Lab 1 – KNN and Cross Validation

K-NN

- It is a local method that follows the idea of predicting the output of a new input reasoning on the outputs of the K-closest points in the input space

Regression

$$\hat{f}_K(x) = \frac{1}{K} \sum_{l \in K_x} y_l$$

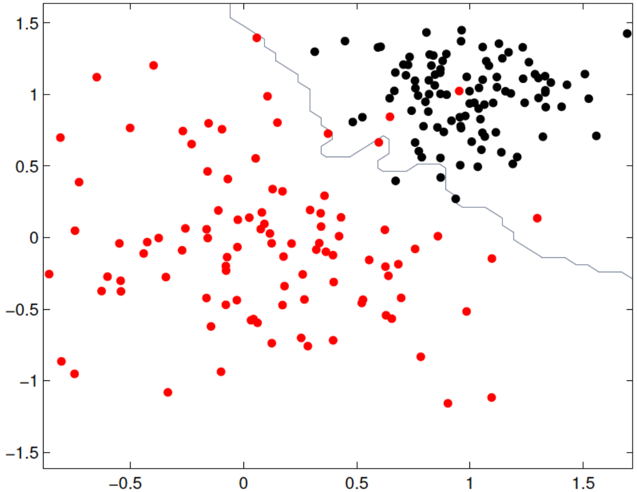
Binary classification

$$\hat{f}_K(x) = sign \sum_{l \in K_x} y_l$$

K-NN: the algorithm

Given x (the new input), S (the training set), and K

- Compute the distances between x and all the points in S
- Sort the distances in increasing order
- Take the outputs of the K closest points
- Compute the predicted output for x according to one of the rules (depending on the task)



Parameter K, noise and number of samples

- K controls the fit and the stability of the function estimated by the KNN algorithm
- We discussed the fact the choice of K influences the “quality” of the estimator, also depending on the amount of noise and the number of samples in the training set
- Today we try and appreciate the effect of its value on the behavior of the K-NN algorithm

Objectives for today

- In the hands-on activity you are asked to provide a (guided) implementation and analysis of K-NN as you change K, the amount of samples in the training set and the amount of noise with specific reference to the properties of fitting and stability
- NOTE
  - To evaluate the fitting, you may consider the prediction ability on the training set
  - To evaluate the stability, we simulate the availability of future data, generating a new test set

The parameter K

- It controls the fit and the stability of the function estimated by the KNN algorithm
- Today we are dealing with the problem of selecting an optimal value for it

Is there an optimal value?

Ideally, we would like to choose K that minimizes the expected error

$$\mathbf{E}_S \mathbf{E}_{x,y}(y - \hat{f}_K(x))^2.$$

You have seen in class how to proceed with a regression problem

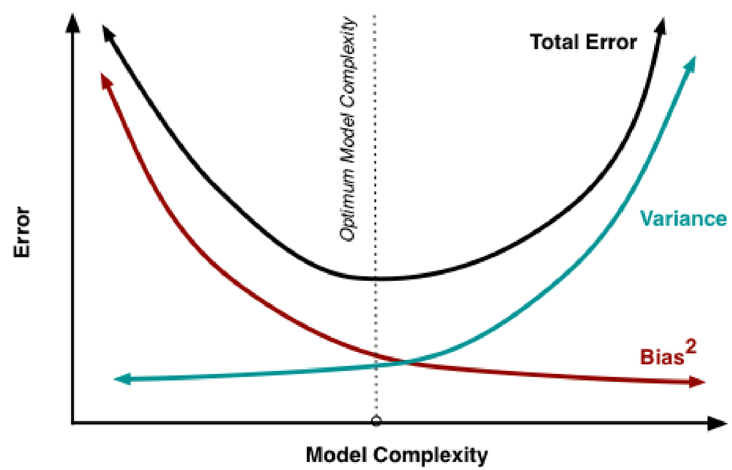
$$y_i = f_*(x_i) + \delta_i, \quad \mathbf{E}\delta_I = 0, \mathbf{E}\delta_i^2 = \sigma^2 \quad i = 1, \ldots, n$$

Is there a optimal value?

After some math...

$$\mathbf{E}_S\mathbf{E}_{y|x}(f_*(x) - \hat{f}_K(x))^2 = \underbrace{(f_*(x) - \mathbf{E}_S\mathbf{E}_{y|x}\hat{f}_K(x))^2}_{Bias} + \underbrace{\mathbf{E}_S\mathbf{E}_{y|x}(\mathbf{E}_{y|x}\hat{f}_K(x) - \hat{f}_K(x))^2}_{Variance}$$

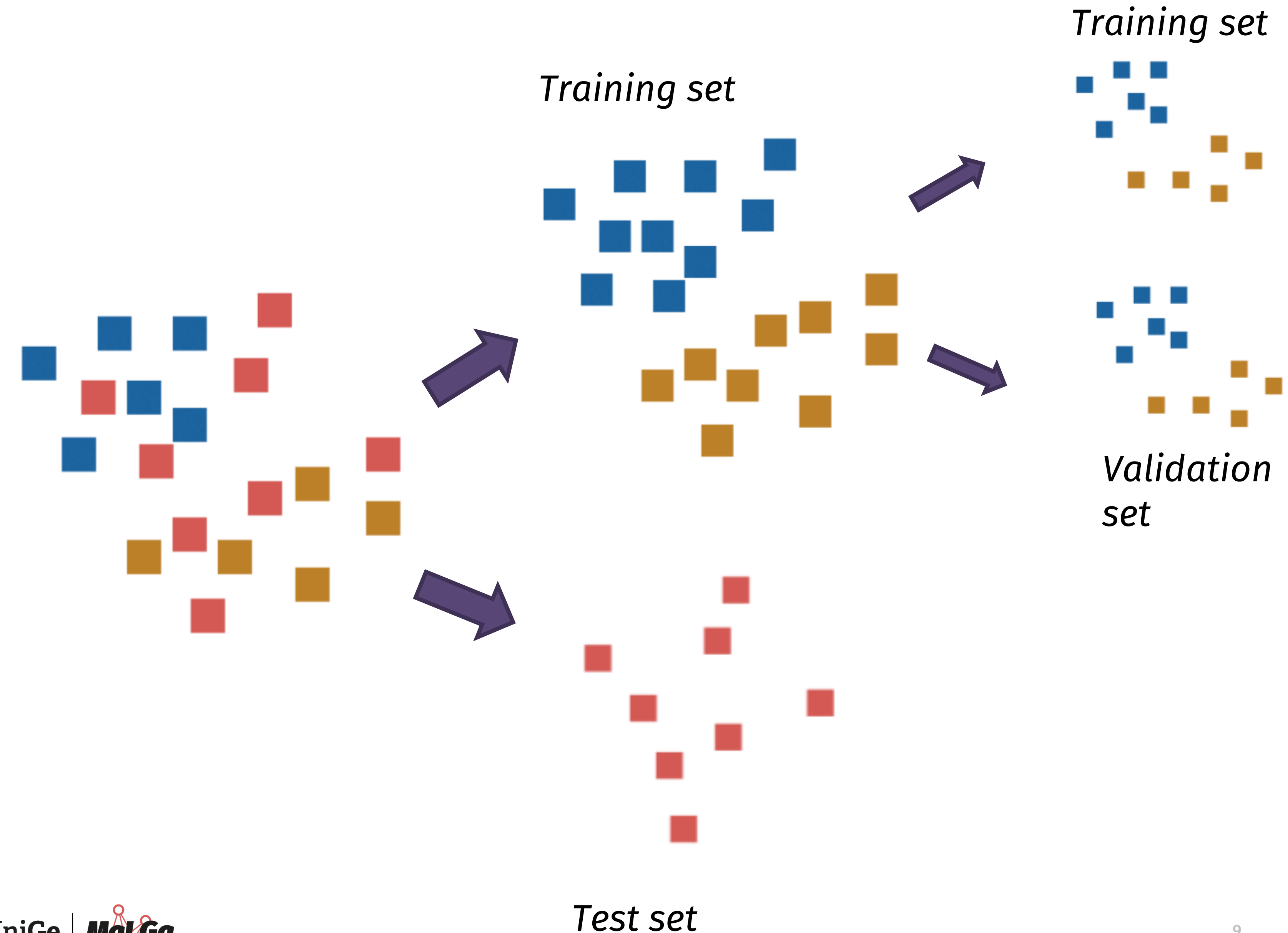
$$(f_*(x) - \frac{1}{K}\sum_{\ell \in K_x} f_*(x_\ell))^2 \qquad \frac{\sigma^2}{K}$$



Is there an optimal value?

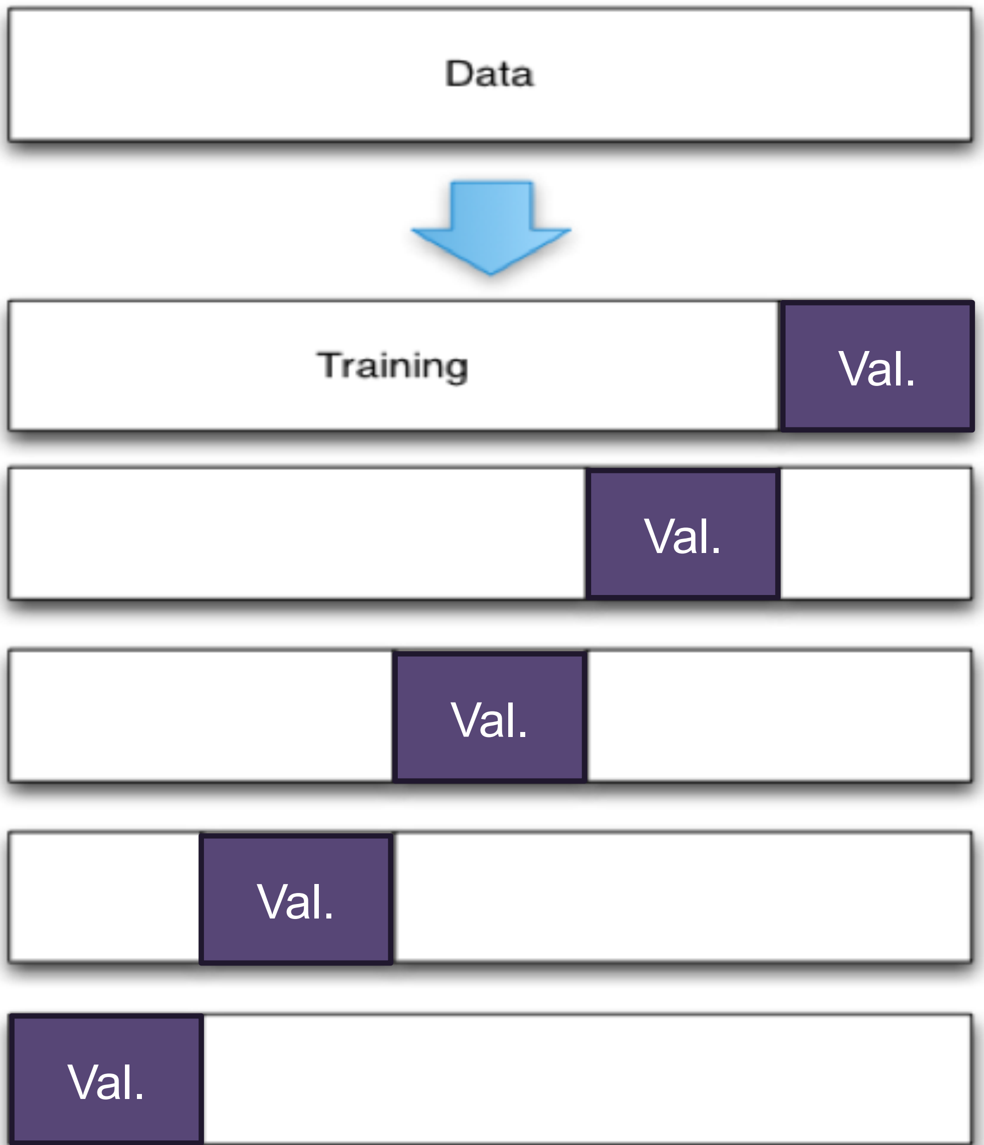
Can it be computed?

Cross Validation



Hold-Out Cross Validation

K-Fold Cross Validation



Your objectives today

Practicing the selection of an appropriate value for the K parameter using Cross Validation, by doing the following

- Pretending to have the test set (and in fact you have it in these examples) and have a look to the trend of the error
- Applying k-Fold Cross Validation

UniGe

MaLGa