K-nearest neightbours.

* Euclidean distance.

$$D_{Euclidean}(x_i, x_j) = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2} .$$

* Manhattan distance (block distance).

$$D_{Manhattan}(x_i, x_j) = \sum_{s=1}^{p} |x_{is} - x_{js}| .$$

* Chebyshev distance.

$$D_{Chebyshev}(x_i, x_j) = \max_{s=1,\ldots,p} |x_{is} - x_{js}| .$$

* Canberra distance.

$$D_{Canberra}(x_i, x_j) = \sum_{s=1}^{p} \frac{|x_{is} - x_{js}|}{|x_{is}| + |x_{js}|} .$$

Note:

* Calculate distance after standardising the variables so the variables are in a comparable scale.

Notation

$\mathcal{N}_i^k$ = a set of index of k observations with the smallest distance to observation i.

$$\mathcal{N}_i^k = \{ j \mid D(x_i, x_j) < \varepsilon_k \} .$$

* Predict with KNN.

Assume that we want to predict a new record with predictor values x_{new}.

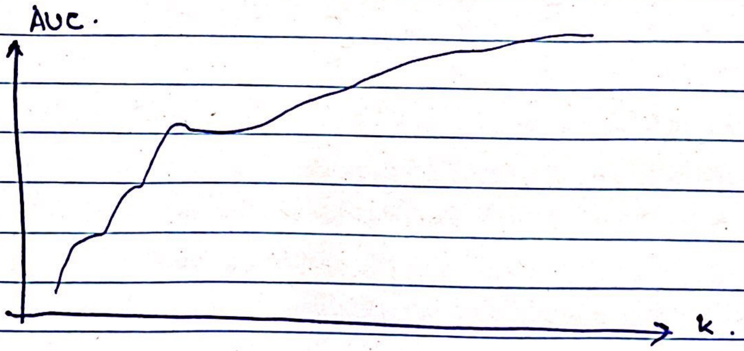
1. Find the k-nearest neighbours $\mathcal{N}_{new}^{k}$.

2. In classification problems, count how many neighbours belong to each class and the predicted class is the one which have majority votes.

- If the response is numerical, the predicted value is equal to the average of the outcome of the neighbours:

$$\hat{y}_{new} = f(x_{new}) = \frac{1}{k} \sum_{j \in \mathcal{N}_{new}^{k}} y_j .$$

* Selecting k.



* You can use elbow method to find an optimal value for k.