Big Data Analytics

Métriques en Non-Supervisé

Métriques en Non-Supervisé

$$coût = \sum_{i} \sum_{j} \delta_{i,j} |x_j - \mu_i|$$

où $\delta_{i,j}$ vaut 1 si le cluster μ_i est le plus proche du point \mathbf{x}_j , 0 sinon

1

Métrique : Silouhette

Points $x = \{x_1, \dots, x_n\}$, Clusters $\mu = \{\mu_1, \dots, \mu_k\}$.

$$a(x_i) = \frac{1}{\#\mu_i - 1} \sum_j |x_i - x_j|$$

$$b(x_i) = \min_{i \neq j} \frac{1}{\#\mu_j} \sum_j |x_i - x_j|$$

où:

 $\#\mu_i$ est le nombre d'éléments de x dans le cluster μ_i L'ensembe d'indice j ne représente que ceux des points appartenant au cluster μ_i

 $a(x_i)$: distance moyenne aux autres points du cluster contenant x_i

 $b(x_i)$: distance moyenne aux points du cluster le plus proche

Métrique : Silouhette

$$s_i = rac{b_i - a_i}{\max\{a_i, b_i\}} \;\;,\;\; s_i = \left\{ egin{array}{ll} 1 - a_i/b_i & ext{if } a_i < b_i \ 0 & ext{if } a_i = b_i \ b_i/a_i - 1 & ext{if } a_i > b_i \end{array}
ight.$$

donc
$$s_i \in [-1,1]$$

 $s_i \approx 1 \iff x_i$ bien clusterisé $s_i \approx 0 \iff x_i$ au bord de 2 clusters $s_i \approx -1 \iff x_i$ mal clusterisé

3

Métrique : etc

- Calinski-Harabaz index
- Davies-Bouldin Index
- ...