Course. Introduction to Machine Learning
Work 1. Clustering Exercise

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

Contents

1. Clustering exercise

1. Preprocess the data
2. Agglomerative Clustering with sklearn
3. K-Means (your own code)
4. K-Modes (your own code)
5. K-Prototype (your own code)
6. Fuzzy clustering (your own code)
7. Validation techniques (using sklearn validation metrics)

Preprocess the data

Data in

- You need to read the .arff file
  - You can implement your own code or use scipy.io.arff.loadarff
- Data needs pre-processing
  - Features may contain **different ranges**
    - Normalize or Standarize the machine learning data
  - Features may have **different types**
    - Categorical, Numerical, and mix-type data
  - Data may contain **missing values**
    - Use the median (for example)

Data pre-processing

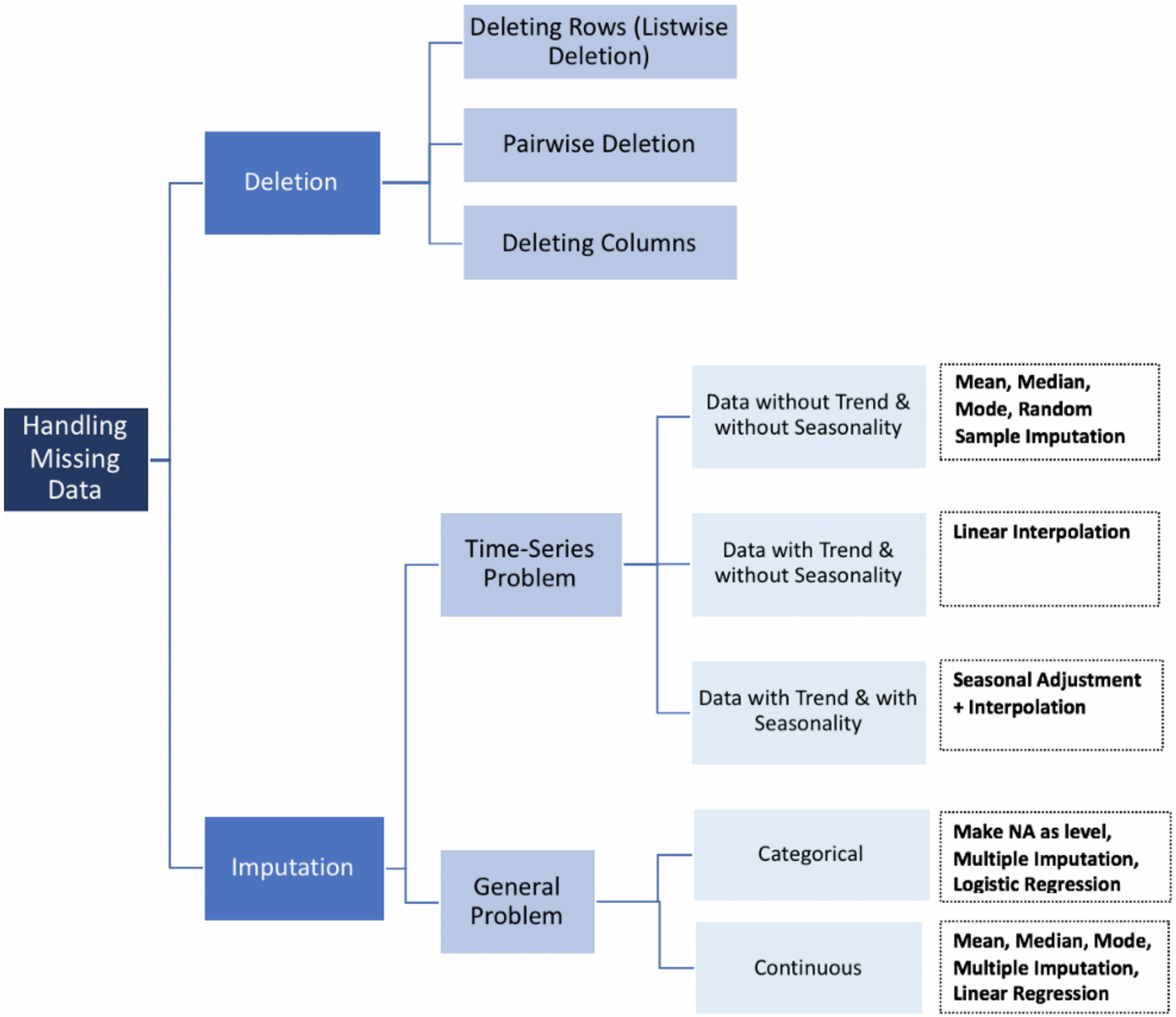
- **To deal with different ranges**
  - Normalize or scale features
- Alternatives
  - **Standardisation**: Standardisation replaces the values by their Z scores. `sklearn.preprocessing.scale`
  - **Mean normalisation**: This distribution will have values between **-1 and 1**with **μ=0**. `sklearn.preprocessing.StandardScaler`
  - **Min-Max scaling**: This scaling brings the value between 0 and 1. `sklearn.preprocessing.MinMaxScaler`
  - **Unit vector**: Scaling is done considering the whole feature vector to be of unit length. `sklearn.preprocessing.Normalizer`

Data pre-processing

- **To deal with different types**
- Alternatives
  - **Label encoding**: convert to a number
    `sklearn.preprocessing.LabelEncoder`
  - **One hot encoding**: where a categorical variable is converted into a binary vector, each possible value of the categorical variable becomes the variable itself with default value of zero and the variable which was the value of the categorical variable will have the value 1.
    `sklearn.preprocessing.OneHotEncoder`

Data pre-processing

- **To deal with missing values**



Agglomerative Clustering

Using sklearn

Some Videos

- https://www.youtube.com/watch?v=VMyXc3SiEqs
- https://www.youtube.com/watch?v=RdT7bhm1M3E
- https://www.youtube.com/watch?v=Cy3ci0Vqs3Y
- http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

K-Means

Implement your own code

K-Means basis

- It is a partitional algorithm that …
  - Assumes instances are **real-valued vectors**
  - Clusters based on ***centroids, center of gravity***, or **mean of points** in a cluster, ***c***:

$$\vec{\mu}(\mathrm{c}) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

  - Reassignment of instances to clusters is **based on distance** to the current cluster centroids
    - Manhattan distance (L_1 norm), Euclidean distance (L_2 norm), Cosine similarity

Discussion on the K-Means method

- K-Means clustering often **terminates at a local optimal**
  - Initialization can be important to find high-quality clusters
- **Need to specify K**, the number of clusters, in advance
  - There are ways to automatically determine the "*best*" K
  - In practice, one often runs a range of values and selected the "*best*" K value
- **Sensitive to noisy data and outliers**
  - Variations: Using K-medians, K-medoids, etc.
- K-Means is applicable only to objects in a **continuous n-dimensional space**
  - Using the K-Modes for **categorical data**
- Non suitable to discover clusters with **non-convex shapes**
  - Using density-based clustering, kernel k-means, etc.

Variations of K-Means

- There are many variants of the K-Means methods, varying different aspects
  - Choosing better initial centroid estimates
    - K-Means++, Intelligent K-Means, Genetic K-Means
  - Choosing different representatives for the clusters
    - K-Medoids, K-Medians, K-Modes
  - Applying feature transformation techniques *(explained at the supervised part of the course)*
    - Weighted K-Means, Kernel K-Means

Initialization of K-Means

- Different initializations may generate rather different clustering results
- Original proposal (MacQueen,1967): selects the k seed randomly
  - Need to run the algorithm multiple times using different seeds
- There are many methods proposed for better initialization of K seeds
  - K-Means++ (Arthur and Vassilvitskii,2007):
    - The first centroid is selected randomly
    - The next centroid selected is the one that is farthest from the currently selected (selection is based on a weighted probability score).
    - The selection continues until K centroids are obtained

Some k-Means references

- MacQueen, J. B. (1967). **Some Methods for classification and Analysis of Multivariate Observations**. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. **(in RACÓ)**

- Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). **A comparative study of efficient initialization methods for the k-means clustering algorithm**. Expert Systems with Applications. 40 (1): 200–210. **(in RACÓ)**

- Arthur, D.; Vassilvitskii, S. (2007). **K-means++: the advantages of careful seeding**. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035. **(in RACÓ)**

K-Modes

K-Modes for categorical data

- K-Means cannot handle non-numerical (categorical) data
  - Mapping categorical value to 1/0 cannot generate quality clusters for high-dimensional data
- K-Modes is a variation of the *K-Means* Method (Huang'98)
  - Replacing means of clusters with <u>modes</u>
  - Using new dissimilarity measures to deal with categorical objects
  - Using a <u>frequency</u>-based method to update modes of clusters

K-Modes basis

- K-Modes: an extension to K-Means by replacing means with **modes**

$$\Phi(x_j, z_j) = 1 - n_j^r/n_l \text{ when } x_j = z_j \text{ ; } 1 \text{ when } x_j \neq z_j$$

where z_j is the categorical value of attribute j in Z_l, n_l is the number of objects in cluster l, and n_j^r is the number of objects whose attribute value is r

- Dissimilarity measure between object X and the center of a cluster Z
- The dissimilarity measure (distance function) is **frequency-based**

$$d(X_i, X_l) \equiv \sum_{j=1}^{m} \delta(x_{i,j}, x_{l,j})$$

where

$$\delta(x_{i,j}, x_{l,j}) = \begin{cases} 0, & x_{i,j} = x_{l,j} \\ 1, & x_{i,j} \neq x_{l,j} \end{cases}$$

K-Modes algorithm

- K-Modes deals with categorical attributes

```
Insert the first K objects into K new clusters.
Calculate the initial K modes for K clusters.
Repeat {
    For (each object O) {
        Calculate the similarity between object O and the
        modes of all clusters.
        Insert object O into the cluster C whose mode is the
        least dissimilar to object O.
    }
        Recalculate the cluster modes so that the cluster
        similarity between mode and objects is maximized.
} until (num_iterations or few objects change clusters).
```

K-Modes

- Algorithm is still based on iterative object cluster assignment and centroid update
- A **fuzzy k-modes** method is proposed to calculate a **fuzzy cluster membership** value for each object to each cluster
- A mixture of categorical and numerical data: Using a **K-prototype** method

UNIVERSITAT DE BARCELONA

References of K-Modes

- Zhexue Huang and Michael K. Ng. 2003. **A Note on K-Modes Clustering**. J. Classif. 20, 2 (September 2003), 257-261. DOI=http://dx.doi.org/10.1007/s00357-003-0014-4 **(in RACÓ)**
- Anil Chaturvedi, Paul E. Green, and J. Douglas Caroll. 2001. **K-Modes Clustering**. J. Classif. 18, 1 (January 2001), 35-55. DOI=http://dx.doi.org/10.1007/s00357-001-0004-3 **(in RACÓ)**
- Zengyou He, **Approximation algorithms for K-Modes clustering**. https://arxiv.org/pdf/cs/0603120.pdf
- Fuyuan Cao, Jive Liang, Deyu Li, Liang Bai, Chuangyin Dang. **A dissimilarity measure for the K-Modes clustering algorithm**. Knowledge-based Systems, Volume 26, 2012, ISSN 0950-7051.DOI= https://doi.org/10.1016/j.knosys.2011.07.011.

  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.652.5571&rep=rep1&type=pdf

K-Prototypes

K-prototypes Algorithm

- To integrate the k-means and k-modes algorithms into the k-prototypes algorithm that is used to cluster the mixed-type objects

- $A_1^r, A_2^r, \ldots, A_p^r, A_{p+1}^c, \ldots, A_m^c$, *m* is the attribute numbers the first *p* means numeric data, the rest means categorical data

K-prototypes Algorithm(cont.)

$$d_2(X,Y)=\sum_{j=1}^{p}(x_j-y_j)^2+\gamma\sum_{j=p+1}^{m}\delta(x_j,y_j)$$

- The first term is the Euclidean distance measure on the numeric attributes and the second term is the simple matching dissimilarity measure on the categorical attributes
- The weight γ is used to avoid favoring either type of attribute

K-prototypes Algorithm(cont.)

Choose clusters

Modify the mode

```
FOR i = 1 TO NumberOfObjects
    Mindistance= Distance(X[i],O_prototypes[1])+ gamma* Sigma(X[i],C_prototypes[1])
    FOR j = 1 TO NumberOfClusters
        distance= Distance(X[i],O_prototypes[j])+ gamma * Sigma(X[i],C_prototypes[j])
        IF (distance < Mindistance)
            Mindistance=distance
            cluster=j
        ENDIF
    ENDFOR
    Clustership[i]=cluster
    ClusterCount[cluster] + 1
    FOR j=1 TO NumberOfNumericAttributes
        SumInCluster[cluster,j] + X[i,j]
        O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]
    ENDFOR
    FOR j=1 TO NumberOfCategoricAttributes
        FrequencyInCluster[cluster,j,X[i,j]] + 1
        C_prototypes[cluster,j]=HighestFreq(FrequencyInCluster,cluster,j)
    ENDFOR
ENDFOR
```

Figure 2. Initial allocation process.

K-prototypes Algorithm(cont.)

UNIVERSITAT DE BARCELONA

```
moves=0
FOR i = 1 TO NumberOfObjects
    …
    (To find the cluster whose prototype is the nearest to object i.  Same as Figure 2)
    …
    IF (Clustership[i]<>cluster)
        moves+1
        oldcluster=Clustership[i]
        ClusterCount[cluster] + 1
        ClusterCount[oldcluster] - 1
        FOR j=1 TO NumberOfNumericAttributes
            SumInCluster[cluster,j] + X[i,j]
            SumInCluster[oldcluster,j] - X[i,j]
            O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]
            O_prototypes[oldcluster,j]= SumInCluster[oldcluster,j]/ClusterCount[oldcluster]
        ENDFOR
        FOR j=1 TO NumberOfCategoricAttributes
            FrequencyInCluster[cluster,j,X[i,j]] + 1
            FrequencyInCluster[oldcluster,j,X[i,j]] - 1
            C_prototypes[cluster,j]=HighestFreq(cluster,j)
            C_prototypes[oldcluster,j]=HighestFreq(oldcluster,j)
        ENDFOR
    ENDIF
ENDFOR
```

Modify the mode

Figure 3. Reallocation process.

References of K-prototypes

- Zhexue Huang, **Clustering large datasets with mixed numerical and categorical values**. https://pdfs.semanticscholar.org/d42b/b5ad2d03be6d8fefa63d25d02c0711d19728.pdf
- Byoungwook Kim. **A Fast K-prototypes Algorithm Using Partial Distance Computation**. https://www.researchgate.net/publication/316348009_A_Fast_K-prototypes_Algorithm_Using_Partial_Distance_Computation

Fuzzy Clustering

Fuzzy Clustering

- Data points are given partial **degree of membership** in multiple nearby clusters
- Central point in the fuzzy clustering is always no unique partitioning of the data in a collection of clusters
- In this membership value is assigned to each cluster. Sometimes this membership has been used to decide whether the data points belong to the cluster or not

Fuzzy C-Means Clustering (FCM)

- Several approximations
  - FCM: Fuzzy C-Means Clustering (Bezdek, 1981)
  - PCM: Possibililistic C-Means Clustering (Krishnapuram - Keller, 1993)
  - FPCM: Fuzzy Possibililistic C-Means (N. Pal - K. Pal - Bezdek, 1997)

References of Fuzzy Clustering

- *C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York*
- **J. C. Bezdek, R. Ehrlich, W. Full (1984). FCM: The fuzzy C-Means Algorithm.**
- James C. Bezdek, James Keller, Raghu Krishnapuram and Nikhil R. Pal (1999), *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, TA 1650.F89.
- R. Krishnapuram and J. M. Keller (1993) A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, Vol. 1, No. 2, pp. 98-110.
- N. R. Pal, K. Pal and J. C. Bezdek (1997), "A mixed c-means clustering model," *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems*, Vol. 1, pp. 11-21.
- Jun Yan, Michael Ryan and James Power, *Using fuzzy logic Towards intelligent systems*, Prentice Hall, 1994.

Validation of clustering

Internal indexes

- Ground truth is rarely available but unsupervised validation must be done.
- Minimizes (or maximizes) internal index:
  - Variances of within cluster and between clusters
  - Rate-distortion method
  - F-ratio
  - Davies-Bouldin index (DBI)
  - Bayesian Information Criterion (BIC)
  - Silhouette Coefficient
  - Minimum description principle (MDL)
  - Stochastic complexity (SC)

Internal indexes

Table B.1: Formulas for internal indexes

Name	Formula
SSW	$SSW = \frac{1}{N}\sum_{i=1}^{N} \left\| x_i - C_{p_i} \right\|^2$
SSB	$SSB = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j \neq i}^{M} \left\| C_i - C_j \right\|^2$
Calinski-Harabasz index	$CH = \frac{SSB/(M-1)}{SSW(N-M)}$
Hartigan	$H_M = (\frac{SSW_M}{SSW_{M+1}} - 1)(N - M - 1)$ $or: H_M = \log(SSB_M / SSW_M)$
Krzanowski-Lai index	$diff_M = (M-1)^{2/D} SSW_{M-1} - M^{2/D} SSW_M$ $KL_M = \lvert diff_M \rvert \,/\, \lvert diff_{M+1} \rvert$
Ball&Hall	$BH_M = SSW_M / M$
Xu-index	$Xu = D \log(\sqrt{SSW_M/(DN^2)}) + \log M$
Dunn's index	$Dunn = \sum_{i=1}^{M} \frac{\max(\left\| x_j - C_i \right\|^2)_{j \in C_i}}{}$
Davies&Bouldin index	$R_{ij} = \frac{S_i + S_j}{d_{ij}}, i \neq j$ $where: d_{ij} = \left\| C_i - C_j \right\|^2, S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left\| x_j - C_i \right\|^2$ $and, R_i = \max_{j=1,\ldots,M} R_{ij}, i = 1, \ldots, M$ $DBI = \frac{1}{M} \sum_{i=1}^{M} R_i$

UNIVER

Internal indexes

Silhouette Coefficients	$a(x_i) = \frac{1}{n_m - 1} \sum_{j=1, j\neq i}^{n_m} \|x_i - x_j\|^2_{x_i, x_j \in C_m}$ $b(x_i) = \min_t \{\frac{1}{n_t} \sum_{j \in C_t} \|x_i - x_j\|^2\}_{x_i \notin C_t}$ $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$ $SC = \frac{1}{N} \sum_{i=1}^{N} s(x_i)$ $b(x_i) = \min\{\sum_{t\neq m} \|C_t - C_m\|^2\}_{x_i \notin C_t} (SC'2008)$
RMSSTD	$RMSSTD = \frac{\sum_{k=1,\ldots,M;\, d=1,\ldots,D} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{k=1,\ldots,M;\, d=1,\ldots,D} (n_{kd} - 1)}$
R-square	$RS = \frac{SST - SSW}{SST} = \frac{\sum_{d\, 1,\ldots,D} \sum_{i=1}^{n_d} (x_i - \overline{x^d})^2 - \sum_{k=1,\ldots,M;\, d=1,\ldots,D} \sum_{i=1}^{n_{kd}} (x_i - \overline{x^d})^2}{\sum_{d=1,\ldots,D} \sum_{i-1}^{n_d} (x_i - \overline{x^d})^2}$
Bayesian Information Criterion	$BIC = L * N - \frac{1}{2} M(D+1) \sum_{i=1}^{M} \log(n_i)$
Xie-Beni	$XB = \frac{\sum_{i-1}^{N} \sum_{k=1}^{M} u_{ik}^2 \|x_i - C_k\|^2}{N \min_{t\neq s}\{\|C_t - C_s\|^2\}}$
Partition Coefficient	$PC = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik}^2 / N$
Partition Entropy	$PE = -(\sum_{i=1}^{N} \sum_{k=1}^{M} u_{ik} \log(u_{ik}))/N$

Soft partitions

External indexes

- Pair counting
  - Chi-Squared Coefficient
  - Rand Index
  - Adjusted Rand Index
  - Fowlkes-Mallows Index
  - Mirkin Metric
- Other measures
  - Information theoretic
    - Mutual Information Metric (MI), Normalized Mutual Information, Variation of Information
  - Set matching
    - Jaccard Index, Normalized Van Dongen, Pair Set Index

Summary of external indexes

Table 1: External Cluster Validation Measures.

	Measure	Notation	Definition	Range
1	Entropy	E	$-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$	$[0, \log K']$
2	Purity	P	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
3	F-measure	F	$\sum_j p_j \max_i [2\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}/(\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j})]$	(0,1]
4	Variation of Information	VI	$-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_i p_j}$	$[0, 2\log\max(K, K')]$
5	Mutual Information	MI	$\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_i p_j}$	$(0, \log K']$
6	Rand statistic	R	$[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2} - \sum_j \binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij}\binom{n_{ij}}{2}/[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - \sum_{ij}\binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}{\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}[\binom{n}{2}-\sum_i\binom{n_{i\cdot}}{2}][\binom{n}{2}-\sum_j\binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2\sum_i\binom{n_{i\cdot}}{2} - 2\sum_j\binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2} - 2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j\binom{n_{\cdot j}}{2}}$	$[0, +\infty)$
12	classification error	ε	$1 - \frac{1}{n}\max_\sigma \sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})/2n$	[0, 1)
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
15	Goodman-Kruskal coefficient	GK	$\sum_i p_i(1 - \max_j \frac{p_{ij}}{p_i})$	[0,1)
16	Mirkin metric	M	$\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2\sum_i\sum_j n_{ij}^2$	$[0, 2\binom{n}{2})$

Note: $p_{ij} = n_{ij}/n$, $p_i = n_{i\cdot}/n$, $p_j = n_{\cdot j}/n$.

Clustering evaluation in Python

- Clustering performance evaluation

```
from sklearn import metrics
```

  - Adjusted Rand index
  - Mutual information based scores
  - Homogeneity, completeness and V-measure
  - Fowlkes-Mallows scores
  - Silhouette Coefficient
  - Calinski-Harabaz Index
  - Davies-Bouldin Index
  - Contingency Matrix

References on cluster validation

1. G.W. Milligan, and M.C. Cooper, “An examination of procedures for determining the number of clusters in a data set”, *Psychometrika*, Vol.50, 1985, pp. 159-179.
2. E. Dimitriadou, S. Dolnicar, and A. Weingassel, “An examination of indexes for determining the number of clusters in binary data sets”, *Psychometrika*, Vol.67, No.1, 2002, pp. 137-160.
3. D.L. Davies and D.W. Bouldin, "A cluster separation measure “, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227, 1979.
4. J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity “, *IEEE Transactions on Systems, Man and Cybernetics*, 28(3), 302-315, 1998.
5. H. Bischof, A. Leonardis, and A. Selb, "MDL Principle for robust vector quantization“, *Pattern Analysis and Applications*, 2(1), 59-72, 1999.
6. P. Fränti, M. Xu and I. Kärkkäinen, "Classification of binary vectors by using DeltaSC-distance to minimize stochastic complexity", *Pattern Recognition Letters*, 24 (1-3), 65-73, January 2003.

References on cluster validation

7. G.M. James, C.A. Sugar, "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach". *Journal of the American Statistical Association*, vol. 98, 397-408, 2003.
8. P.K. Ito, Robustness of ANOVA and MANOVA Test Procedures. In: Krishnaiah P. R. (ed), *Handbook of Statistics 1: Analysis of Variance*. North-Holland Publishing Company, 1980.
9. I. Kärkkäinen and P. Fränti, "Dynamic local search for clustering with unknown number of clusters", *Int. Conf. on Pattern Recognition (ICPR'02)*, Québec, Canada, vol. 2, 240-243, August 2002.
10. D. Pellag and A. Moore, "X-means: Extending K-Means with Efficient Estimation of the Number of Clusters", *Int. Conf. on Machine Learning* (ICML), 727-734, San Francisco, 2000.
11. S. Salvador and P. Chan, "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms", *IEEE Int. Con. Tools with Artificial Intelligence* (ICTAI), 576-584, Boca Raton, Florida, November, 2004.
12. M. Gyllenberg, T. Koski and M. Verlaan, "Classification of binary vectors by stochastic complexity ". *Journal of Multivariate Analysis*, 63(1), 47-72, 1997.

References on cluster validation

13. M. Gyllenberg, T. Koski and M. Verlaan, "Classification of binary vectors by stochastic complexity ". *Journal of Multivariate Analysis*, 63(1), 47-72, 1997.
14. X. Hu and L. Xu, "A Comparative Study of Several Cluster Number Selection Criteria", *Int. Conf. Intelligent Data Engineering and Automated Learning* (IDEAL), 195-202, Hong Kong, 2003.
15. Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. *John Wiley and Sons, London*. ISBN: 10:0471878766.
16. [1.3] M.Halkidi, Y.Batistakis and M.Vazirgiannis: Cluster validity methods: part 1, SIGMOD Rec., Vol.31, No.2, pp.40-45, 2002
17. R. Tibshirani, G. Walther, T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J.R.Statist. Soc. B*(2001) 63, Part 2, pp.411-423.
18. T. Lange, V. Roth, M, Braun and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*. Vol. 16, pp. 1299-1323. 2004.

References on cluster validation

19. Q. Zhao, M. Xu and P. Fränti, "Sum-of-squares based clustering validity index and significance analysis", *Int. Conf. on Adaptive and Natural Computing Algorithms (ICANNGA'09*), Kuopio, Finland, LNCS 5495, 313-322, April 2009.
20. Q. Zhao, M. Xu and P. Fränti, "Knee point detection on bayesian information criterion", *IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI)*, Dayton, Ohio, USA, 431-438, November 2008.
21. W.M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, 66, 846–850, 1971
22. L. Hubert and P. Arabie, "Comparing partitions", *Journal of Classification*, 2(1), 193-218, 1985.
23. P. Fränti, M. Rezaei and Q. Zhao, "Centroid index: Cluster level similarity measure", Pattern Recognition, 2014.