ML INFOSEC
5: k Nearest Neighbors

April 17, 2019

Reminder 1: Euclidean distance in $\mathbb{R}^n$

For $a = (a_1, ..., a_n), b = (b_1, ..., b_n) \in \mathbb{R}^n$

$$d(a, b) = \left(\sum_{j=1}^{n} (a_j - b_j)^2 \right)^{1/2}$$

is called the Euclidean distance between a and b.

Reminder 2: Metric spaces

Let X be non-empty set. A function $d : X \times X \to [0, \infty)$ such that

$$\begin{aligned} d(a,a) &= 0 \quad \forall a \in X, \\ d(a,b) &> 0 \quad \forall a,b \in X,\; a \neq b, \\ d(a,b) &= d(b,a) \quad \forall a,b \in X, \\ d(a,c) &\leq d(a,b) + d(b,c) \quad \forall a,b,c \in X. \end{aligned}$$

is called a metric on X and (X, d) a metric space. The Euclidean distance is a metric.

The setting

Let $\mathcal{A}_1$, $\mathcal{A}_2$, ..., $\mathcal{A}_n$ be sets of real-valued **attributes**, w.l.o.g. $= \mathbb{R}$, C a finite set of **classes** classes and $T \subset \mathbb{R}^N$ a finite set of **instances**. Moreover, let

$$F : T \to C$$

a function, i.e. each instance x is classified as class $F(x)$.

The kNN classification algorithm

Let $x_1, ..., x_k$ be the k instances in T that are nearest to a with respect to d.

kNN Classifier

For $a \in \mathbb{R}^n$, the k Nearest Neighbor classifier is given by

$$\begin{aligned} c_{kNN}(a) &= \mathrm{argmax}_{c \in C} \#\{j \mid 1 \leq j \leq k, F(x_j) = c\} \\ &= \mathrm{argmax}_{c \in C} \sum_{j=1}^{k} \delta(F(x_j), c) \end{aligned}$$

where

$$\delta(a, b) = \begin{cases} 1, & a = b \\ 0, & else \end{cases}$$

Weighted and Distance-weighted NN

Weighted kNN Classifier

For $a \in \mathbb{R}^n$ and weights $w_1, \ldots, w_k > 0$, the weighted k Nearest Neighbor classifier is given by

$$c_{kNN,w}(a) = \operatorname{argmax}_{c \in C} \sum_{j=1}^{k} w_i \delta(F(x_j), c)$$

If

$$w_j = \frac{1}{d(x, x_j)},$$

it is called distance weighted NN.