1. Describe the regression task

- We have a domain $X \in \mathbb{R}^d$ and a label set $Y \in \mathbb{R}$
- The hypothesis set is $\mathcal{H}_{reg} : \mathbb{R}^d \to \mathbb{R}$

OBJECTIVE : $$h^* = \underset{h \in \mathcal{H}_{reg}}{\arg\min}\ L_d(h)$$

2. Introduce the linear regression model class and derive the optimal solution

- Linear hypothesis class: $\mathcal{H}_{lin} = \{ x \to \langle w, x \rangle + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$
- Squared loss: $l_2(h, (x,y)) \overset{def}{=} (h(x) - y)^2$

ERM: $$L_s(h) = \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)^2$$

HOM. COORD. $$w^* = \underset{w, \cancel{b}}{\arg\min}\ \frac{1}{m} \sum_{i=1}^{m} (\langle w, x_i \rangle - y_i)^2$$

SET GRADIENT TO 0

$$\frac{\partial L_s}{\partial w} = \frac{2}{m} \sum_{i=1}^{m} [\langle w, x_i \rangle x_i - x_i y_i]$$

$$\frac{\partial L_s}{\partial w} = 0 \Rightarrow \sum_{i=1}^{m} \langle w, x_i \rangle x_i = \sum_{i=1}^{m} y_i x_i$$

$$A = \sum_{i=1}^{m} x_i x_i^T$$

$$b = \sum_{i=1}^{m} y_i x_i \qquad Aw = b \Rightarrow w = A^{-1} b$$

3. Describe how the process can be extended using regularization to avoid large coefficients

We define the regularization function $R(w) = \lambda \|w\|^2$

↳ new problem: $$w^* = \underset{w}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2$$

$$\frac{\partial L_s}{\partial w} = \frac{1}{m} \sum_{i=1}^{m} [\langle w, x_i \rangle x_i - x_i y_i] + 2\lambda w$$

$$\frac{\partial L_s}{\partial w} = 0 \Rightarrow 2\lambda m w + \sum_{i=1}^{m} \langle w, x_i \rangle x_i = \sum_{i=1}^{m} y_i x_i$$

$$(2m\lambda I + A) w = b \Rightarrow w = (2m\lambda I + A)^{-1} b$$

. Define the clustering problem

$X \subseteq \mathbb{R}^d$: domain

$d: X^2 \to \mathbb{R}^+$: distance function

→ symmetric : $d(x,y) = d(y,x) \ \forall x,y \in X$

→ positive

→ triangular inequality : $d(x,z) \leq d(x,y) + d(y,z) \ \forall x,y,z \in X$

OUTPUT : $C = (C_1, \dots C_k)$

→ $\bigcup_{i=1}^{k} C_i = X$

→ $C_i \cap C_j = \emptyset \ \forall i,j \in \{1,\dots k\}, i \neq j$

SOMETIMES k IS GIVEN AS INPUT, SOMETIMES IT IS FREE.

2. Introduce the cost function for k-means and describe Lloyd's algorithm

$d(x,y) = \|x-y\|_2$ (Euclidean distance)

Cluster cost : $\sum_{x \in C_i} d(x, \mu_i)^2$ → CENTROID $\qquad \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$

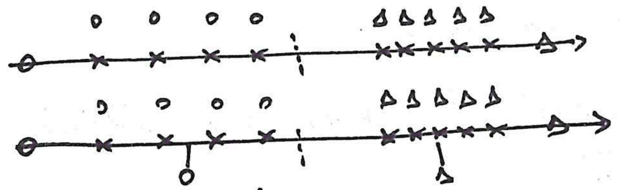
Lloyd's algorithm :

1. Select k random centroids

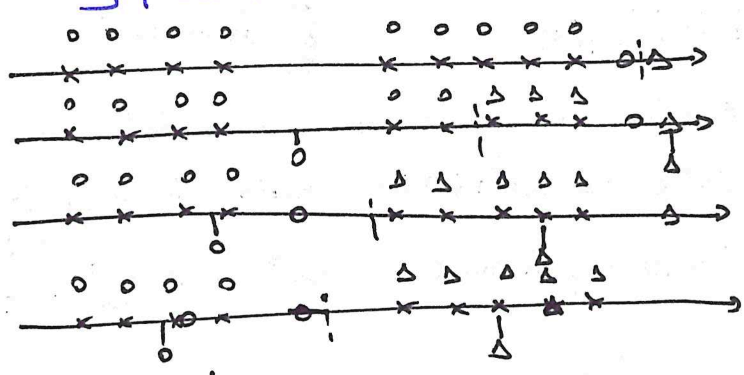
REPEAT

2. Partition the points : $c(x) = \arg\min_{i \in \{1,\dots k\}} d(x, \mu_i) \ \forall x$
3. Recompute the centroids
4. Convergence: no change / threshold error / number of iterations

3. Use Lloyd's algorithm to solve the clustering problem



CONVERGED!



CONVERGED!

Exercise

. Describe the classification task

X : domain set

$\mathcal{Y}$: label set

$\mathcal{H}: X \to \mathcal{Y}$: hypothesis set

S: training set $((x_1, y_1), \dots (x_m, y_m))$

f: function to be learned (unknown)

D: sampling distribution over X

$$L_{D,f} \overset{def}{=} P_{x \sim D}[h(x) \neq f(x)]$$

$$L_S \overset{def}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(x) \neq f(x))$$

$$\text{ERM: } h^* = \underset{h \in \mathcal{H}}{\arg\min}\ L_S(h)$$

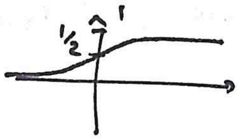
2. Describe logistic regression

$\mathcal{H}: \phi_{sig} \circ L_d$

↳ LINEAR

↳ SIGMOID

$$\phi_{sig}(z) = \frac{1}{1+e^{-z}}$$



$$h_w(x) = \frac{1}{1+e^{-\langle w, x\rangle}}$$

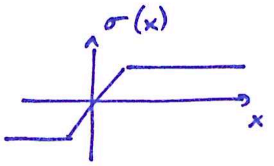
↳ HOMOG. COORD.

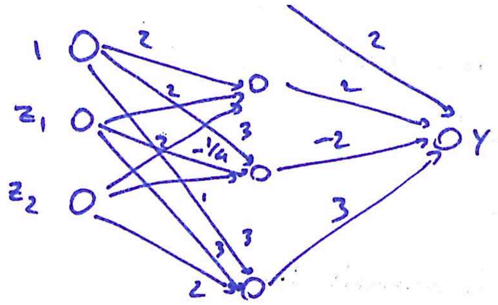
$$\ell(h_w, (x,y)) = \log(1+e^{-y\langle w,x\rangle})$$

$$\text{ERM: } w^* = \underset{w \in \mathbb{R}^d}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} \log(1+e^{-y_i \langle w, x_i\rangle})$$

Exercise 4

$$\sigma(x) = \begin{cases} 1 & x \geq 1 \\ x & -1 \leq x \leq 1 \\ -1 & x \leq -1 \end{cases}$$





$$w^{(1)} = \begin{pmatrix} 2 & 3 & 3 \\ 2 & -1/4 & 3 \\ 2 & 1 & 2 \end{pmatrix} \qquad w^{(2)} = \begin{pmatrix} 2 \\ 2 \\ -2 \\ 3 \end{pmatrix}$$

$$z = (1 \ \ 3)$$

$$a_{1,1} = \langle w_1^{(1)}, [1, z] \rangle = (2\ 2\ 2) \cdot (1\ 1\ 3)^T = 7 \qquad o_{1,1} = \sigma(a_{1,1}) = 1$$

$$a_{1,2} = \langle w_2^{(1)}, [1, z] \rangle = (3\ -\tfrac{1}{4}\ 1) \cdot (1\ 1\ 3)^T = \frac{23}{4} \qquad o_{1,2} = \sigma(a_{1,2}) = 1$$

$$a_{1,3} = \langle w_3^{(1)}, [1, z] \rangle = (3\ 3\ 2) \cdot (1\ 1\ 3)^T = 12 \qquad o_{1,3} = \sigma(a_{1,3}) = 1$$

$$a_{2,1} = \langle w^{(2)}, [1, o_1] \rangle = (2\ 2\ -2\ 3) \cdot (1\ 1\ 1\ 1)^T = 5 \qquad y = \sigma(a_{2,1}) = 1$$

. Introduce the concept of kernel and its use in SVMs

Feature space transf.: $\psi : X \to F$ ↪ Hilbert space

$S = ((x_1, y_1) \dots (x_m, y_m)) \to \hat{S}((\psi(x_1), y_1) \dots (\psi(x_m), y_m))$

$\mathcal{H}$: linear

Kernel: $K(x, y) = \langle \psi(x), \psi(y) \rangle$

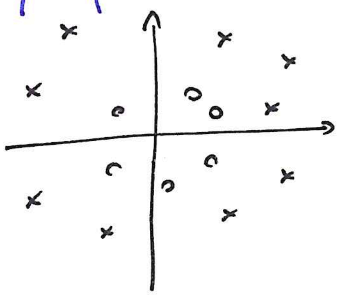
↳ can often be computed without going through ψ

SVM: $w^* = \underset{w}{\text{argmin}} \left(f(\langle w, \psi(x_1) \rangle \dots \langle w, \psi(x_m) \rangle) + R(\|w\|) \right)$

Representer theor.: $\exists \alpha \in \mathbb{R}^m : w^* = \sum_{i=1}^{m} \alpha_i \psi(x_i)$

$\alpha^* = \underset{\alpha}{\text{argmin}} \left(f\left(\sum_{j=1}^{m} \alpha_j K(x_j, x_1), \dots \sum_{j=1}^{m} \alpha_j K(x_j, x_m) \right) + R\left(\sqrt{\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)} \right) \right)$

2. Consider the configuration of training data and a scalar function $\Phi(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ that makes data linearly separable and relate the map to a kernel



$\Phi(x) = \|x\|_2 = x_1^2 + x_2^2$

ONLY POSITIVE

$\langle \Phi(x), \Phi(y) \rangle = x_1^2 y_1^2 + x_1^2 y_2^2 + x_2^2 y_1^2 + x_2^2 y_2^2$

$K_\Phi(x, y) = \langle \langle x, x \rangle, \langle y, y \rangle \rangle$

$(\langle x, y \rangle)^2 = (x_1 y_1 + x_2 y_2)^2 = (x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 y_1 x_2 y_2)$

Exercise 6

1. Let $X = [x_1, \dots, x_n]$, $x_i \in \mathbb{R}^n$ be the data matrix. Introduce PCA

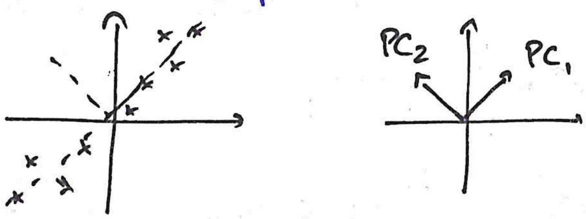
$W \in \mathbb{R}^{n \times d}$ $\qquad y = Wx$

$W = U^T$ (U is an orthonormal base of $\mathbb{R}^n$)

$A = \sum_{i=1}^{n} x_i x_i^T = VDV^T$ (SVD, A is symm. semidef. pos.)

PCA: $w^* = \left(\underset{U \in \mathbb{R}^{d \times n},\, U^T U = I}{\arg\max} \operatorname{tr}(U^T A U) \right)^T$

2. Find the 2 PCs for the graph and describe how PCA can be used to simplify linear regression



Exercise +

1. Describe the linear SVM for classification in the case of non-linearly separable data.

ξ_i : slack variables

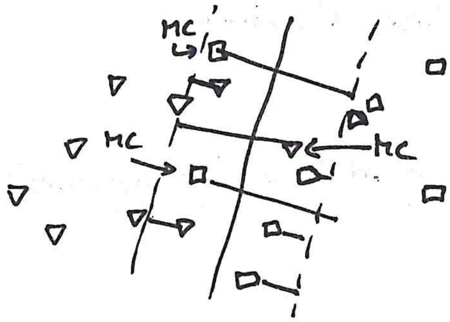
$\rightarrow \xi_i \geq 0$

$\rightarrow 1 - \xi_i \leq y_i (\langle w, x_i \rangle + b)$

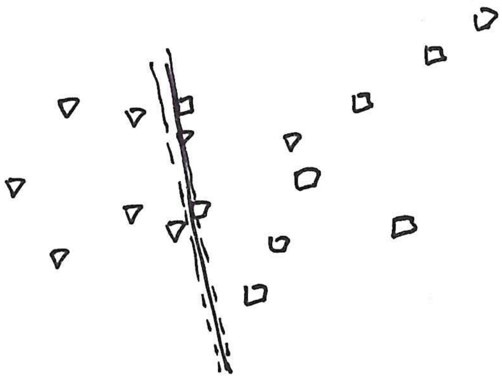
$$\underset{w,b,\xi}{\operatorname{argmin}} \left(\lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right)$$

s.t. $\xi_i \geq 0$, $1 - \xi_i \leq y_i (\langle w, x_i \rangle + b)$

2. Label the misclassified points and draw the non-zero slack variables. Discuss what happens if C increases.



C increase = λ decreases



Exercise 0

1. Describe the concepts of training and generalization errors in supervised learning

TRUE LOSS: $E_{x \sim D}\left[L(h, x)\right]$

TRAINING ERROR: $\frac{1}{m} \sum_{i=1}^{m} L(h, x_i)$

overfitting: training error is minimized, but true loss grows

2. How would you state the final goal of supervised learning?

→ making predictions that generalize to unseen data: using available data to minimize L_D as well as possible (even if the actual value of L_D is unknowable)

3. What role does k-fold CV play?

Training errors → optimize parameters in a class of hypotheses/algorithm

Validation errors → optimize hyperparameters

k-fold CV (properly implemented) allows to avoid overfitting by having a large hypothesis set.

Exercise

1. What do you need to show that $VC(\mathcal{H}) = d$?

$$C \in \mathcal{X}^m$$

$$\mathcal{H}_C = \{(h(c_1), \dots, h(c_m)), h \in \mathcal{H}\}$$

$$h_C \in \mathcal{H}_C : C \to \{0,1\}^m$$

$\mathcal{H}$ shatters C if $\mathcal{H}_C$ contains all possible functions : ~~$\mathcal{H}_C = \{0,1\}$~~

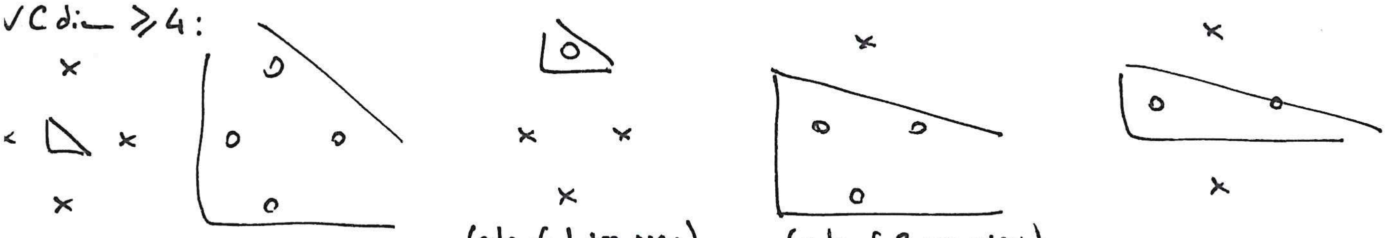
$$\exists h_C : C \to \{0,1\}^m \notin \mathcal{H}_C$$

$\boxed{VC(\mathcal{H}) = d}$

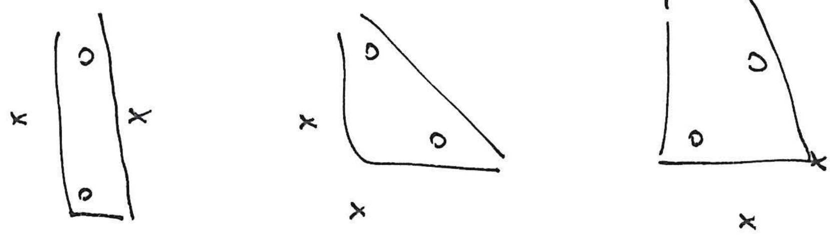
1. $\exists\, C \in \mathcal{X}^d$ shattered by $\mathcal{H}$
2. $\nexists\, C \in \mathcal{X}^{d+1}$ shattered by $\mathcal{H}$ ($\forall C \in \mathcal{X}^{d+1}$, C is not shattered)

2. Find the VC dim of right triangles whose legs are parallel to the axes and whose right angle is in the lower left corner

VC dim ≥ 4:

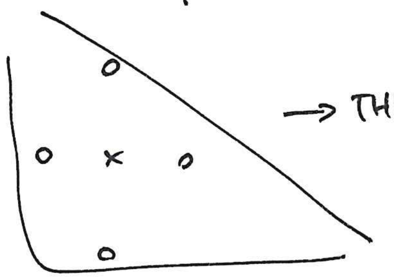


(sets of 1 are easy) (sets of 3 are easy)



the set is shattered

5th point: one point is inside the quadrangle of the other 4



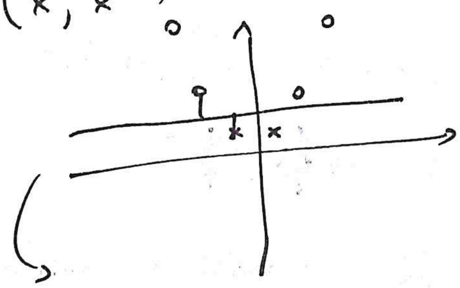
→ THIS CONFIG IS IMPOSSIBLE! VCd < 5

$\boxed{VCd = 4}$

Exercise 10

x	y
-3	1
-2	1
-1	-1
1	-1
2	1
3	1

$\phi: x \to (x, x^2)$



BOUNDARY: $x^2 = 2.5$

MARGIN: 1.5

x	x^2	y
-3	9	1
-2	4	1
-1	1	-1
1	1	-1
2	4	1
3	9	1