Explaining Hubness by the Expected k-occurrences

Summary

- A dataset is said to present **hubness** if there are points that exhibit a large k-**occurrence**.
- Assuming that the dataset is sampled from i.i.d. random variables, the expected k-occurrence has a close form.
- We present preliminary ideas on how this expectation can be used to explain hubness.

Dataset model

Given a dataset $\mathcal{X}$ we define

- The k-**neighbourhood relationship** for $x, c \in \mathcal{X}$

$$Neigh^k_{\mathcal{X}}(x;c) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } x \text{ is in the } k\text{-nn} \\ 0 & \text{otherwise} \end{cases}$$

- The k-**occurrence** of a point $x \in \mathcal{X}$

$$Occ^k_{\mathcal{X}}(x) \overset{\text{def}}{=} |\{c \in \mathcal{X} : Neigh^k_{\mathcal{X}}(x;c) = 1\}|$$

Closeness value

- The **closeness value** of a point x with respect to a reference point c is defined as

$$\mathcal{C}(x;c) = \int_{B^c(r_{xc})} f_X(y)\, dy$$

where $B^c(r_{xc})$ denotes the ball centered on c with radius the distance between x and c.

Example

We aim to compute the expected 30-occurrences of points in a set of 1000 points sampled uniformly in the square. The empirical and theoretical computations agree.

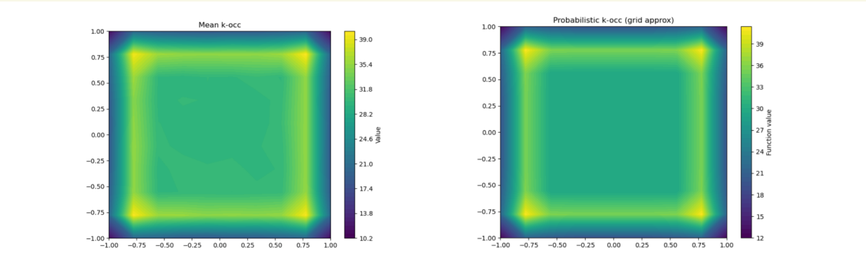

Figure: Empirical mean k-occurrence (left) and the theoretical computation (right). The results closely match.

Probabilistic model

Given a pdf f_X we define

- The probability of the k-neighbourhood relation

$$\mathcal{N}^{k:N}(x;c) \overset{\text{def}}{=} \mathcal{P}_{x_2 \dots x_N \overset{\text{i.i.d.}}{\sim} f_X}\left(Neigh^k_{\{x,c,x_2 \dots x_N\}}(x;c) = 1\right)$$

- The expected value of the k-occurrence

$$O^{k:N}(x) \overset{\text{def}}{=} \mathbb{E}_{x_1 \dots x_N \overset{\text{i.i.d.}}{\sim} f_X}\left[Occ^k_{\{x,x_1 \dots x_N\}}(x)\right]$$

in a dataset sampled from i.i.d. random variables distributed as f_X.

Computation of $O^{k:N}(x)$

Based on the closeness value, we can compute:

- The probability of the k-neighbourhood relationship:

$$\mathcal{N}^{k:N}(x;c) = \sum_{i=1}^{k} \binom{N-1}{i-1} \mathcal{C}(x;c)^{i-1} \left(1 - \mathcal{C}(x;c)\right)^{N-i}$$

- The expectation of k-occurrence:

$$O^{k:N}(x) = N \int_{\mathbb{R}^D} \mathcal{N}^{k:N}(x;c)\, f_X(c)\, dc$$

Findings about Hubness

- The total number of points (N) and of neighbours (k) are necessary in order to define hubs.
- The closeness value relates hubness with dimensionality and change of gradients in the pdf (such as borders of the support).
- The centrality of a point does not necessarily relate with a higher k-occurrence.

Victor Reyes

Victor.Reyesmartin@unige.ch

University of Geneva - SISAP 2025