# Machine Learning and Statistical Learning

## K-Nearest Neighbors

Ewen Gallic
ewen.gallic@gmail.com

MASTER in Economics - Track EBDS - 2nd Year



## K-Nearest Neighbors

When facing real data, we do not know the conditional distribution of $Y$ given $X$. Hence, computing the Bayes classifier is not possible.

Here, we will look at a classifier that estimates the conditional distribution of $Y$ given $X$, namely the **K-nearest neighbors** (KNN) classifier.

## K-Nearest Neighbors

The basic idea of the KNN classifier is, as follows:

- from a given positive integer $K$ and a test observation $x_0$, identify the $K$ points in the training data that are **closest** to $x_0$, represented by $\mathcal{N}_0$
- estimate the conditional probability for class $k$ as the fraction of points in $\mathcal{N}_0$ whose response values equal $k$:

$$\mathbb{P}(Y = k \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{1}_{(y_i = k)}. \tag{1}$$

- once the conditional probabilities for each of the $K$ classes are estimated, apply Bayes rule and assign $x_0$ to the class with the highest probability.

## Example

Let us consider again a response variable that can take two values: either “blue” or “orange”.

For the example, we draw 100 points in a unit square and randomly assign them a class.

Then, we consider a point at coordinates $(0.75, 0.5)$ and try to predict the class for this point using a KNN classifier.

We vary the number of nearest neighbors to consider: $K = \{3, 5, 10\}$.

# Example

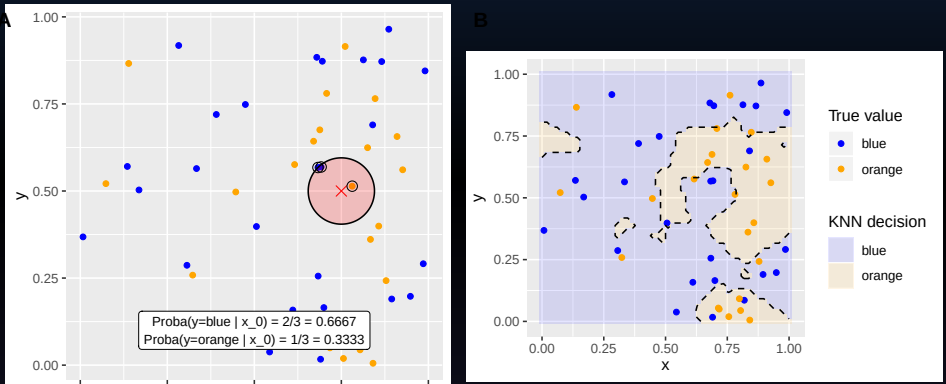


Figure 1: KNN approach, with $K = 3$.

# Example

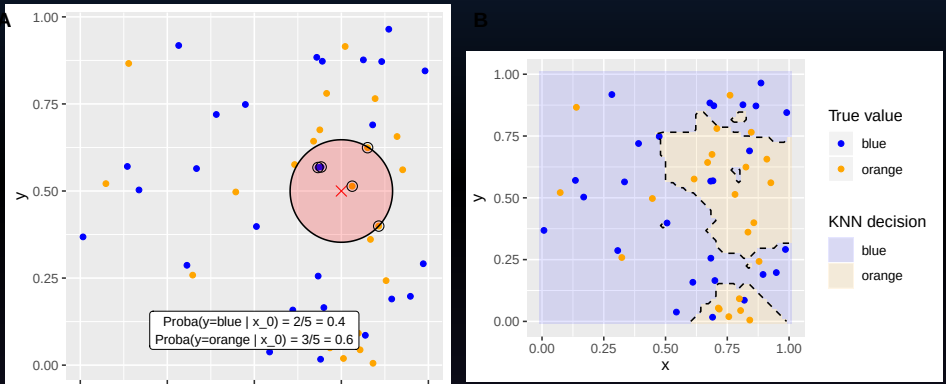


Figure 2: KNN approach, with $K = 5$.

# Example

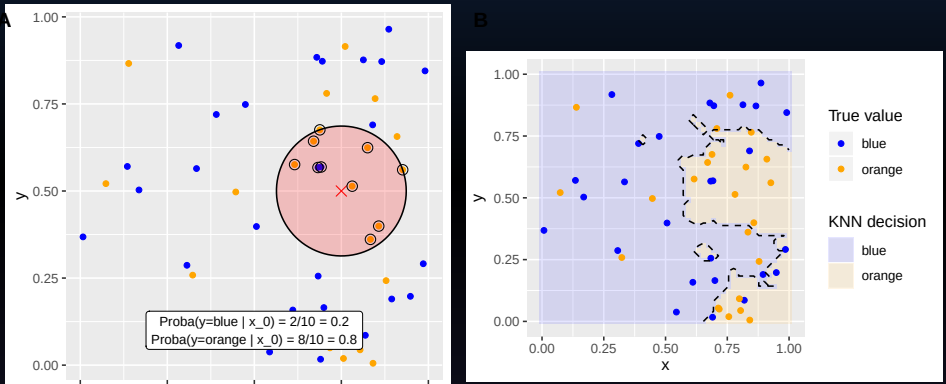


Figure 3: KNN approach, with $K = 10$.

## Exercise

Lab exercise.