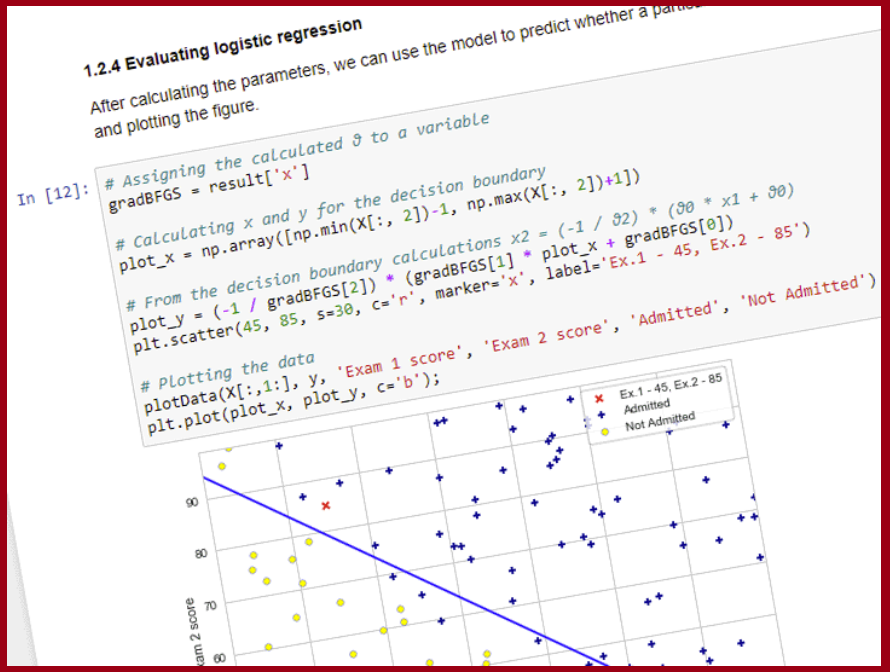
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA



Sample Exercises

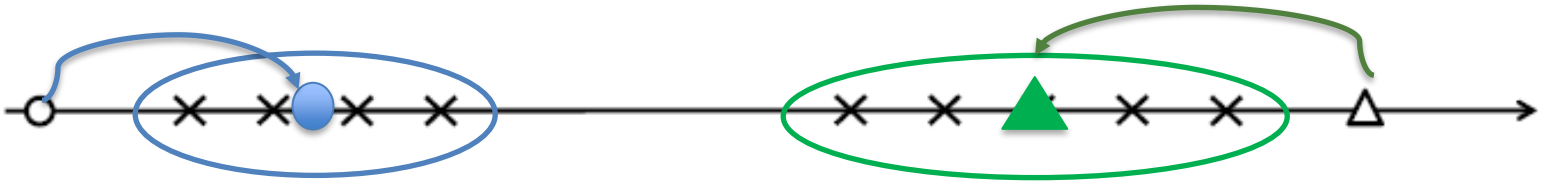
Machine Learning 2023-24
P. Zanuttigh

Exercise 2

1. Define the clustering problem.
2. Introduce the cost function for the K-means clustering problem and describe Lloyd's iterative algorithm.

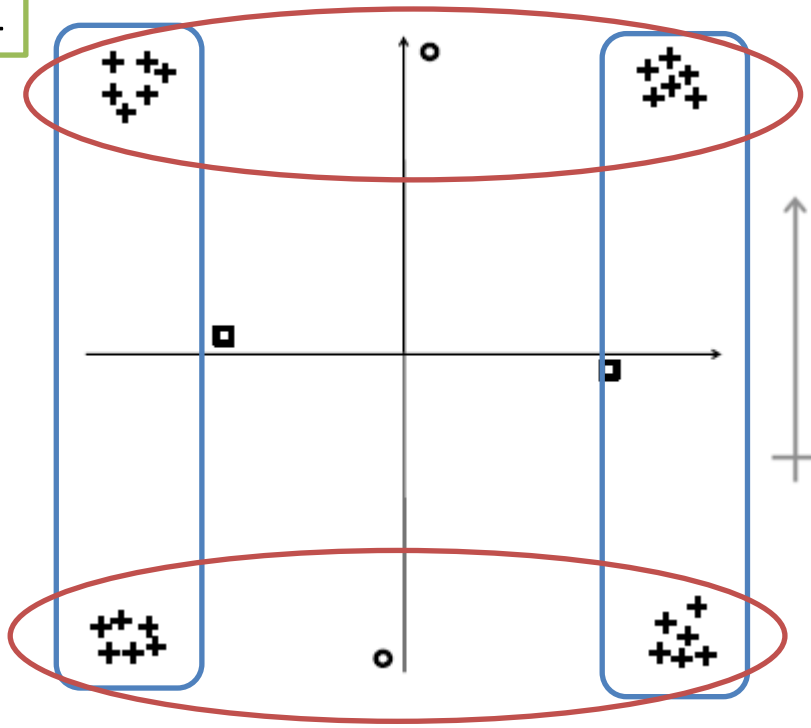
$$G_{km}\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} d(\boldsymbol{x}, \boldsymbol{\mu}_i)^2$$

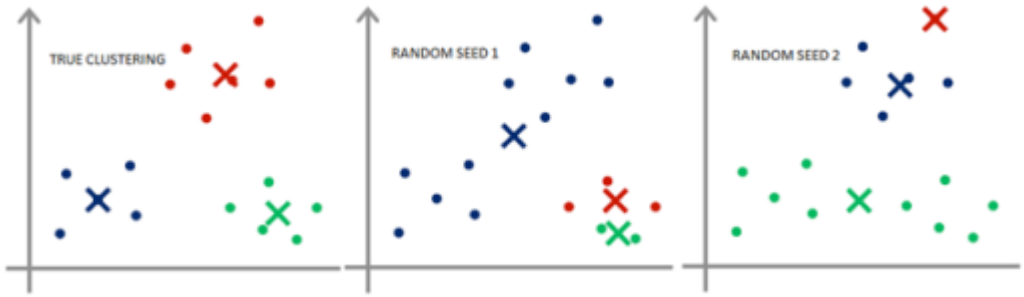
3. Mark approximately in the graph below the solution (clusters and centers) found by Lloyd algorithm for the 2 clusters (K = 2) problem, when the data ($x_i \in \mathbb{R}$) are the crosses in the figure below and the algorithm is initialized with center values indicated with the circle (cluster 1) and triangle (cluster 2) shown in the figure.



K-Means Initialization

1





2

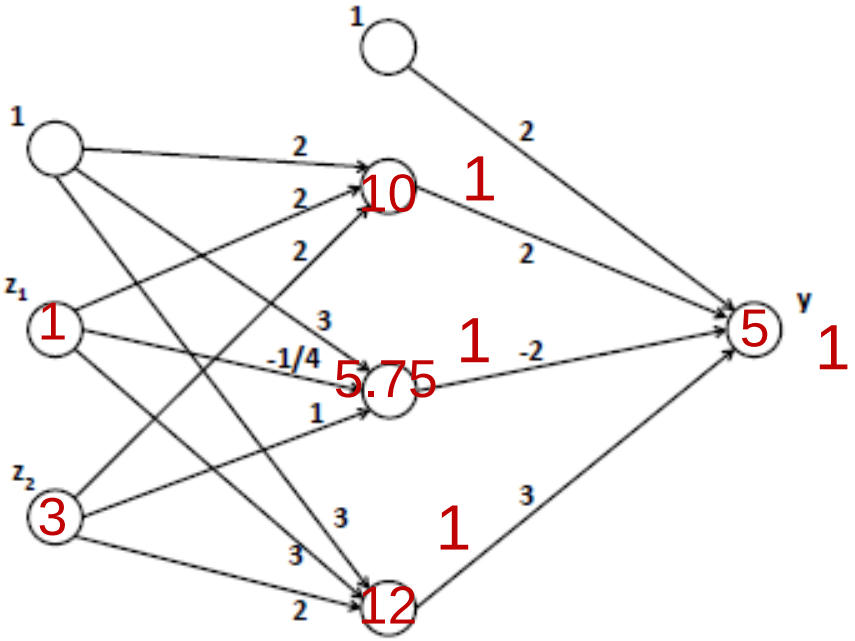
- Initialize with squares or with circles (K=2)
- Different initializations leads to different results
- Case 1: no big differences
- Case 2: a poor initialization can lead to bad result

Exercise 4

Consider the neural network in the figure and assume the activation function $\sigma(x)$ is defined as:

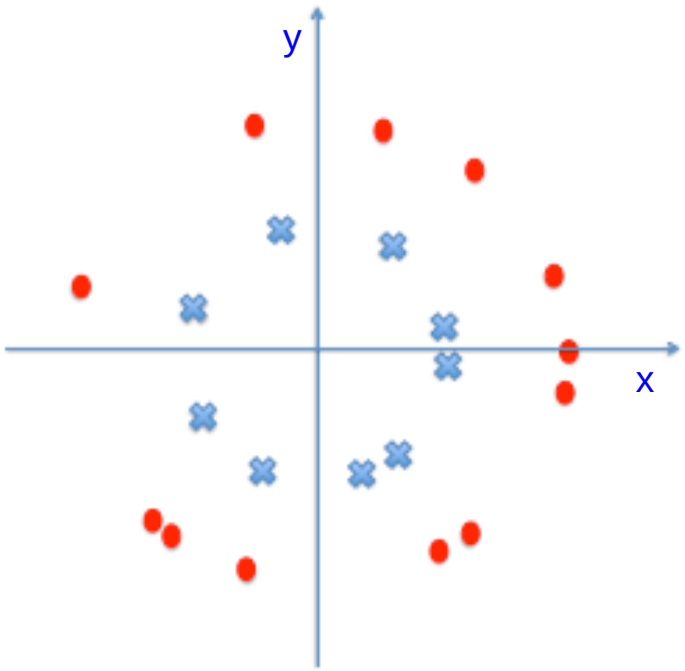
$$\sigma(x) = \begin{cases} 1 & x \geq 1 \\ x & -1 \leq x < 1 \\ -1 & x < -1 \end{cases}$$

Compute the value of the output y when the input **z** is **z** = [1 3]



Exercise 5

1. Introduce the concept of Kernel and its use in SVM for classification.
2. Consider the configuration of training data points (crosses for class 0 and circles for class 1) in the figure below and a scalar function (feature map) $\Phi(\cdot)\colon \mathbb{R}^2 \to \mathbb{R}$ such that the data become linearly separable after the map Φ has been applied.
3. Relate the map Φ to a kernel.

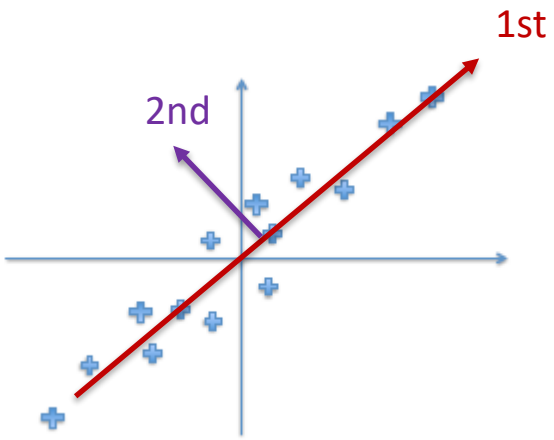


$$\Phi(\cdot)\colon \mathbb{R}^2 \to \mathbb{R} \ : z = x^2 + y^2$$

Polynomial kernel

Exercise 6

1. Let $X = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}]; \quad \boldsymbol{x_i} \in \mathbb{R}^p$ be the data matrix. Introduce the Principal Component Analysis in the context of unsupervised learning.
2. With reference to the figure below draw approximately the first and second right singular vectors of *X*.
3. Describe how PCA can be used in the context of linear regression to reduce the complexity of the model.



Exercise 7

Soft-SVM

1. Describe the linear support vector machine for classification in the case of non linearly separable data.

2. The figure shows the results (separating hyperplane and margin) of linear SVM for binary classification on the data points (in $\mathbb{R}^2$) in the figure, where the class of each point is represented by its shape (triangle or square). Mark with circles the misclassified points and the ones violating the margin and draw the segments which length corresponds to the non-zero slack variables.

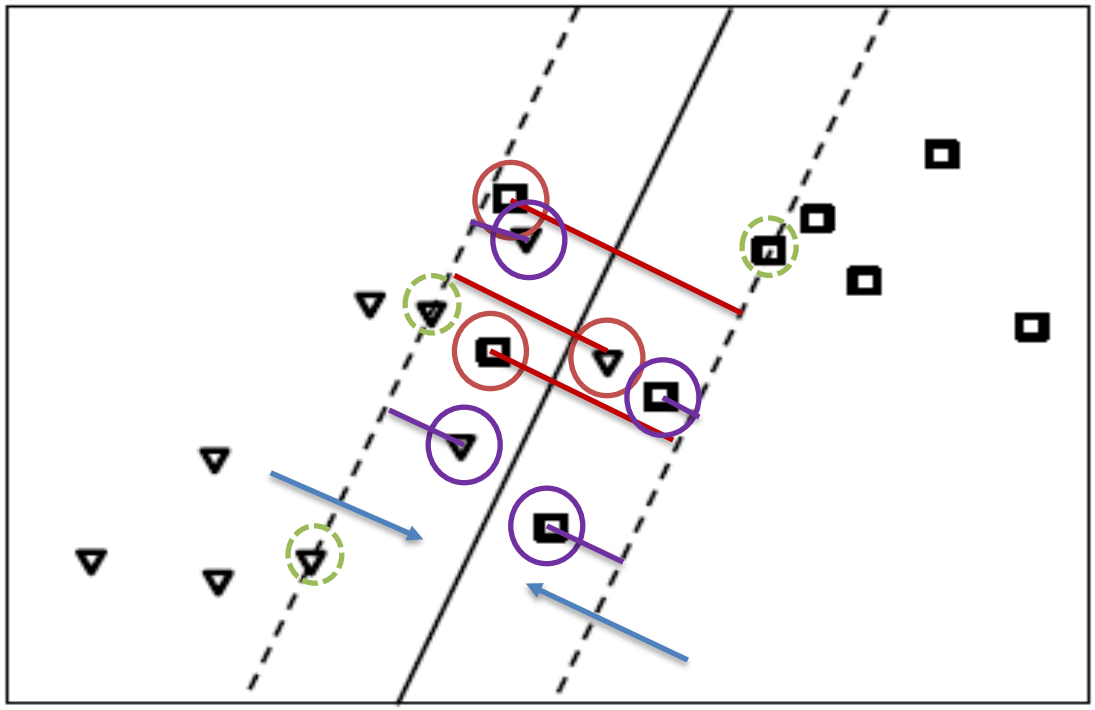
3. Discuss how the solution (margin width and slack variables ξ_i) changes if the value of $C = \frac{1}{\lambda}$ in the object. function $\frac{1}{2}\|\beta\|^2 + C\sum_i \xi_i$ increases.

Red: classification errors ($\xi_i > 1$)

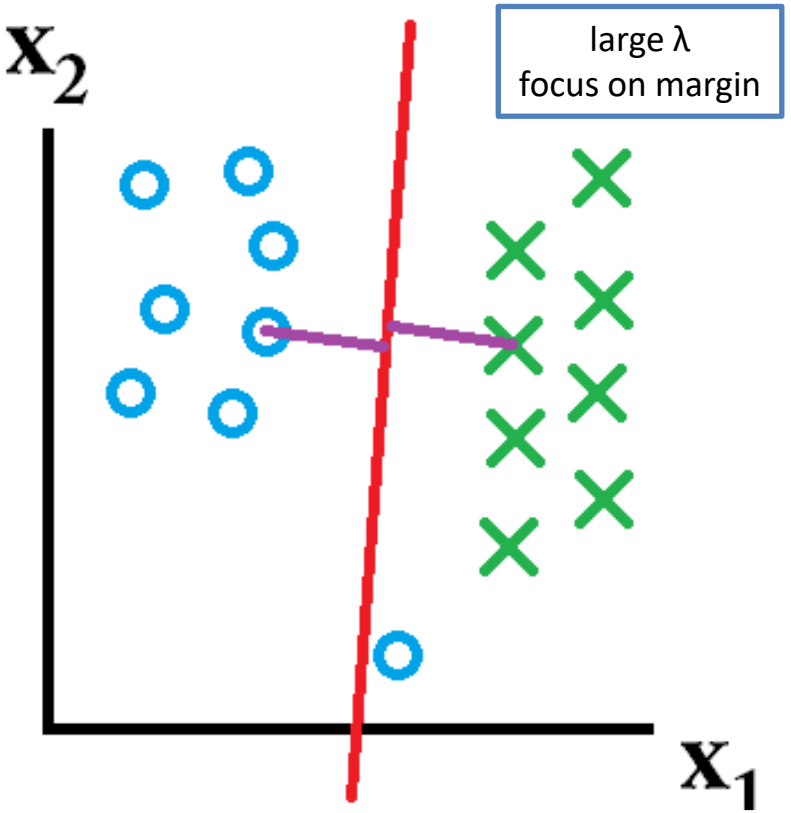
Purple: correct classification but violating margin ($0 < \xi_i < 1$)

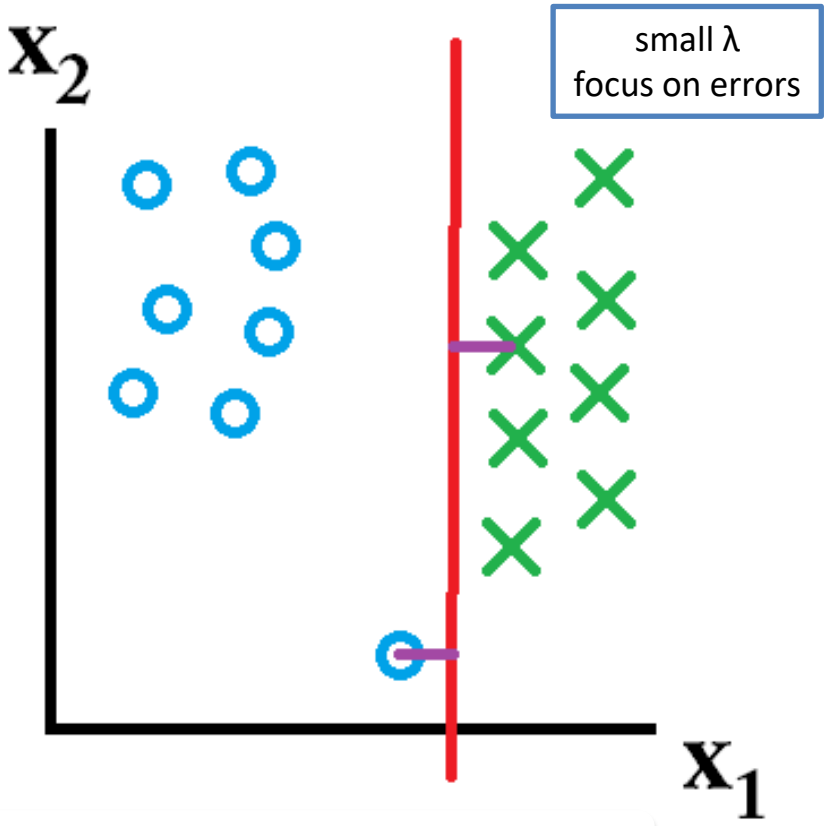
Green: support vectors (→margin)

Larger C: more penalty for misclassified samples → less errors but smaller margin



Small or Large λ ?





$$\min_{\mathbf{w}} \left(\lambda\|\mathbf{w}\|^2 + L_S^{\text{hinge}}(\mathbf{w})\right)$$

C= $\frac{1}{\lambda}$ (opposite effect)

Exercise 9

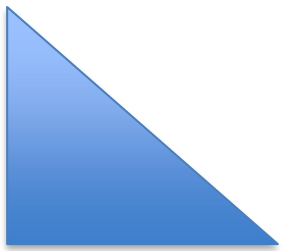
1. Consider an hypothesis class $\mathcal{H}$. What do you need to show in order to demonstrate that *VCdim($\mathcal{H}$)=d ?*

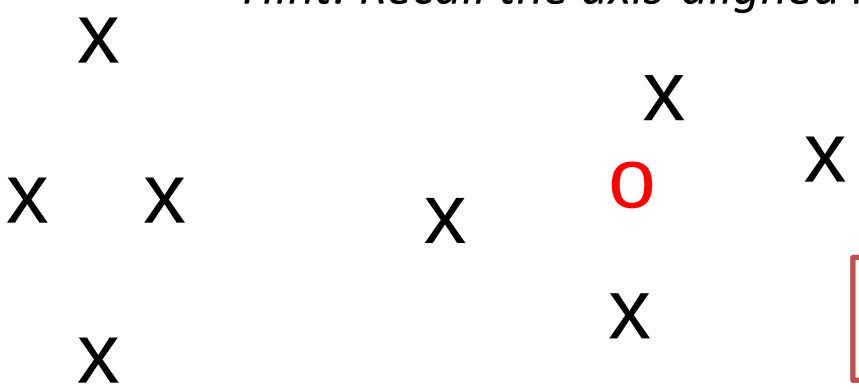
   To show that $VCdim(\mathcal{H}) = d$ we need to show that:

   1. $VCdim(\mathcal{H}) \geq d$: there exists a set C of size *d* which is shattered by $\mathcal{H}$
   2. $VCdim(\mathcal{H}) < (d+1)$: **every** set of size *d + 1* is not shattered by $\mathcal{H}$

2. Consider right triangles in the plane with the sides adjacent to the right angle both parallel to the axes and with the right angle in the lower left corner. Which is the VC-dimension of this family?

Hint: Recall the axis-aligned rectangle demonstration





VCdim($\mathcal{H}$)=4

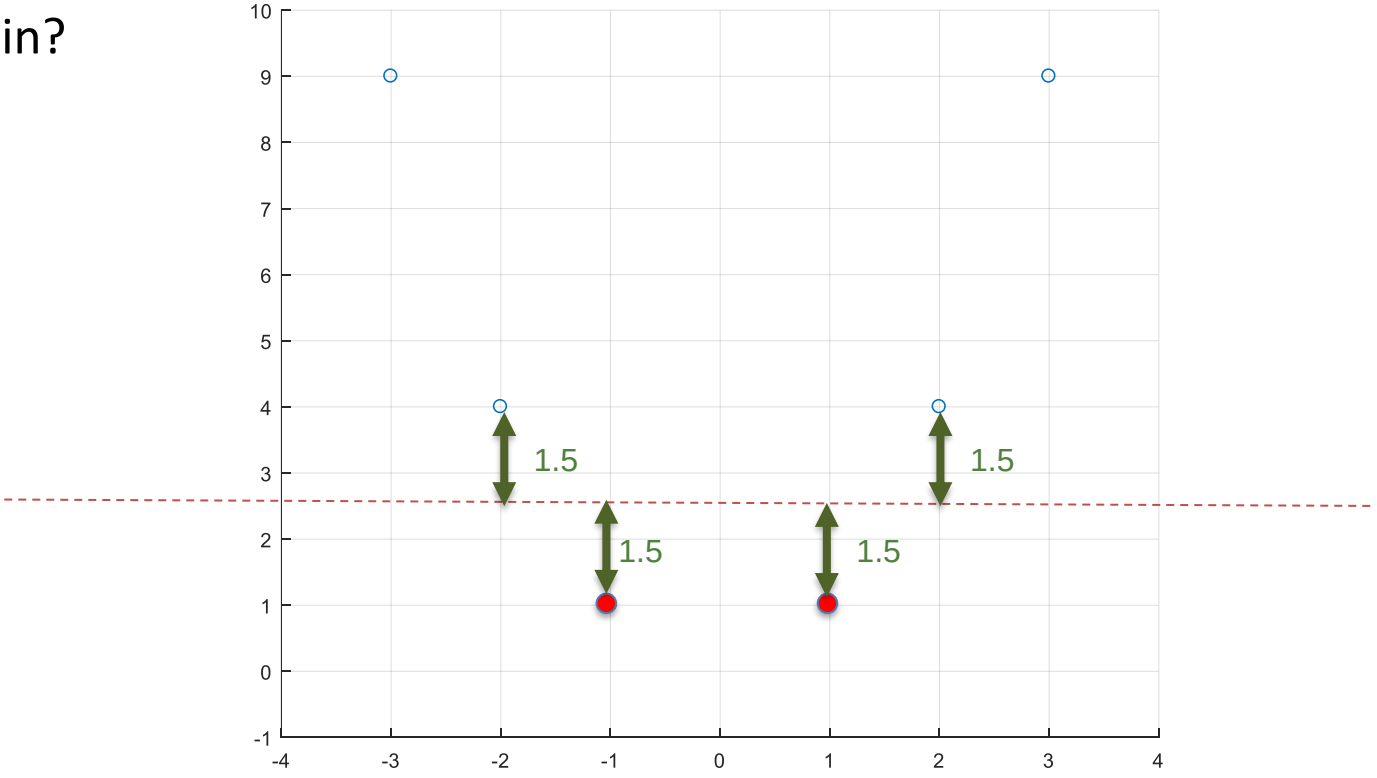
Exercise 10

Consider a dataset with the following six 1-dimensional points (the first element of the couple is the value *x* while the second is the label *y*):

$\{(x_i, y_i)\} = \{(-3, +1), (-2, +1), (-1, -1), (1, -1), (2, +1), (3, +1)\}$

Consider mapping these points to 2 dimensions using the mapping $\phi: x \rightarrow (x, x^2)$

Which is the maximum margin decision boundary for a linear classifier ? Which is the corresponding margin?



Regularization and Overfitting-Underfitting

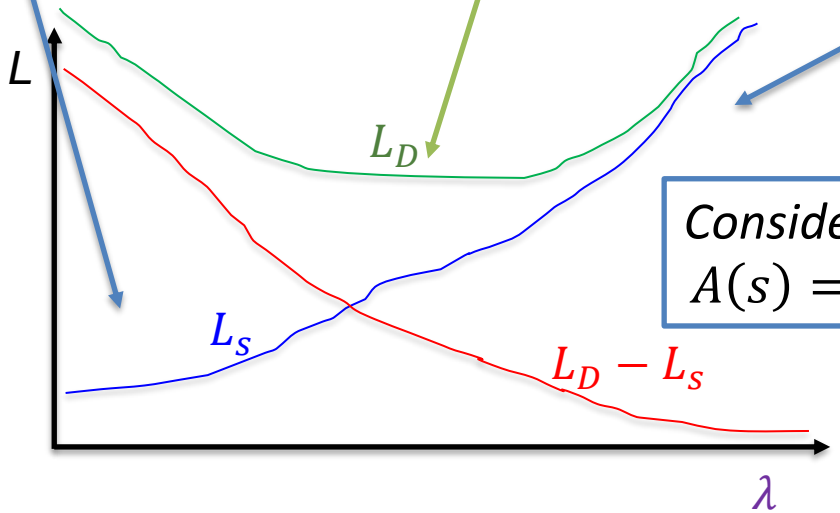
$$E_S[L_D(A(S))] = E_S[L_S(A(S))] + E_S[L_D(A(S)) - L_S(A(S))]$$

- $E_S[L_S(A(S))]$: how well A fits the training set S
- $E_S[L_D(A(S)) - L_S(A(S))]$: measures overfitting, bounded by stability of A

Small λ: focus on training error
Training error L_S : small
Difference $L_D - L_S$: large
Overfitting the training data

Good trade-off in the middle

Large λ: focus on regularization
Training error L_S : large
Difference $L_D - L_S$: small
Underfitting the training data

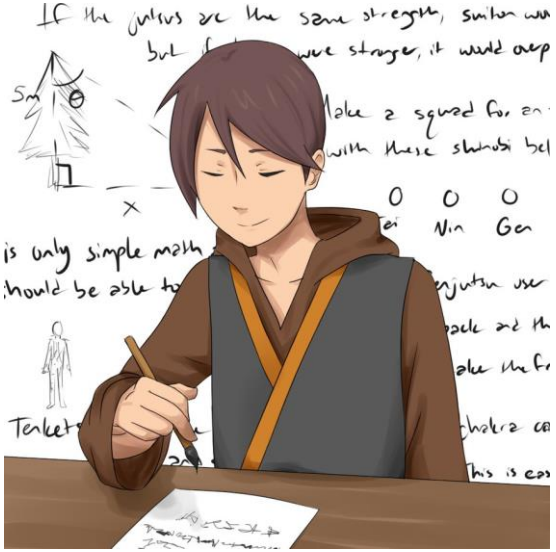


Consider a regularized loss model:
$A(s) = argmin_{\boldsymbol{w}}(L_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2)$

Written Exam

- Written exam in classroom
- No orals; No online exams
- Final mark is the written exam score + the homework score
- 4 exercises, 7 points each (28pts + 4hw)
- Dates for the exams:

1. 30/1/2024 h 10.00 (subscribe before 23/1)
2. 19/2/2024 h 14.30 (subscribe before 12/2)
3. 28/6/2024 h 10.00
4. 12/7/2024 h 10.00
5. 10/9/2024 h 10.00



Check the exam dates
No out-of-session exams