Analogizers

Analogizers are **classification algorithms** that **compare the new instance** to be classified with **instances in the training set**. The most famous analogizer is the **k-Nearest Neighbors** algorithm.

KNN (k-Nearest Neighbors) Algorithm

- Records are represented as **points** in the Euclidean space;
- **Training** consists of **storing the records** in the training set;
- **Classification** consists of **finding the k nearest neighbors** of the new instance to be classified and **assigning the most frequent class** among them to the new instance.

The KNN can be represented using a **Voronoi diagram**:

- Each point in the diagram represents a record in the training set;
- The **region** of each point is the **set of points** that are **closer to that point** than to any other point in the diagram.
- The **decision boundary** is the **line** that separates the regions of two classes.

Choosing the Value of k

- Usually, **k is odd** to avoid ties;
- Choose the number of neighbors **experimentally** - start with one and increase it until the accuracy stops improving.

Comparing Records

- How to classify an object with **equally distant neighbors**?
  - If the neighbors are all from the same class, the object is classified as that class;
  - Otherwise, the object cannot be classified, and if this happens with a considerable frequency, the similarity function should be changed.
- Other problem if the **different scales** of the attributes;
  - **Normalize** the attributes to the same scale;
- Other problem are **correlated variables** - **redundant** information;
  - **Remove** the redundant variables.