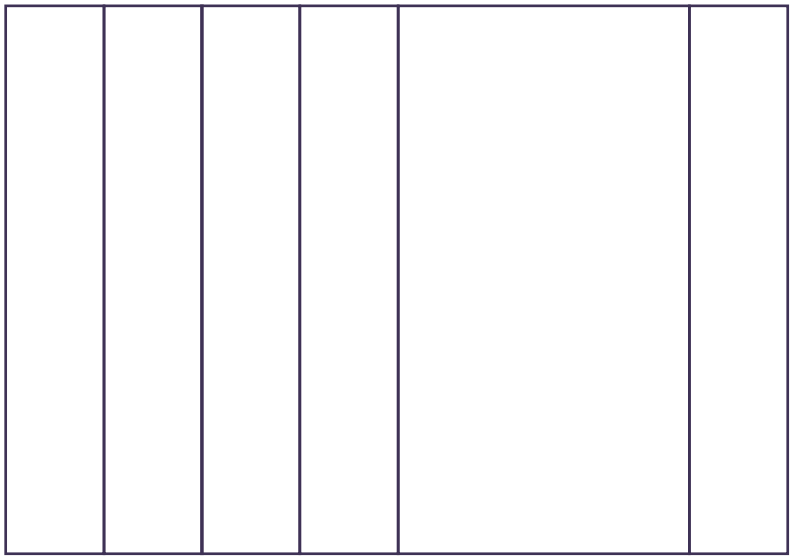
UniGe | MaLGa

Lab 7 – Variable Selection

Our problem today

- How many of the input variables are important?
- And what are the important variables?



$\hat{X}$ $n \times d$

$\hat{Y}$ $n \times 1$

Why is it relevant?

Orthogonal Matching Pursuit

$$I_0 = \emptyset \quad w_0 = [...0...] \quad r_0 =$$

1. Initialize index set, score vector, res
2. Find the single “best” variable
3. Update the index set
4. Update the score vector
5. Update the residual

… repeat 2-5 f … of iterations…

Select the column j that maximizes*

$$\frac{(r_{t-1}^T \hat{X}^j)^2}{||\hat{X}^j||^2}$$

$$I_t = I_{t-1} \cup \{j^*\}$$

$$\hat{X}_t = \hat{X}_{|I_t}$$

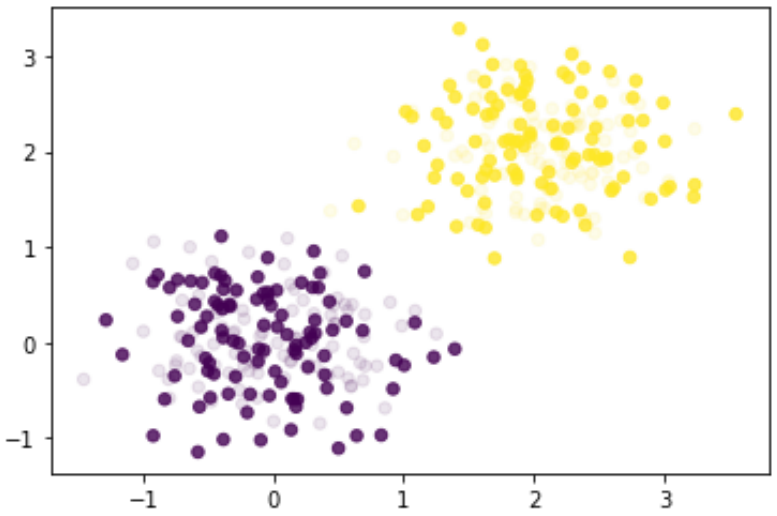
$$w_t = (\hat{X}_t^T \hat{X}_t)^{-1} \hat{X}_t \hat{Y}$$

$$r_t = \hat{Y} - \hat{X} w_t$$

Your objectives today

- Implementing OMP
- Practicing its use on synthetic data for binary classification
- In the notebook you will find a suggested working path

A note on the synthetic data: how to generate an appropriate dataset?



- *We generate synthetic data as usual*
- *We add to each sample a certain amount of random features*

UniGe