ECE 6143
Lecture 7

Reminders

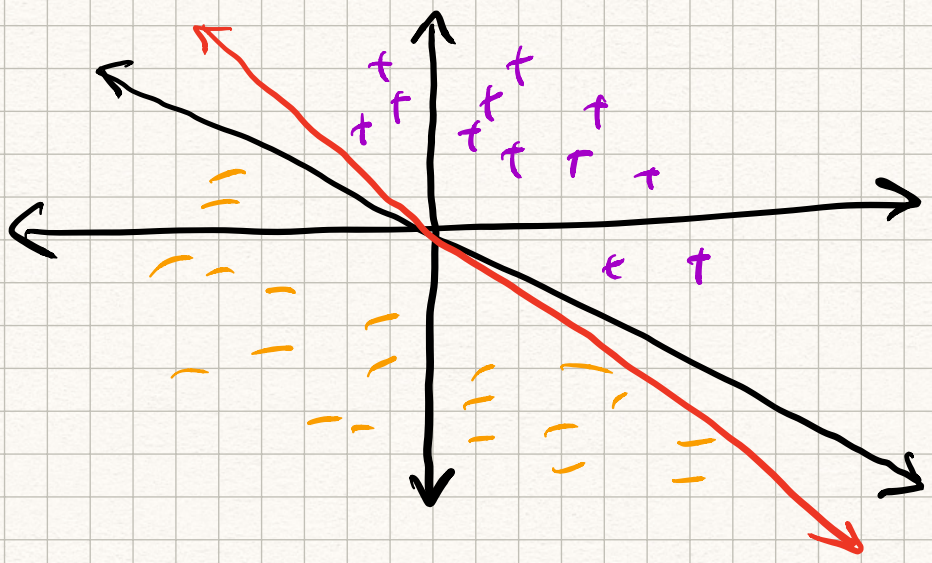
- All lectures will be online until further notice.
- All office hours will be online too (time TBD).
- Midterm postponed to Mar 31
- HW4 will be posted soon, due Apr 2.

Recap

- k NN
- The perceptron algorithm
- Support vector machine
- Kernel methods.

} Classification

Support Vector Machines (SVMs).



Q. What is the "best" separator b/w the classes?

Revisiting the perceptron

Dataset $\{(x_1, y_1), (x_2, y_2) \cdots, (x_n, y_n)\}$.

Find w

s.t. $\text{sign}\langle w, x_i \rangle = y_i$

Rewriting this:

$$l_i(w) = \begin{cases} 0 & y_i \langle w, x_i \rangle \geq 1 \\ 1 - y_i \langle w, x_i \rangle & \text{otherwise} \end{cases}$$

Intuition: 1) If y_i & $\langle w, x_i \rangle$ are both same sign AND $\langle w, x_i \rangle$ is large, then no loss.

2) If y_i & $\langle w, x_i \rangle$ are of opposite signs, then loss ≥ 1

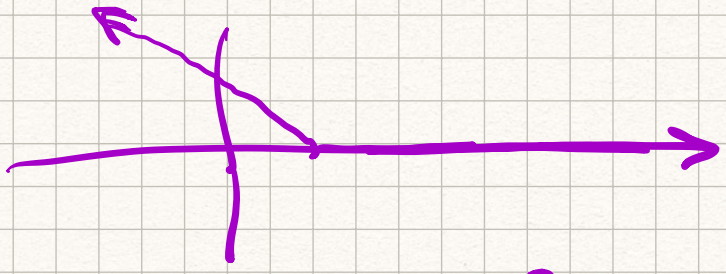
3) Everything else: in between 0 & 1.

$$L(w) = \sum_{i=1}^{n} l_i(w)$$

$$L_{hinge}(w) = \sum_{i=1}^{n} \max(0, 1 - y_i \langle w, x_i \rangle)$$

Hinge loss

$f(z) = \max(0, 1-z)$



$$l_i(w) = \begin{cases} 0 & y_i \langle w, x_i \rangle \geq 1 \\ 1 - y_i \langle w, x_i \rangle & \text{otherwise.} \end{cases}$$

$$\partial l_i(w) = \begin{cases} 0 & y_i \langle w, x_i \rangle \geq 1 \\ -y_i x_i & \text{otherwise.} \end{cases}$$

(Sub) gradient descent:

$$w_{k+1} \leftarrow w_k - \eta_k \partial L(w_k)$$

$$= \begin{cases} w_k & y_i \langle w, x_i \rangle \geq 1 \end{cases}$$

$\{ w_k + y_i x_i$ otherwise

→ Piecewise linear, therefore gradient can be computed in each piece.

* $\partial l_i(w) = 0$ when $l_i(w) = 0$
* when $l_i(w) = 1 - y_i \langle w, x_i \rangle$

$$\partial l_i(w) = \partial (1 - y_i \langle w, x_i \rangle)$$
$$= -y_i \partial \langle w, x_i \rangle$$
$$= -y_i x_i$$

SVM

$$L_{SVM}(w) = L_{hinge}(w) + \frac{\lambda}{2} \|w\|_2^2$$
$$= \sum_{i=1}^{n} \max(0, 1 - y_i \langle w, x_i \rangle) + \frac{\lambda}{2} \|w\|_2^2 .$$

"primal SVM"

(Primal) $\min_w L_{SVM}(w)$

Solution: w^*.

(Dual)

$$\max_{\alpha} \; -\sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\text{s.t.} \quad 0 \le \alpha_i \le \frac{1}{\lambda}.$$

Solution: α^*

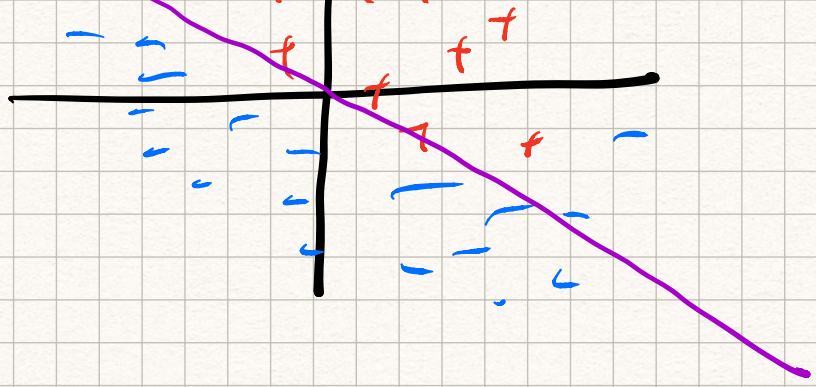
$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i y_i x_i}$$

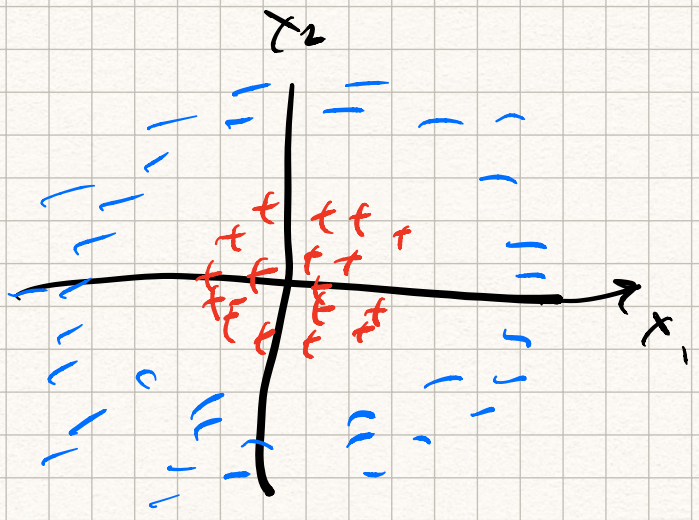
Intuition: SVM model is a linear combination of the data points.

Wherever $\alpha_i^* = 0$, solution w^* does not depend on x_i.

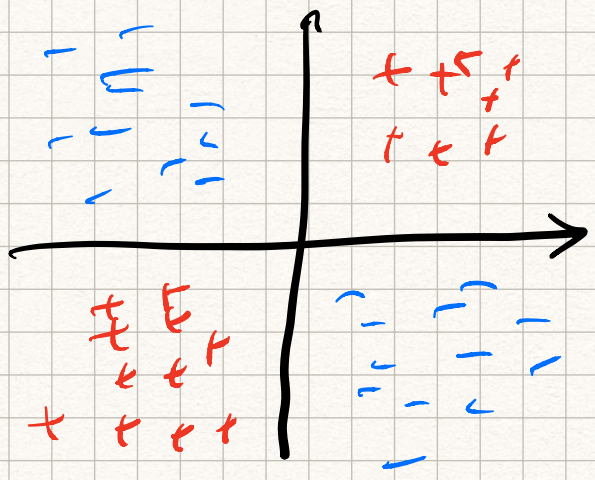
∴ the data points x_i for which $\alpha_i^* \neq 0$ are called the support vectors.

Separability



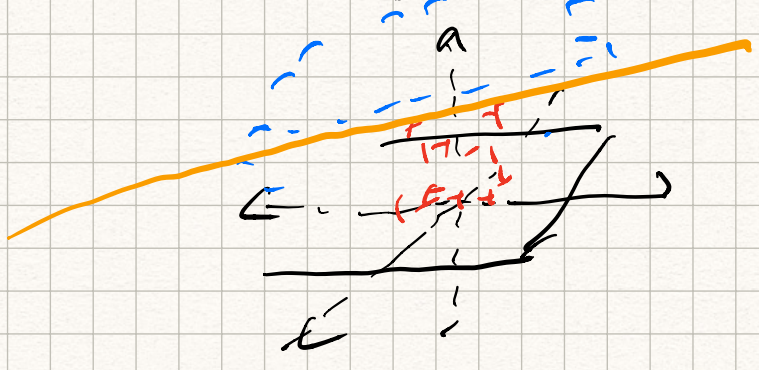


Ex 1



Ex 2

Ex 1 : $(x_1, x_2) \rightarrow (x_1, x_2, x_1^2 + x_2^2)$



Ex 2 : $(x_1, x_2) \rightarrow x_1 x_2$.

Idea: Embed the data into a different feature space in which

the points become linearly separable.

Kernel methods

$$x \longrightarrow \phi(x)$$

$\phi(\cdot)$ is a feature transformation of the data called the kernel mapping.

Typical feature transformations:

- Original features
- Quadratic functions.

$$(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1 x_2).$$

- Higher order polynomials

$$(x_1, x_2) \rightarrow (x_1^2, x_2^2, x_1 x_2, x_1^3, x_2^3, x_1^2 x_2, x_1 x_2^2, x_1, x_2).$$

etc.

$$(x_1, x_2, \cdots, x_d) \xrightarrow{\text{Quadratic}}$$

$$(x_1^2, x_2^2, \cdots, x_d^2, x_1 x_2, x_1 x_3, \cdots, x_1 x_d, \cdots \cdots x_{d-1} x_d)$$

Approximately d^2.

Going back to circles example:

$$(x_1, x_2) \rightarrow (x_1, x_2, x_1^2, x_2^2, x_1 x_2).$$

Circles $w = (0, 0, 1, 1, 0)$.

$$\langle w, \phi(x) \rangle = x_1^2 + x_2^2.$$

$$f(x) = \text{sgn}(\langle w, \phi(x) \rangle - \text{radius}).$$

XOR $w = (0, 0, 0, 0, 1)$.

$$f(x) = \text{sgn}(\langle w, \phi(x) \rangle).$$

What is important is <u>not</u> the feature mapping, but the ability to take dot products with the feature mapping.

→ "kernel trick".

Implicitly define feature mapping via the <u>dot product</u>.

→ kernel dot product

or kernel inner product.

Examples of kernel dot products

* Regular dot product
* Quadratic dot product

$$K(x,y) \rightarrow (1 + \langle x,y \rangle)^2$$

$$(x_1, x_2) \rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$$

$$(y_1, y_2) \rightarrow (1, \sqrt{2}y_1, \sqrt{2}y_2, \sqrt{2}y_1y_2, y_1^2, y_2^2)$$

* Cubic dot product

$$K(x,y) \rightarrow (1 + \langle x,y \rangle)^3$$

Dot product $= 1 + 2x_1y_1 + 2x_2y_2 + 2x_1y_1x_2y_2 + x_1^2y_1^2 + x_2^2y_2^2$

$$= (1 + x_1y_1 + x_2y_2)^2$$

$$= (1 + \langle [x_1, x_2], [y_1, y_2] \rangle)^2$$

* Exponential

$$K(x,y) = \exp(-\|x-y\|^2)$$

"Gaussian kernel"

"Radial basis function" / RBF.