification:

- Input Features, output one bit
- (more complicated models later)

example:

input: height **output:** gender



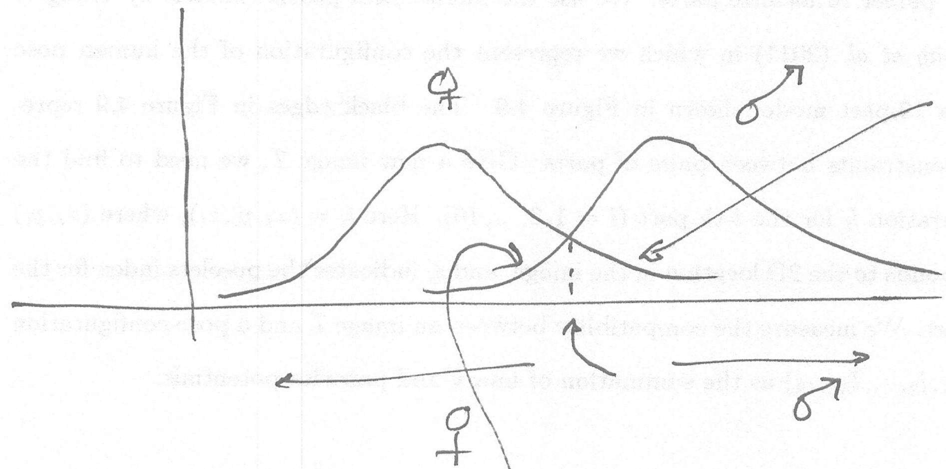
← histograms of height w/ gender

- For the moment, assume that ♀ ♂ are evenly dist.

- how would one classify?
  - choose a height threshold
  - mistakes are inevitable
  - we need to choose the least expensive.

(1a)

Notice guaranteed error



we will get these ♀'s wrong.

we will get these ♂'s wrong

Strategy

$h > t$ ♂
otherwise ♀

Cases:

- males, females equally common
  t at x
- males very common ♀'s rare
  $t \ll x$
- ♂ rare, ♀ common
  $t \gg x$

Q: reasonable way to set t

Choose t to produce the minimum expected-cost of errors

- two types of error
  (♂ → ♀)
  (♀ → ♂)

we assume reward for right answer is 0

③

we have a feature x.

$$\text{if we say } \female \text{, we get } \begin{cases} 0 & \text{right} \\ L(\male \to \female) & \text{wrong} \end{cases}$$

if we do this many times,
we get 0 with frequency

$p(\female|x)$

and $L(\male \to \female)$ with freq $p(\male|x)$

So expected loss of $\female$ is

$$0 \cdot p(\female|x) + L(\male \to \female)\, p(\male|x)$$

Similarly, expected loss of $\male$ is

$$0\, p(\male|x) + L(\female \to \male)\, p(\female|x)$$

so *in principle* we have a rule ④

at x,

cost of ♀ is : $L(♂ \to ♀)\, p(♀|x)$

cost of ♂ is : $L(♀ \to ♂)\, p(♂|x)$

. choose the least expensive, say that.

. But where do we get $p(♂|x)$, $p(♀|x)$?

1) $P(♀|x) = 1 - p(♂|x)$ (only 2 options)

2) ~~p(♀~~ $p(♂|x) = \dfrac{p(x|♂)\, p(♂)}{p(x)}$.

$$= \frac{p(x|♂)\, p(♂)}{\left[p(x|♂)\, p(♂) + p(x|♀)\, P(♀)\right]}$$

prior

we could read this off the histograms.

$[1 - p(♂)]$

We can now build one useful form
of classifier.

- measure $p(x|♀)$, $P(♀) = \pi$, $p(x|♂)$
  (say, histogram)

- say

  ♀ if $L(♀ \to ♂) . p(♀|x) > L(♂ \to ♀) . p(♂|x)$

  ♂ $<$

  doesn't matter $=$

now, consider

$$L(♀ \to ♂)\, p(♀|x) = L(♂ \to ♀)\, p(♂|x)$$

all that matters is $R = \dfrac{L(♀ \to ♂)}{L(♂ \to ♀)}$

So we care about

$$R \,.\, p(\text{♀}|x) = p(\text{♂}|x)$$

i.e.

$$\frac{R \,.\, p(x|\text{♀})\,\pi}{p(x)} = \frac{p(x|\text{♂})\,(1-\pi)}{p(x)}.$$

i.e

$$\underbrace{\frac{p(x|\text{♀})}{p(x|\text{♂})}}_{\text{likelihood ratio}} = \frac{(1-\pi)}{\pi} \cdot \frac{1}{R}.$$

i.e. if

$$\frac{p(x|\text{♀})}{p(x|\text{♂})} \begin{array}{l} > \; g(\pi, R) \quad \text{say ♀} \\ = \quad \text{doesn't matter} \\ < \quad \text{say ♂.} \end{array}$$

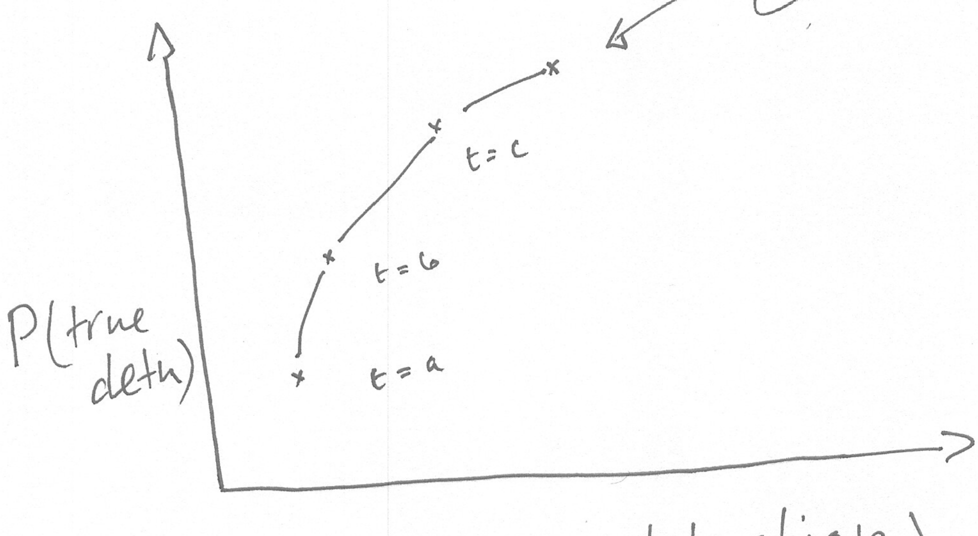
Now equations give a way to determine g. but we can manage without.

Rule: $\frac{p(x|♀)}{p(x|♂)} > t$, say ♀

Plot errors w/ varying t

Receiver operating curve.

plot for different t.



<u>Model:</u> we are trying to detect ♂'s in a population of ♀'s (or vice versa).

$$P(\text{false detect}) = \frac{\#\text{ of } ♀\text{'s we called } ♂\text{'s}}{\#\text{ of times we classified}}$$

$$P(\text{true det}) = \frac{\#\text{ of } ♂\text{'s we called } ♂\text{'s}}{\#\text{ of } ♂\text{'s in population}}$$

We evaluate this on <u>test</u> data.

<u>Training</u>:

- form histograms $p(x|♀)$ etc.
- can be hard to do with high dimensional x.

Major Problem:

- a 1-D histogram w/ n cells in each dir has n cells
- 2D ... n^2
- 3D ... n^3
- dD ... n^d

We cannot build such histograms:

Strategies:

- Simplify the model
- model $p(\omega|x)$ directly
- Search for Decision Boundary directly.

Ⓐ Simplify model

model

$$p(x_1, x_2 \cdots x_n \mid ♀)$$
$$= p(x_1 \mid ♀)\, p(x_2 \mid ♀)\, p(x_3 \mid ♀) \cdots p(x_n \mid ♀)$$

This model is usually wrong.

- But it's convenient
- and works surprisingly well

Naive Bayes:

method:
- one histogram each in each direction.
- form likelihood ratio
- test against threshold.

(B) Model $p(q|x)$ directly.

- parametric models, perhaps later
-

(C) <u>Find decision boundary</u>:

- By continuity reasoning (Nearest neighbours)
- By search.

<u>Nearest neighbours</u>:

<u>alg</u>:

- find $x_i \in$ examples such that $\|x_i - x\|^2$ is smallest
- class of x is class of x_i

k - l nearest neighbors:

alg: • find the k $x_n \in$ examples that are closest to x.

• find the most common class in these neighbors

• if there are l in this class, classify x with that class, otherwise, don't know.

Properties:

• with enough examples, error rate is no more than 2x best possible

• we must worry about scaling dimension

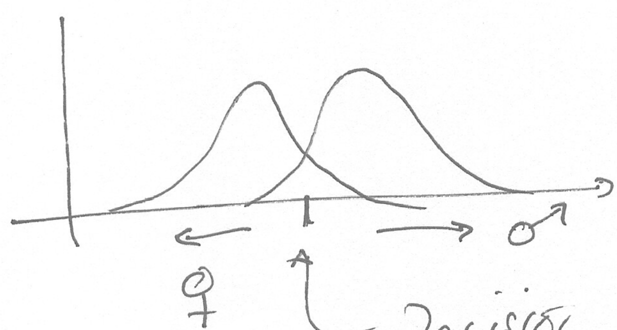
•

• Algorithmicly complex.

(B) model $P(♀|x)$ directly
(we'll talk about this later)

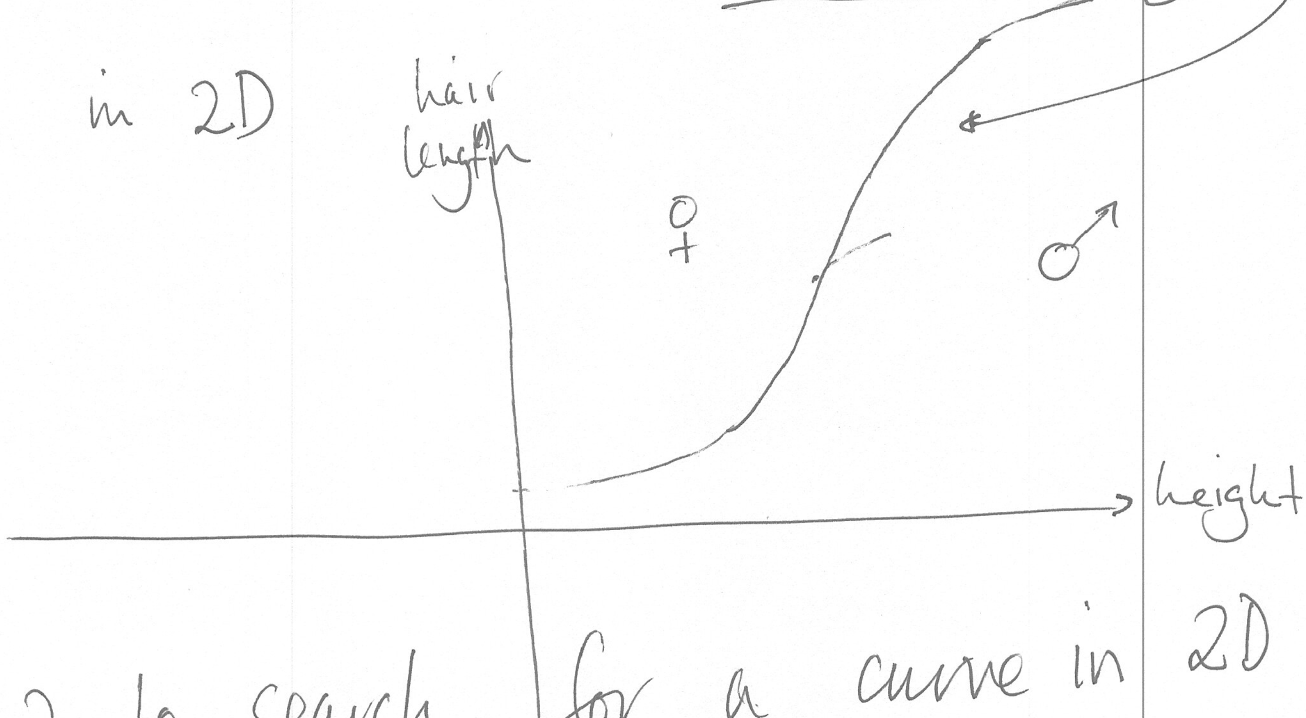
(C) <u>Find decision boundary directly</u>

- recall



<u>Decision boundary</u>

- eg in 2D

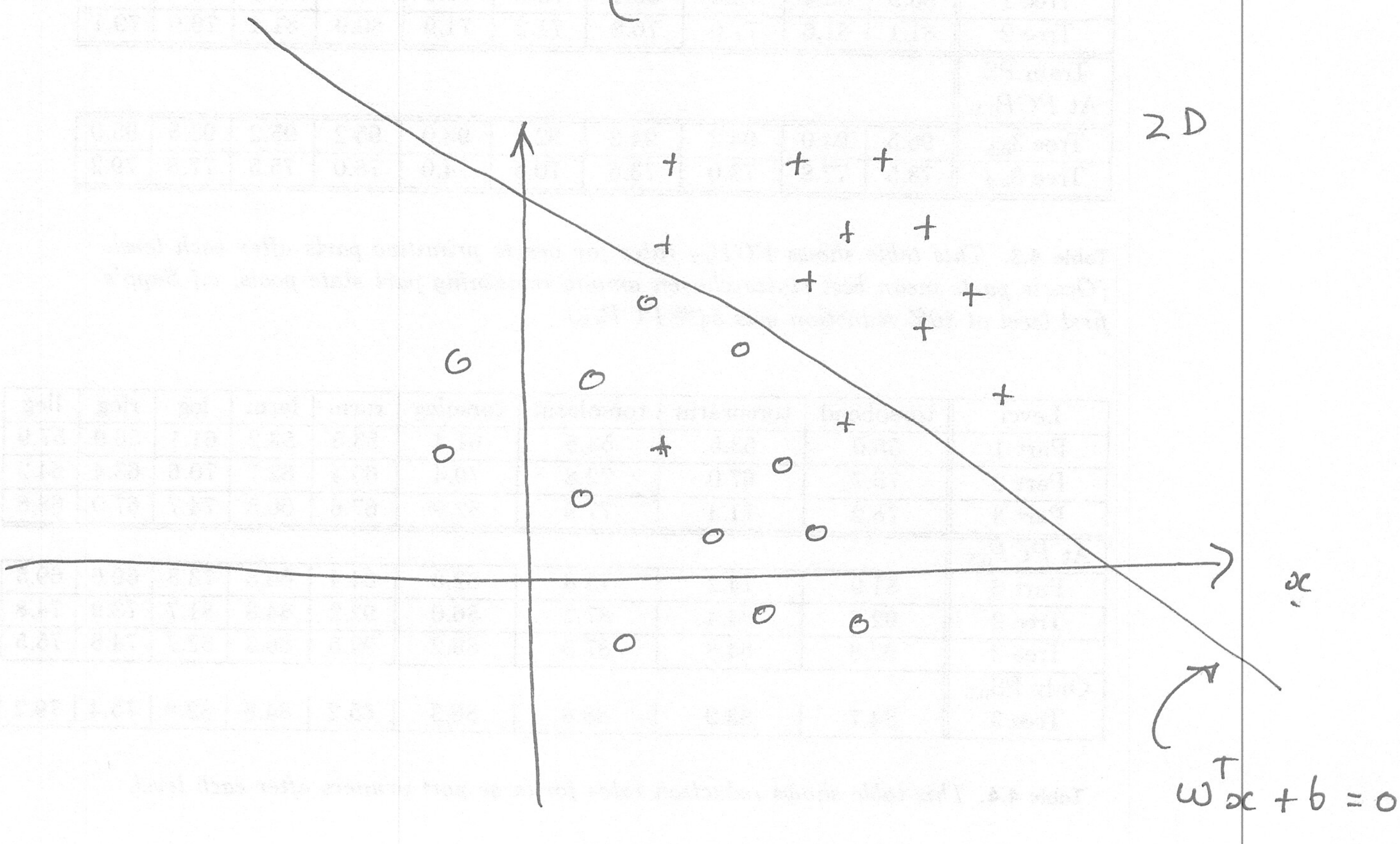


- Hard to search for a curve in 2D or more-D

Easy, highly successful strategy

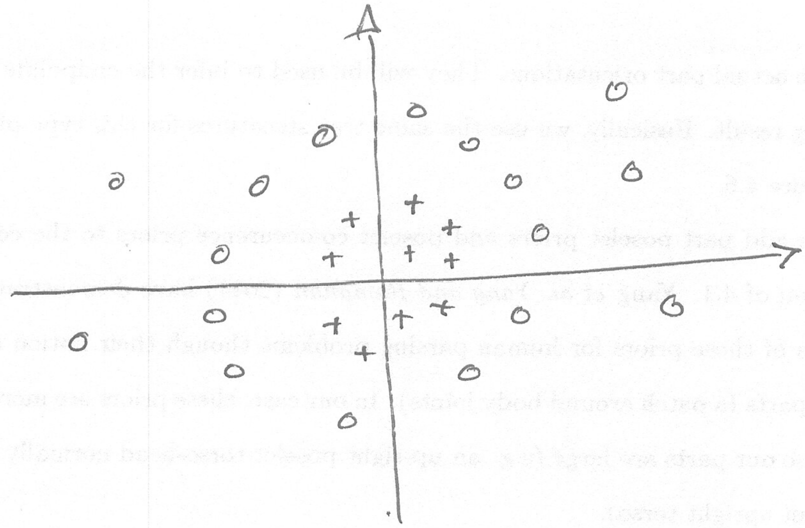
- decision boundary is a flat (line, plane, hyperplane)
- i.e.

$$(w^T x + b)\begin{cases} > 0 & \text{class } 1 \\ = 0 & \\ < 0 & \text{class } 2 \end{cases}$$



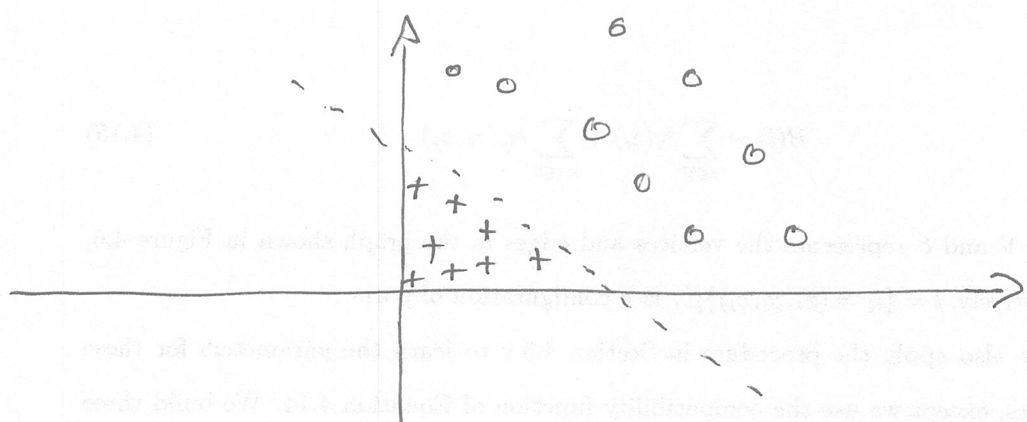
clearly, this won't work always

eg

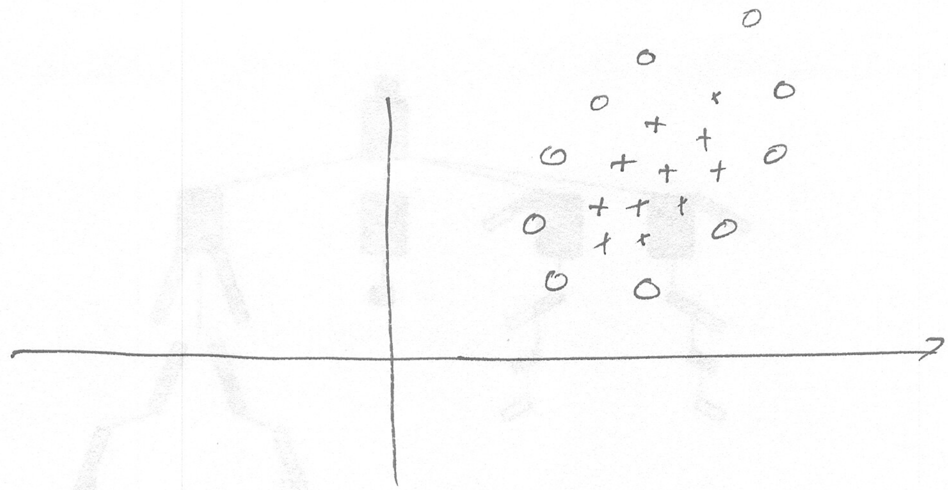


~~by~~ <u>but</u>: map these points by

$$(x, y) \longrightarrow (x^2, y^2)$$



again



<u>but</u>

$$(x, y) \longrightarrow (x^2, xy, y^2, x, y)$$

(recall general ellipse is $ax^2 + bxy + cy^2 + dx + ey + f = 0$)

and we get linear boundary.

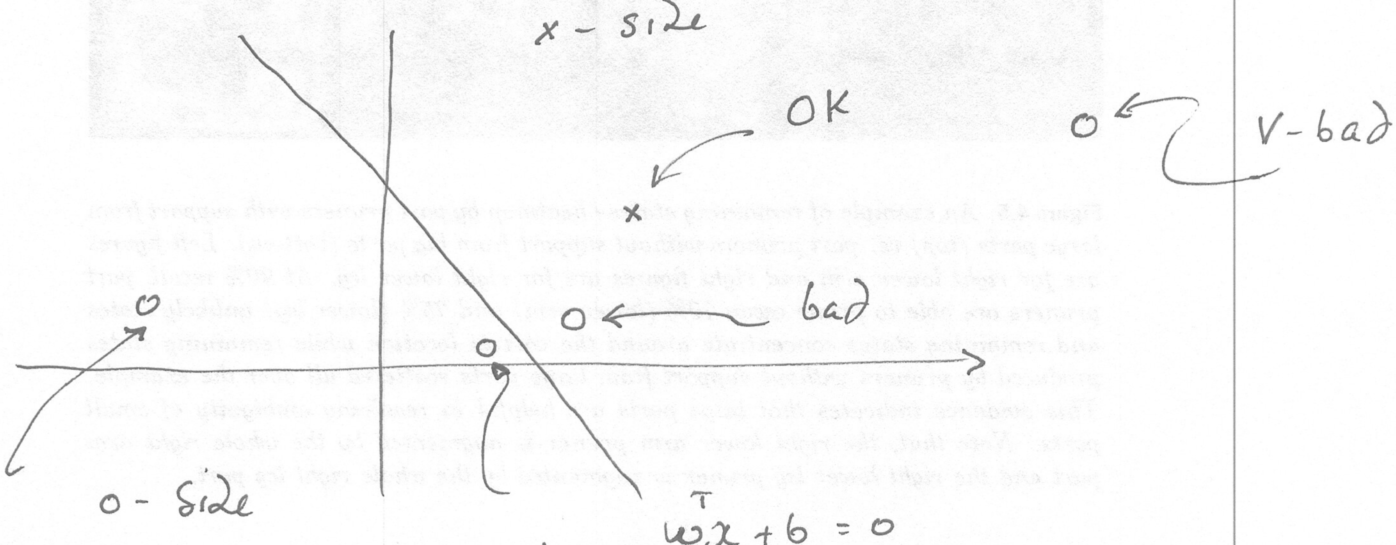
<u>General principle here:</u>

- with enough features a linear classifier will behave well.

How to choose a linear classifier

Q: what is w, b ?

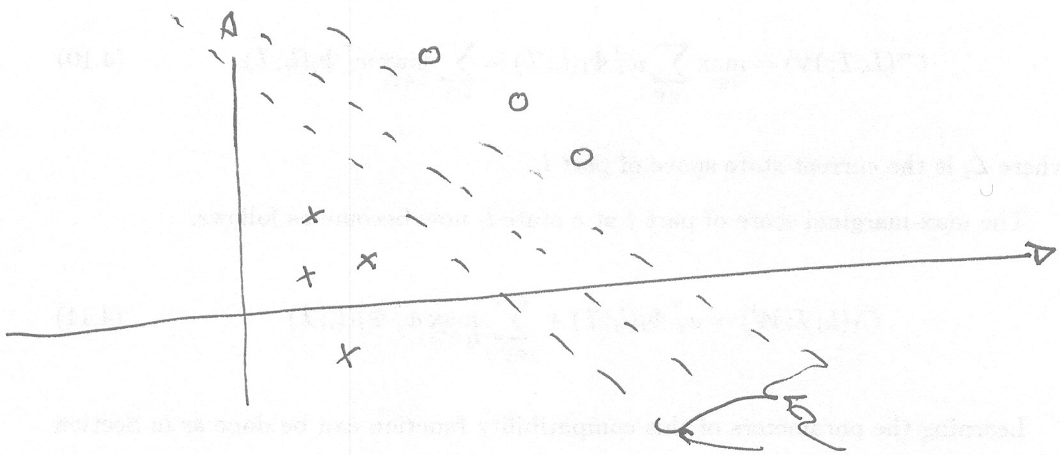
A: minimize <u>loss</u> of using classifier



OK.

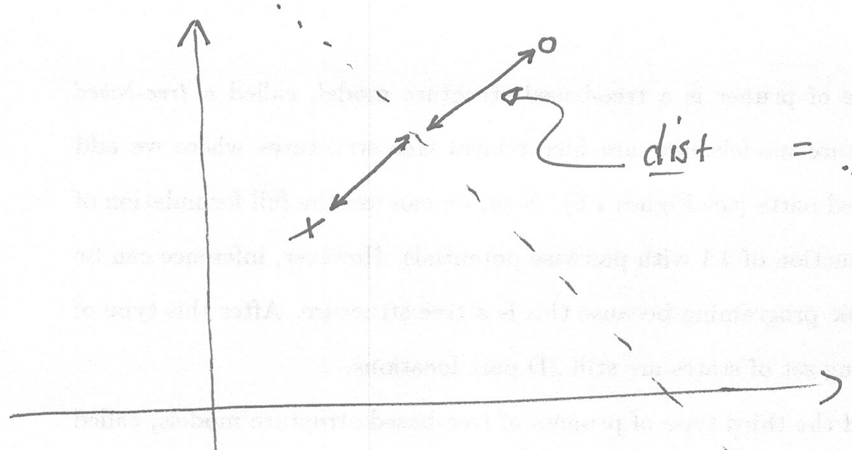
might become bad w/small change.

We need to deal w/ easy cases



each of these lines looks OK — but which do we choose?

- Good choice
  - closest examples should be as far away as possible



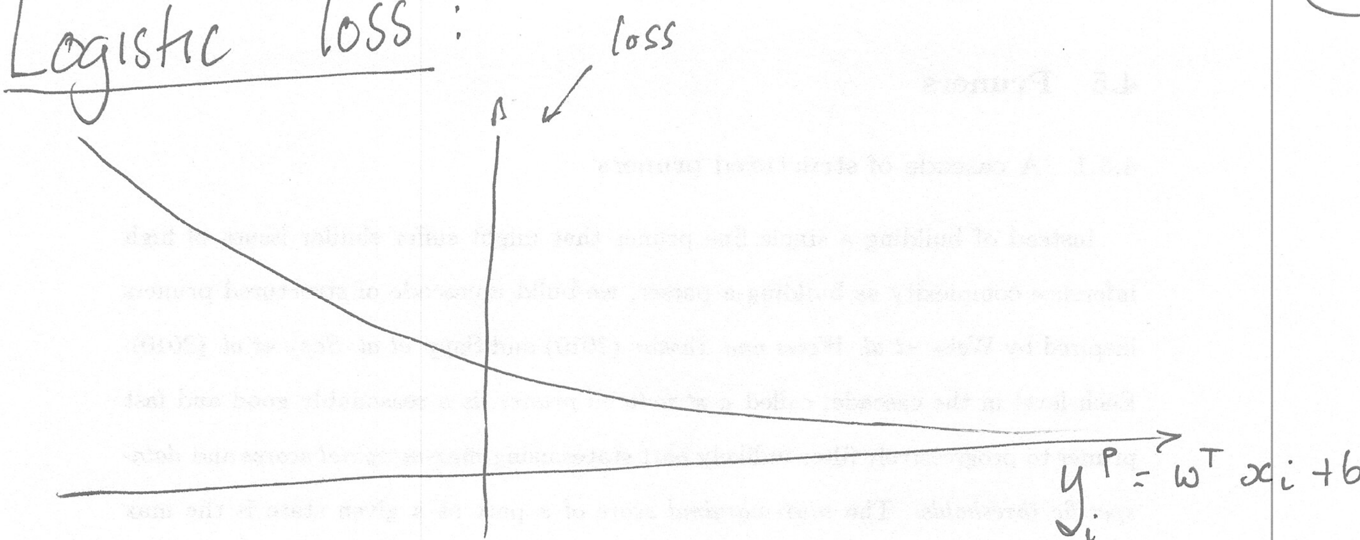
$$\text{dist} = \frac{|w^T x + b|}{\|w\|}$$

- hence, if loss is zero, we would like to minimize $\|w\|^2$
- if loss is non zero, small $\|w\|^2$ is a good idea
- Hence minimize

$$\text{Loss} + \theta \|w\|^2$$

weighting parameter choose later

Logistic loss :



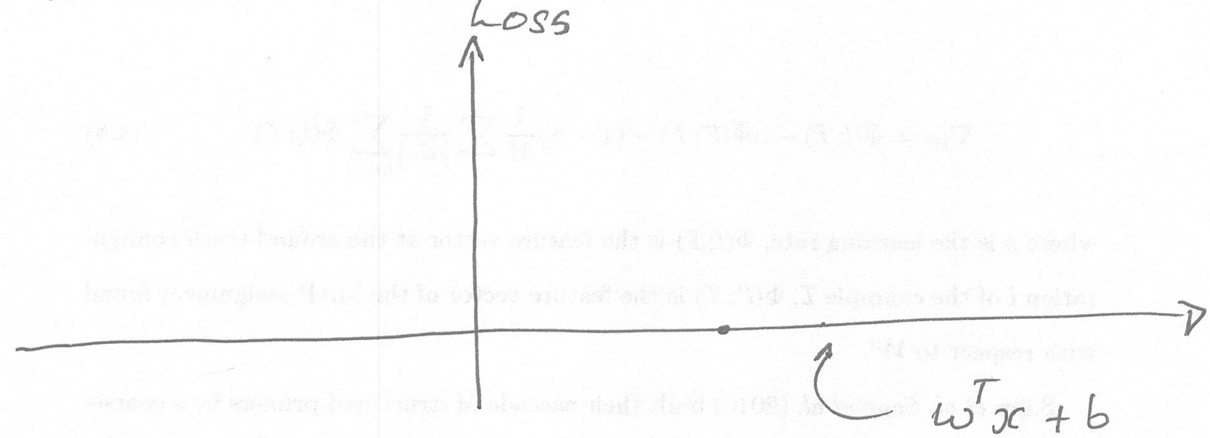
$$\log\left[1 + \exp(-y_i \cdot y_i^p)\right]$$

- leads to

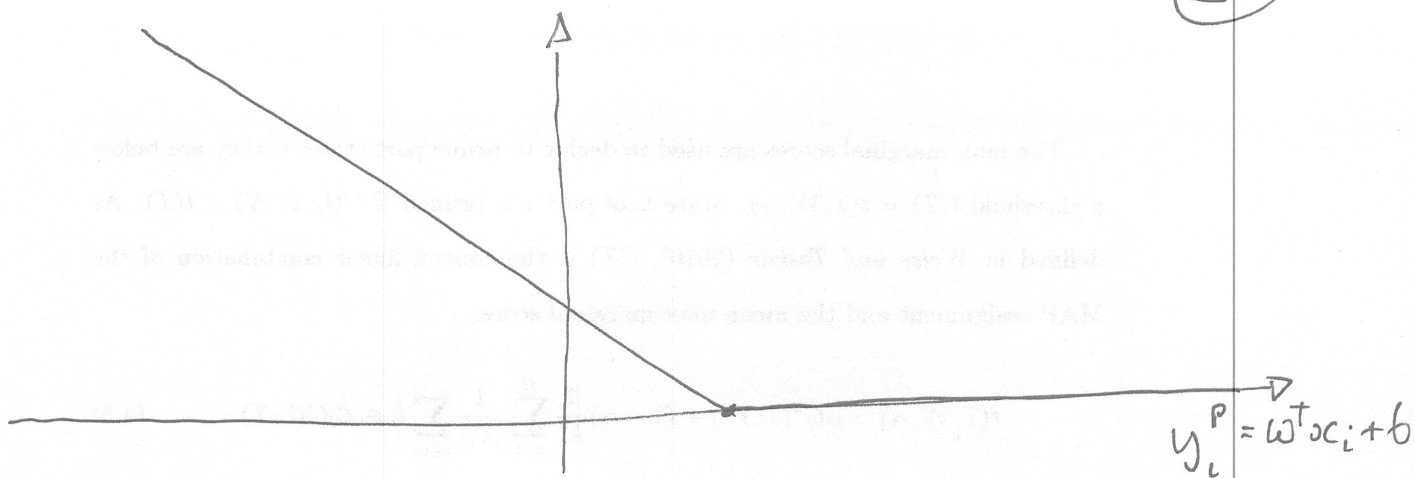
$$\min \theta \|w\|^2 + \sum_i \log\left[1 + \exp(-y_i \cdot y_i^p)\right]$$

Loss:

- consider an example of class 1
- we want $w^T x + b > 0$



- for $w^T x + b \gg 0$, Loss $= 0$
- for $w^T x + b < 0$, loss is big
- for $w^T x + b \ll 0$, bigger
- loss should not grow too fast, otherwise one example dominates
- loss should be non-zero for small +ve $w^T x + b$



Hinge loss:

- example has label $y_i \in \{1, -1\}$
- we predict $y_i^p = w^T x_i + b$
- loss is

$$\max\{0, 1 - y_i \cdot y_i^p\}$$

- plotted above for $y_i = 1$.
- leads to

$$\min \quad \theta \|w\|^2 + \sum_i \max\{0, 1 - y_i y_i^p\}$$

which is hard.