University of Warsaw
Faculty of Economic Sciences

Unsupervised Learning

Winter Semester, 2025/2026

Unsupervised learning: clustering quality

Silhouette index

SEE THE PREVIOUS LECTURE

GAP statistic

GAP statistic

For each potential number of clusters *k*, perform clustering (e.g., k-means) on the original dataset.

Calculate the total within-cluster variation W_k, typically the sum of squared distances between data points and their cluster centroids.

Create *B* reference datasets by sampling uniformly within the range of the observed data. These datasets represent data without any inherent cluster structure.

For each reference dataset and each *k*, perform clustering and compute the within-cluster dispersion W_{kb}. For each *k*, compute the difference between the log within-cluster dispersion of the reference data and the observed data (select *k* where the gap statistic is maximized).

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} logW_{kb} - logW_k$$

Shadow statistics

Shadow statistics

Shadow statistics very close to silhouette. More in Leisch F (2009) Neighborhood Graphs, Stripes and Shadow Plots for Cluster Visualization

https://pdfs.semanticscholar.org/4d41/34253b34c3425b6cf29595d3f84c96a76892.pdf

"The main difference between silhouette values and shadow values is that we replace average dissimilarities to points in a cluster by dissimilarities to point averages (=centroids)." https://rdrr.io/cran/flexclust/man/shadow.html

definition of shadow: twice the distance to the closest centroid divided by the sum of distances to closest and second-closest centroid.

interpretation: shadow~0 → points are close to their centroids

shadow~1 → points are equidistant to the two centroids

Good cluster: many points with small shadow values

Hopkins statistic

Hopkins statistic

In simple words: “how well the data can be clustered”

Hopkins statistics: total y / (total x + total y)

total x → average distance to nearest neighbor between real data

total y → average distance to nearest neighbor between real point

and uniformly generated random point (with the same variance as real data)

Because of randomly generated data statistics may differ.

Null hypothesis: the dataset is uniformly distributed (i.e., no meaningful clusters)

Alternative hypothesis: the dataset is not uniformly distributed (i.e., contains meaningful clusters)

Hopkins statistic

- The ususal interpretation:

h~0	h~0.5	h~1
- accept the null hypothesis **- unlikely that there are statistically significant clusters** **- no clusters are visible, uniformly distributed data**	Random data	- reject the null hypothesis - dataset is significantly a clusterable data - some clusters are visible

Rand index

Rand index

Rand index → it is expressed as **R=(a+b)/(a+b+c+d)**, where all pairs of observations are compared in two periods (t_0 and t_1), and checked if they are in the same or different clusters:

a - in t_0 the same, in t_1 the same,

b - in t_0 different, in t_1 different,

c - in t_0 the same, in t_1 different,

d - in t_0 different, in t_1 the same;

thus, the counter is always the same (a) and always different (b) clusters, and denominator are all possible outcomes (a,b,c,d). Rand Index=1 means that partitions always agree (c and d are NULL) and clusterings are the same, while Rand Index=0 means that partitions migrate and do not agree for even a single pair. Worth to remember, it checks paris of pairs of points. It is insensitive to relabelling. As the random observations may be clustered in the same partitions, the pure Rand Index may not reach zero – for this reason, one uses the **Adjusted Rand Index (ARI)**, rescaled in a way that eliminates random assignments.

Fowlkes-Mallows is almost the same as Rand Index, but A is number of pairs in the same cluster divided by the geometric mean of the sums of the number of pairs in each cluster of the two partitions.

Rand index

period 1

cell A cluster 1	cell B cluster 1
cell C cluster 2	cell D cluster 2

initial pattern

period 2

cell A cluster 2	cell B cluster 2
cell C cluster 1	cell D cluster 1

relabeling of clusters
no change in pattern

period 3

cell A cluster 1	cell B cluster 2
cell C cluster 1	cell D cluster 2

shift of clusters
change in pattern

comparison of pairs of cells in two periods	period 1 (t0) & period 2 (t1)		period 1 (t0) & period 3 (t1)	
a (in t0 the same, in t1 the same)	A:B, C:D	2	---	0
b (in t0 different, in t1 different)	A:C, B:D, A:D, B:C	4	A:D, B:C	2
c (in t0 the same, in t1 different)	---	0	A:B, C:D	2
d (in t0 different, in t1 the same)	---	0	A:C, B:D	2

Rand Index	(a+b)/(a+b+c+d)	=6/6=100%	(a+b)/(a+b+c+d)	=2/6=33%
Jaccard similarity	a/(a+c+d)	=2/2=100%	a/(a+c+d)	=0/4=0%

Jaccard similarity

Jaccard similarity

Jaccard similarity → it is expressed as **J=a/(a+c+d)**, thus it omits a number of events which are always in different clusters (b), both in the counter and denominator. It measures the ratio of overlap and union of two vectors, shapes or any other datasets. Using notation presented for Rand Index, Jaccard similarity is interpreted in a similar way to the Rand Index, but it is concentrated only on pairs that are connected, being a zoom compared to the Rand Index.

Calinski-Harabasz & Duda-Hart

Calinski-Harabasz index

Its construction is as follows:

counter: BGSS/(K-1) → between-group sum of squares (for K clusters)

nominator: WGSS/(N-K) → within-cluster sum of squares (sum of the within-cluster dispersions for all clusters) (for N observations)

The higher statistics the better. The statistic is usually used for comparing solutions for alternative number of clusters.

Duda-Hart index

It is well defined for *kmeans* class. Its hypotheses are as follows:

H0: homogeneity of cluster (data within cluster as similar)

H1: heterogeneity of cluster (one can easily split the cluster)

Statistics dh: ratio of within-cluster sum of squares for two clusters and overall sum of squares.

verification: cluster1=FALSE (H0 of homogeneity rejected, accept H1)

verification: when dh statistics is lower than “compare” (critical value), accept H1

Inertia

Intertia

- **Intra-cluster inertia W**, assuming the existence of a P_K partition, is the sum of I (C_K) inertia in all available K (*k = 1*, .., K) clusters and is expressed by

$$W = \sum_{k=1}^{K} I(C_K)$$

- where the individual intra-cluster inertia are determined as:

$$I(C_K) = \sum_{i \epsilon C_K} w_i d_i^2 \, (x_i, g_k)$$

- where d_i is the distance between the observation x_i and the center of the cluster g_k, while w_i is the weight assigned to the observation - which in particular may be 1/*n* for *n* observations. Intra-cluster inertia for a single cluster is the sum of the weighted square distances between the observations and the center of the cluster; measures heterogeneity within clusters - the lower the inertia and thus the heterogeneity, the more coherent clusters

Intertia

- **Between-clusters inertia B**, measures the separation between clusters and is expressed as the sum of the weighted squared distances d_k between the centers of g_k clusters and the center g of all observations considered together. Hence the inter-cluster inertia is given as:

$$B = \sum_{k=1}^{K} \mu_k d_k^2 (g_k, g)$$

- where μ_k is the sum of the weights assigned to the observations inside the given cluster k:

$$\mu_k = \sum_{i \epsilon C_k} w_i$$

Intertia

- **Total inertia T** is the sum of the weighted squared distances d_g between individual observations x_i and the center *g* of all observations taken together:

$$T = \sum_{i=1}^{n} w_i d_g^2 (x_i, g)$$

- The total inertia does not depend on the division into clusters. The total *T* inertia can also be expressed as the sum of intra-clustering *W* and inter-clustered *B*:

$$T = W + B$$

- This implies that for a given total inertia, independent of the division into clusters, the reduction of inertia (diversity, heterogeneity) inside the cluster translates into an increase in inter-cluster inertia (moves the clusters away from each other)

Intertia

- Good division into clusters is characterized by high inter-cluster inertia (diversity) and low intra-cluster inertia (homogeneity). The measure of the division quality is the percentage of total inertia explained by the division of P_k, which is equal to Q = (1-W/T). This percentage is 100% when each observation is its own cluster (so-called *singletons*), and is 0% when all observations are in one cluster
- In order to select the number of clusters in the literature, Q is also used, defined as $Q = (I_{n+1} - I_n) / I_{n+1}$, where I is a measure of intra-cluster inertia

Thank you!