



Clustering

Evaluation Metrics

- Withholding ground truth
- Unknown ground truth

Withholding Ground Truth

Problem Definition

We define a problem having k clusters and n samples with k pre-determined classes. C is the grouping of samples based on their classes and P is the grouping based on their predicted clusters.

Contingency Matrix

$X \backslash Y$	Y_1	Y_2	\dots	Y_s	Sums
X_1	n_{11}	n_{12}	\dots	n_{1s}	a_1
X_2	n_{21}	n_{22}	\dots	n_{2s}	a_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
X_r	n_{r1}	n_{r2}	\dots	n_{rs}	a_r
Sums	b_1	b_2	\dots	b_s	

Adjusted Rand Index

$$rand = \frac{w + d}{\binom{n}{2}},$$

where w is the number of pairs that are within the same group in both C and P , and d is the number of pairs that are in different groups in both C and P .

Adjusted Rand Index

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} + \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] / 2 - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}.$$

Adjusted Rand Index

$$\begin{aligned} \text{ARI} &= \frac{\sum_{ij} \binom{n_{ij}}{2} + \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] / 2 - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}} \\ &= \frac{\text{Index} + \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} \end{aligned}$$

Adjusted Mutual Information - Entropy

$$H(U) = - \sum_{i=1}^{|U|} P_U(i) \log(P_U(i)).$$

Adjusted Mutual Information - MI

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P_{U,V}(i,j) \log\left(\frac{P_{U,V}(i,j)}{P_U(i)P_V(j)}\right).$$

Adjusted Mutual Information - NMI

$$NMI(U, V) = \frac{MI(U, V)}{\sqrt{H(U)H(V)}}.$$

Adjusted Mutual Information

$$AMI = \frac{MI - \text{Expected MI}}{\text{Max Entropy of U or V} - \text{Expected MI}}.$$

Homogeneity, Completeness, V-Measure - Entropy

$$H(C) = - \sum_{i=1}^{|C|} \frac{n_i}{n} \log\left(\frac{n_i}{n}\right).$$

Homogeneity, Completeness, V-Measure - Conditional Entropy

$$H(C \mid P) = - \sum_{i=1}^{|C|} \sum_{j=1}^{|P|} \frac{n_{ij}}{n} \log\left(\frac{n_{ij}}{n_j}\right).$$

Homogeneity, Completeness, V-Measure

$$\text{Homogeneity} = 1 - \frac{H(C | P)}{H(C)}$$

$$\text{Completeness} = 1 - \frac{H(P | C)}{H(P)}$$

$$\text{V-Measure} = \frac{2 \times \text{Homogeneity} \times \text{Completeness}}{\text{Homogeneity} + \text{Completeness}}$$

Unknown Ground Truth

$$\hat{w}_{i,k} = ||x_i - \hat{x}_k||_p^r, \hat{d}_{i,k} = ||x_i - \hat{x}_l||_p^r,$$

where \hat{x}_k is the mean of all the points in cluster k

$$\text{Silhouette}_i = \frac{\hat{d}_{i,k} - \hat{w}_{i,k}}{\max(\hat{w}_{i,k}, \hat{d}_{i,k})}$$

Methods

K Means - Objective

$$\min_{\mu} \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2,$$

where μ_i is the centroid of cluster i and C_i is the subset of points that belong to cluster i .

Algorithm 1: K Means Solution Algorithm

Input: Features from data

Output: Centroids of the k clusters

Initialize cluster centroids (randomly or using a particular strategy)

while *cluster centroids update not converge* **do**

 Determine cluster for each point based on distance to centroids

 Update centroids' location as the mean of points in the cluster

Simple Demo

Involved Demo

Questions

These slides are designed for educational purposes, specifically the CSCI-470 Introduction to Machine Learning course at the Colorado School of Mines as part of the Department of Computer Science.

Some content in these slides are obtained from external sources and may be copyright sensitive. Copyright and all rights therein are retained by the respective authors or by other copyright holders. Distributing or reposting the whole or part of these slides not for academic use is HIGHLY prohibited, unless explicit permission from all copyright holders is granted.