Context

- Linear Regression
- Logistic Regression

} parametric

$y = Xw + \epsilon$

$\ln\left(\frac{p}{1-p}\right) = Xw + \epsilon$

→ parameters

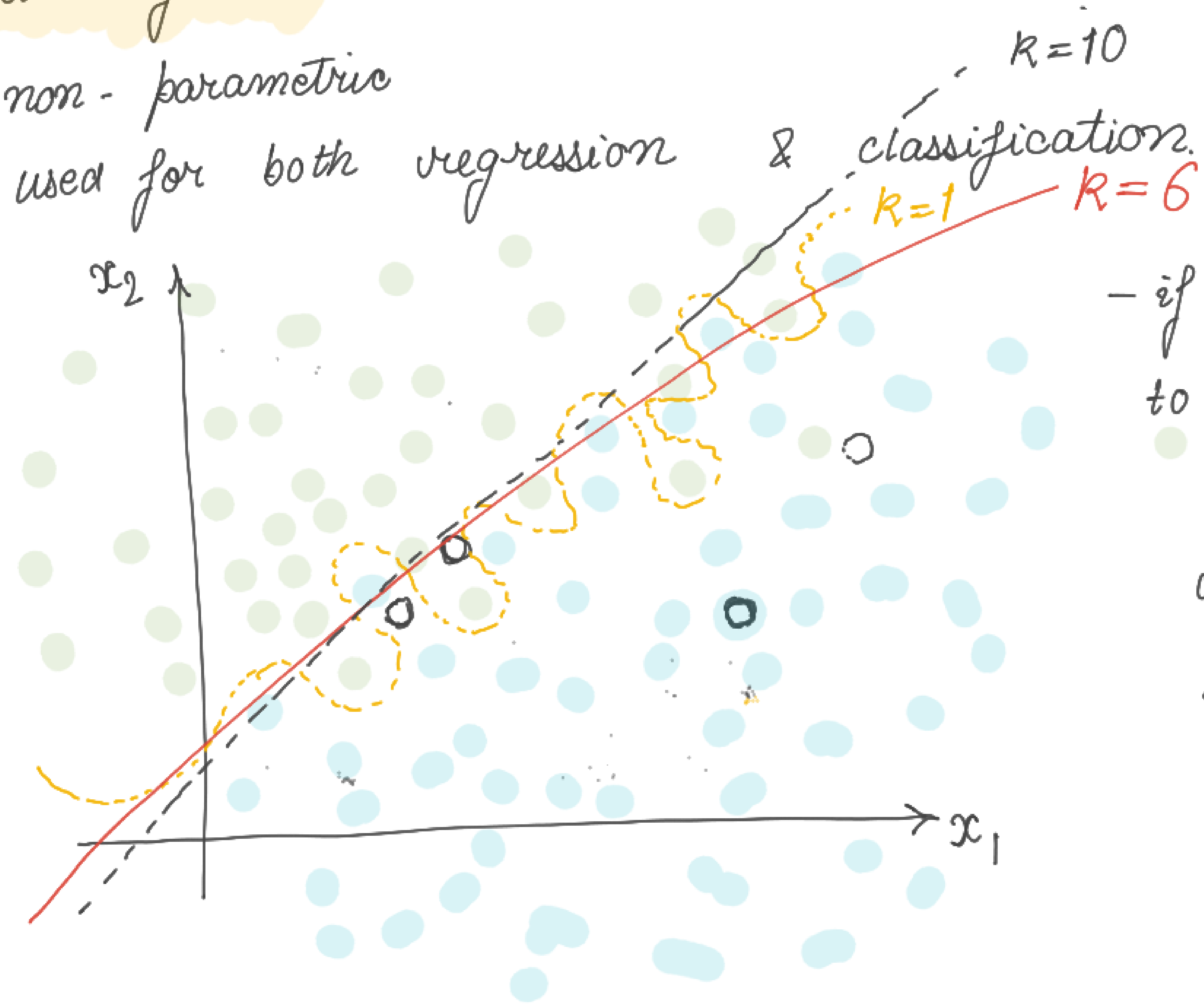
- Decision Trees
- Random Forest
- K Nearest Neighbours

} non-parametric → no statistical qty is being estimated

(Can be used for Regression & Classification)

K-Nearest Neighbours

- non-parametric
- used for both regression & classification.



- if you want your pt to be influenced by only a small number of closest points; then value of k should be small.

k=1 ⇒

k=3 ⇒

KNN- classifier

- smaller value of k
  - ⇒ complex decision boundary
  - ⇒ overfit model
  - ⇒ more susceptible to outliers.
- larger value of k
  - ⇒ simpler (linear-ish) decision boundary
  - ⇒ underfit model
  - ⇒ ignores subtle patterns present in data.

⇒ How do I select k?

How to select k ?

- Divide DS → Train, Test
- For k = 1 → k = 20 ;

  ⇒ Calculate evaluation metrics for each value of k

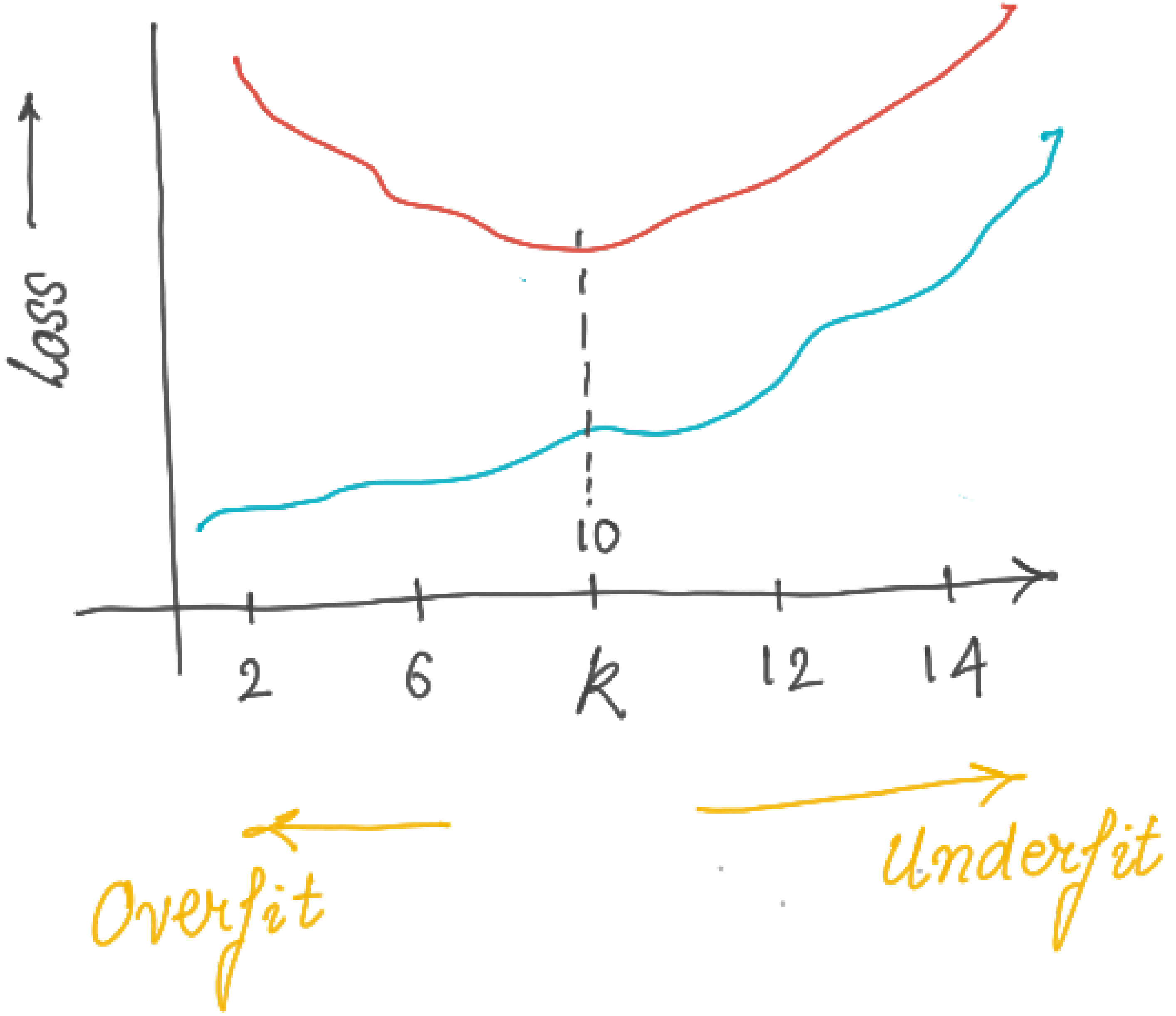
  ↳ accuracy, recall, precision, misclassif. rate

  ↳ mape, mae, rmse

  for train & test separately.

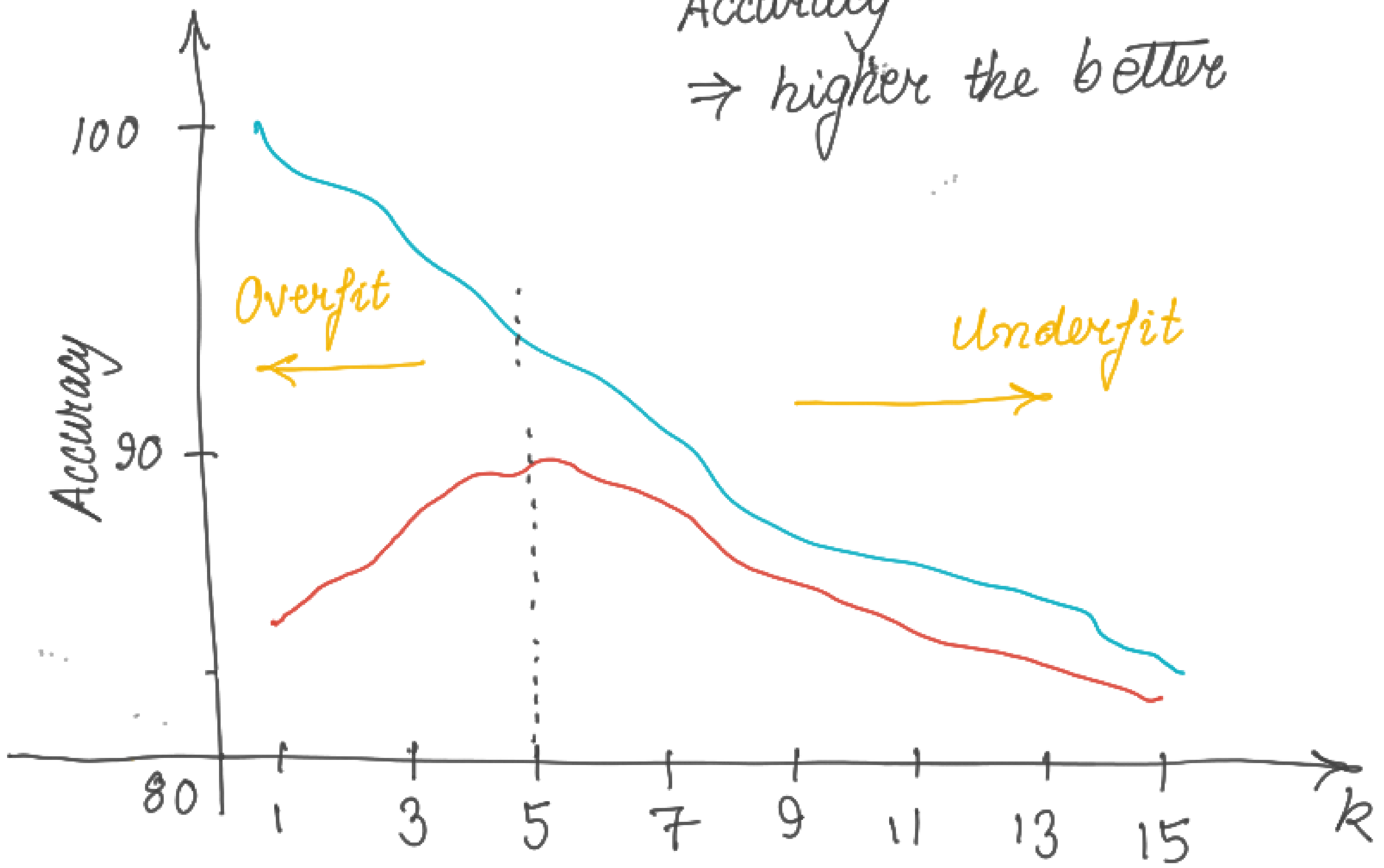
- Choose that value of k that gives you best eval. metric for test.

— Test
— Train

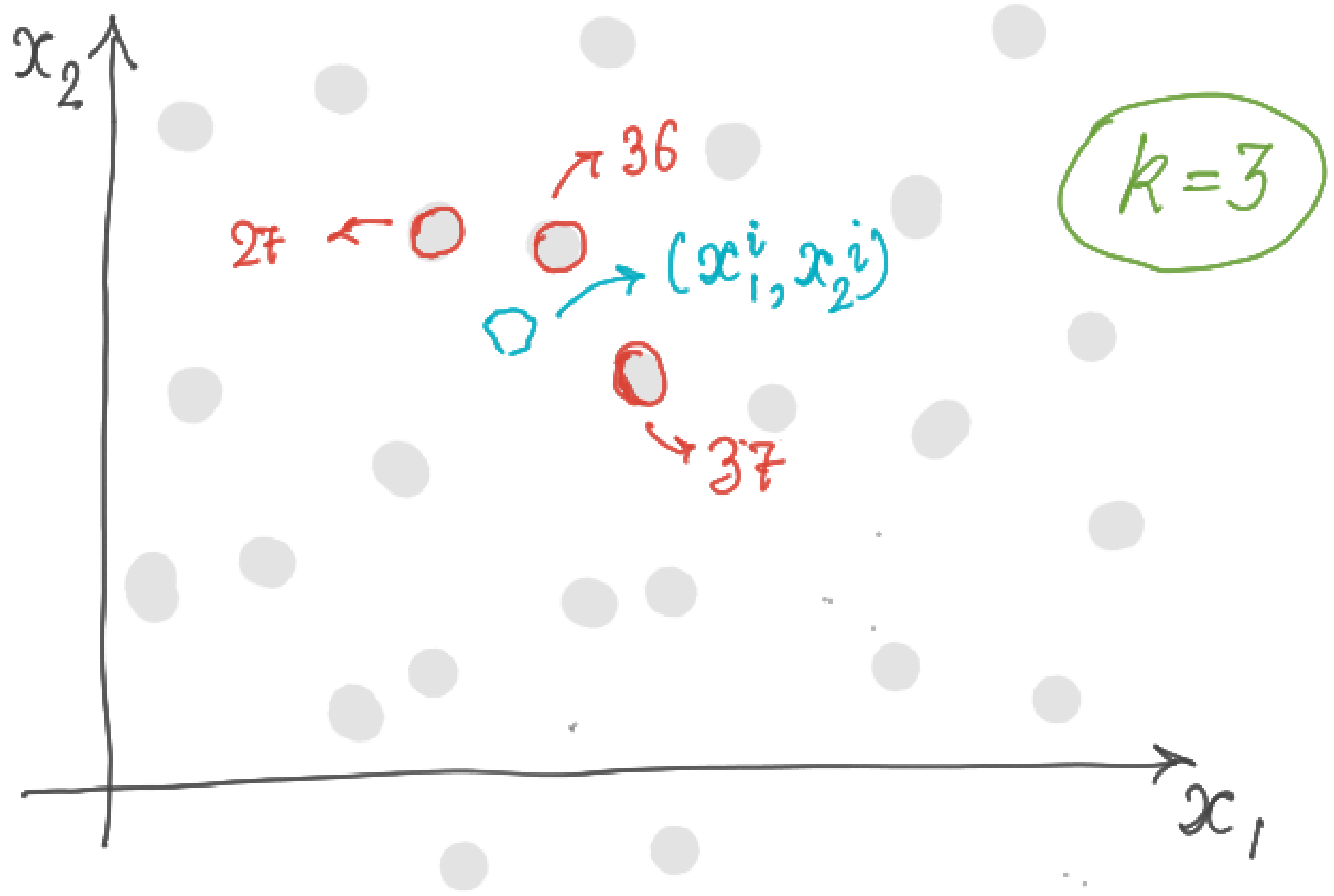


Loss
⇒ Lower the better

Accuracy
⇒ higher the better



KNN for Regression



$$\begin{array}{c|ccc}
 & x_1 & x_2 & y \\
1 & & & \\
2 & & & \\
3 & & & \\
\vdots & & & \\
n & & & \\
 & \multicolumn{2}{c}{\underbrace{\hspace{3em}}_{X}} & \underbrace{\hspace{1em}}_{y}
\end{array}$$

$(x_1^i, x_2^i) \rightarrow y^i$?

$$y^i = \frac{37+36+27}{3} = 33.\overline{3}$$

KNN for Regression

- smaller k
  - ⇒ predictions influenced by outliers
  - ⇒ overfit model

- larger k
  - ⇒ predictions insensitive to subtle patterns in dataset
  - ⇒ underfit model

We choose k as before ⇒ best eval. metric value of test dataset

Misc

Minkowski distance $= \left\{ \sum_{i=1}^{n} |x_i - y_i|^p \right\}^{1/p}$

$D_p(x, y)$

$\downarrow$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$D_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

= Euclidean distance

$$D_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

= Manhattan distance