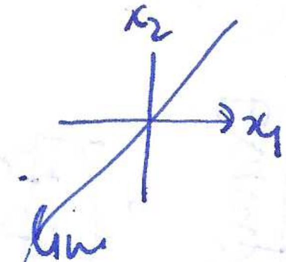
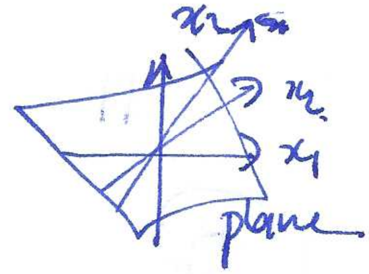
Support vector machines

hyperplane: $\mathbb{R}^{d-1} \subseteq \mathbb{R}^{d}$.



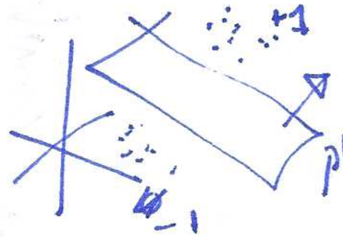


$a_0 + a_1x_1 + a_2x_2 = 0$ $\qquad$ $a_0 + a_1x_1 + a_2x_2 + a_3x_3 = 0$

in $\mathbb{R}^d$: $\underbrace{a_0 + \sum_{i=1}^{d} a_i x_i}_{\text{function of } \underline{x} = (x_i)} = 0$.

$\left.\begin{array}{l} f(\underline{x}) > 0 \\ f(\underline{x}) < 0 \end{array}\right\}$ two sides of hyperplane

Classification.



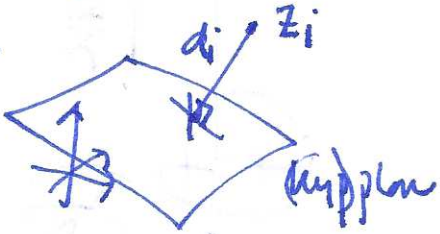
$\underline{n}$ normal vector is $\underline{n} = (a_i)$.

plane separates data.

get a bit more info $|f(x)| >> 0$ means point is far from the plane

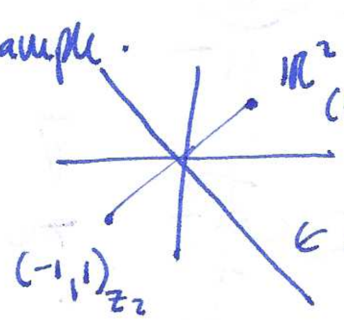
Maximal margin hyperplane / optimal separating hyperplane

$Z = \{z_i\} \subseteq \mathbb{R}^d$.

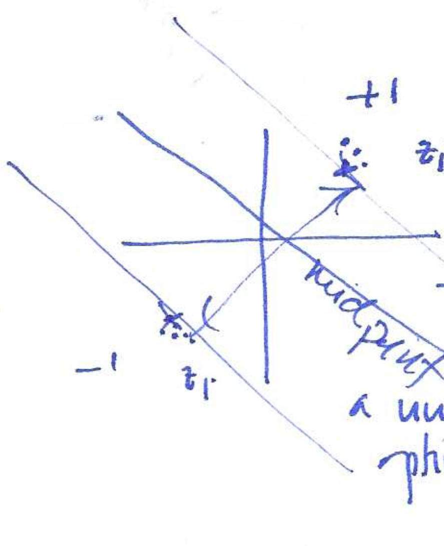


d_i is called the margin.

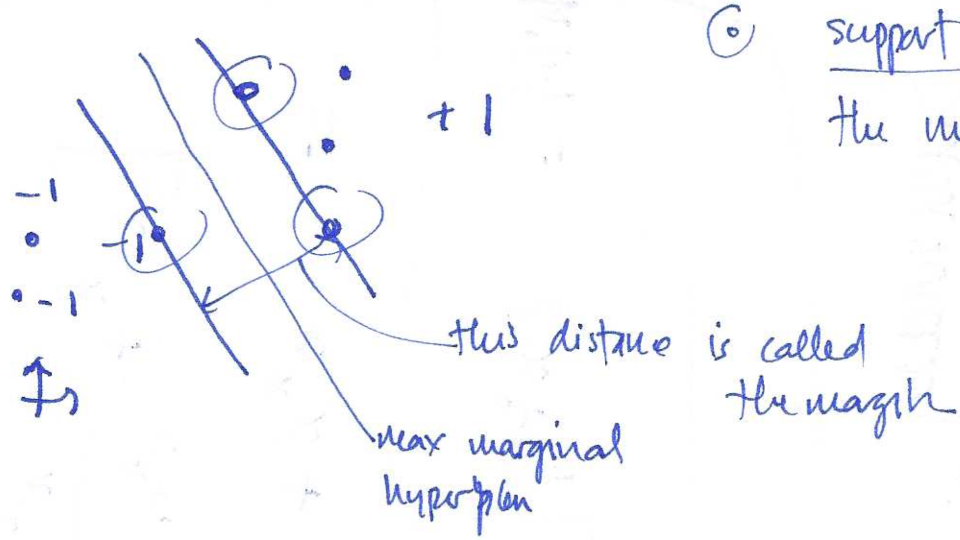
simplest example.



← any plane through origin works.

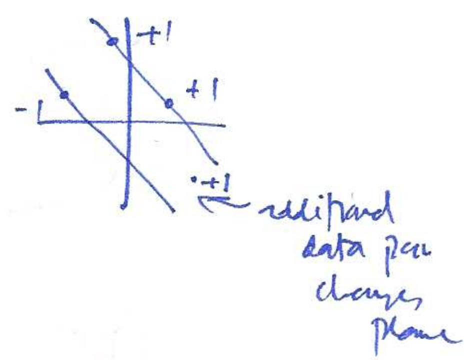


this will determine a unique optimal plane.



support vectors. these determine the maximal margin hyperplane!!

unstable!



Constructing the maximal margin classifier

setup $Z = \{z_i\} \subseteq \mathbb{R}^d$ labelled $\{\underset{y_i}{\pm 1}\}$. separated by a plane.

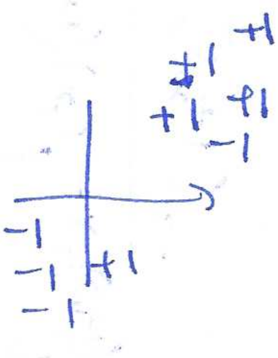
$$\underset{\{a_i\}_{i=0}^{d}}{\text{maximize}} \quad M \qquad \text{subject to} \quad \sum_{i=1}^{d} a_i^2 = 1.$$

$$y_i(a_0 + a_1 z_{i1} + \cdots + a_d z_{id}) \geq M.$$

$$\overset{\parallel}{(\text{label}) \times f(z_i)}.$$

Fact: this optimization problem can be solved $\square$.

General case $z = (z_i)$ not separated by a hyperplane



Support vector classifiers / soft margin classifier

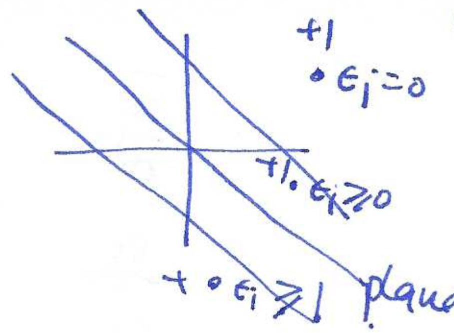
$$\underset{\substack{a_i \\ 1 \le i \le d}}{\max} \; \underset{\epsilon_j \le n}{} M \quad (\text{set } \textstyle\sum a_i^2 = 1)$$

subject to $y_i(a_0 + a_1 z_{i1} \cdots a_d z_{id}) \geq M(1 - \epsilon_i)$

$$\epsilon_i \geq 0 \qquad \sum_{i=1}^{n} \epsilon_i \leq C$$

$\epsilon_i > 0$ means z_i on wrong side of margin

$\epsilon_i > 1$ z_i on wrong side of plane



$C = 0$ original case, i.e. all points must be separated by hyperplane

C tuning parameter ← chosen through cross validation.

in general: C small plane classifier strongly fit to data
Long low bias high variance

C large looser fit, more bias hopefully lower variance.

<u>key fact</u> z_i with $\epsilon_i = 0$ don't effect the choice of plane!

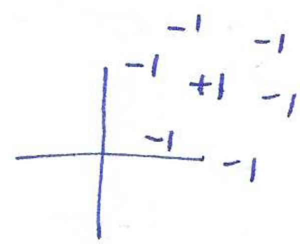
observe the plane only depends on data points w/ $\epsilon_i > 0$

i.e. lying in the margin, or on the wrong side of the plane.

These are called the <u>support vectors</u>.

<u>Support vector machines</u>.

aim: deal with non-linear boundaries.



<u>fact</u>: finding the plane depends only on the <u>inner products</u> of the data points

$\underline{a}.\underline{b} = (a_1, a_2, \ldots, a_d).(b_1, \ldots, b_d) = a_1b_1 + a_2b_2 + \ldots + a_db_d$.

• the linear support vector classifier can be represented as

$f(x) = a_0 + \sum_{i=1}^{n} a_i \langle x, x_i \rangle$ n parameters not d ! (4)

to find the a_i need the $\binom{n}{2}$ inner products $\langle x_i, x_i' \rangle$ for $(x_i) \in X$.

$= \frac{n(n-1)}{2} \sim n^2$.

- $a_i = 0$ unless x_i is a support vector. Let S be set of indices of support vectors.

$$f(x) = a_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle$$ typically $|S| << \cancel{d}\ n$?

we can replace the inner product with ~~the~~ a generalized inner product called a kernel $K(x_i, x_i')$

standard inner product is $K(x_i, x_i) = x_i \cdot x_i' = \sum_j x_{ij} x_{ij}'$.
standard gives linear hyperplane classifier.

other choices : • $K(x_i, x_i') = \left(1 + \sum_{j=1}^{d} x_{ij} x_{ij}'\right)^k$ $\cancel{d}$.

polynomial kernel of degree k.

a support vector machine is a classifier with a (possibly non-linear) kernel $K(x_i, x_i')$ and the classifying function will have the form $f(x) = a_0 + \sum_{i \in S} a_i K(x, x_i)$

• radial kernel $K(x_i, x_i') = \exp\left(-\gamma \sum_{j=1}^{d} (x_{ij} - x_{ij}')^2\right)$.

suppose there are more than two outcomes, not just $\{\pm 1\}$?

suppose there are $\{1, 2, \dots, K\}$ outcomes.

- can do $\binom{K}{2}$ pairwise comparisons and pick most frequent (one vs one).
- (one vs all) build K SVMs identifying i against all other. $\{1, \dots K\} \setminus i$ $\rightsquigarrow$ each of the K classifiers gives a weight/margin $f_i(x)$ $\leftarrow$ choose biggest one.

---

relation to logistic regression.

can rewrite criterion for finding the support vector classifier

$f(x) = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_d x_d$ as

$$\underset{a_i}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\{0, 1 - y_i f(x_i)\} + \lambda \sum_{j=1}^{d} a_i^2 \right\}$$

λ: non-negative tuning parameter

λ large, a_i are small, more margin violations, high bias low variance

λ small a_i large few margin violations low bias, high variance

General setup : "Loss + Penalty"

$$\underset{a_i}{\text{minimize}} \left\{ L(X, y, a_i) + \lambda P(a_i) \right\}$$

λ: tuning parameter

L: Loss function / error metric — measures how closely model fits data

P: penalty function on parameters.

Lasso : $L(X, y, a) = \sum_{i=1}^{n} \left(y_i - a_0 - \sum_{j=1}^{d} x_{ij} a_j \right)^2$ (**)

$P(a) = \sum_{i=1}^{n} |a_i|$.

note (*) zero for non-support element

(**) very small for far away elements.