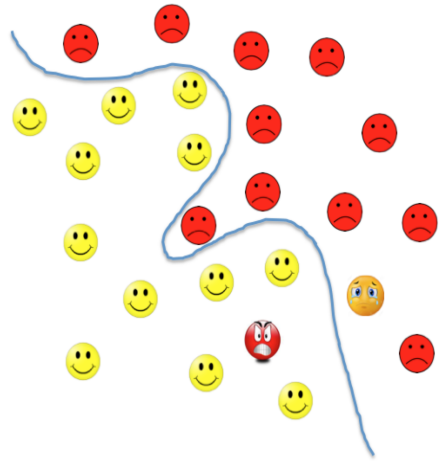
# Introduction to machine learning
# Non Parametric classification
## K Nearest Neighbors (KNN)
## Classification Trees

Mathilde Mougeot

ENSIIE

2018

# K Nearest Neighbors



# K Nearest Neighbors

**Setup**

- $X$ quantitative features, $X \in R^p$
- Y qualitative target with $Q$ modalities $Y \in \{G_1, G_2, ..G_Q\}$
- $n$ observations,

**Data set**

- $\mathcal{D}_n = \{(x_i, y_i), i = 1..n, x_i \in R^p,\ y_i \in \mathcal{X}\}$

**The KNN estimation of $x$ is given by**

$$\hat{f}_K(x) = G_{q_0} \text{ with } q_0 = \underset{q=1..Q}{\arg\max} \sum_{y_i \in V_K(x)} \mathbb{1}_{y_i = G_q}$$

$V_K(x)$ is defined by the $K$ nearest neighbors of $x$.

# K Nearest Neighbors

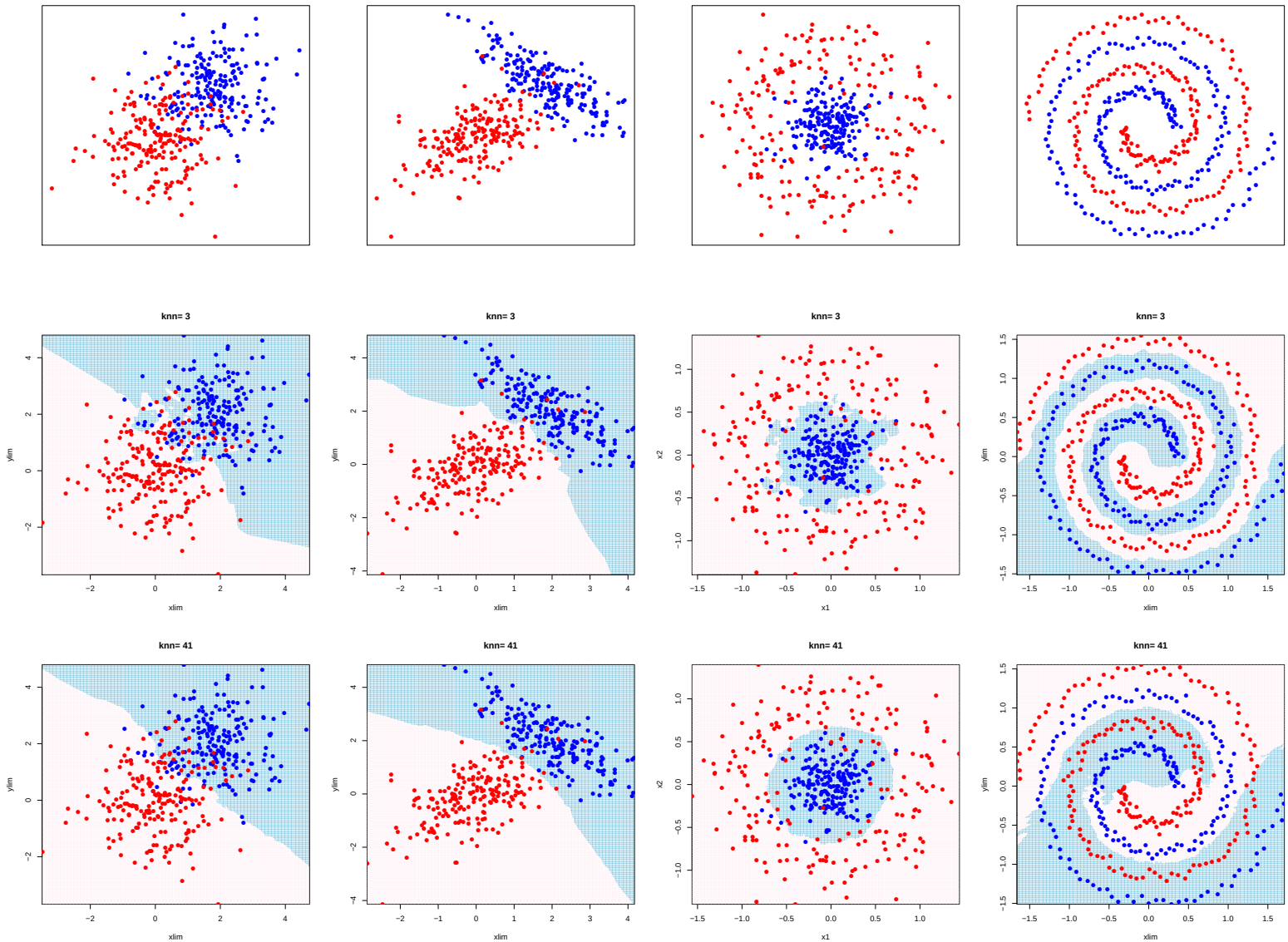
**The KNN estimation of $x$ is given by**

$$\hat{f}_K(x) = G_{q_0} \text{ with } q_0 = \underset{q=1..Q}{\arg\max} \sum_{y_i \in V_K(x)} \mathbb{1}_{y_i = G_q}$$

$V_K(x)$ is defined by the $K$ nearest neighbors of $x$.

**Remark.**
Different values of $K$ lead to different classifiers.

# K Nearest Neighbors Classifier



# K Nearest Neighbors

- $K$ ("K"NN) is a key parameter, which has a direct impact on $\hat{f}_K(x)$
- To favor the Generalization, the value of $K$ is usually chosen by cross-validation
- Impact of the metric
  - The $K$ nearest Neighbors are computed given a metric.
  - Each feature $X^j$, $j = 1..p$ has its own unit.
  - The $\ell_2$ norm is usually used to compute the distance between $x$ and $x_j$ :

$$d(x, x_j)^2 = ||x - x_j||^2 = \sum_{\ell} (x_\ell - x_\ell^j)^2$$

  $\rightarrow$ The inputs are in general renormalized to erase the influence of the units