Machine Learning

Clustering

Karol Przystalski

March 27, 2018

Department of Information Technologies, Jagiellonian University

First things first

You can find me in C-2-36.

Reach me via email: kprzystalski@gmail.com or karol.przystalski@uj.edu.pl

Twitter: @kprzystalski

Skype: kprzystalski

Agenda

1. Introduction
2. Distributed clustering
3. Hierachical clustering
4. Density-based clustering
5. Quality metrics

Introduction

What is clustering?

This group of learning methods are also known under different names. It depends on the context where it is used.

Unsupervised learning can be called as learning without a teacher. It is the opposite to learning with a teacher – supervised learning.

Unsupervised learning is also known as partitioning, segmentation, typology, numerical taxonomy or clustering. The last term is one of the most common used aside from unsupervised learning.

Goal

A cluster is a set of elements/objects of the same label. Comparing to supervised methods, the label used here is based on similarities between elements of each cluster. It means that some elements are more similar to some than to other elements.

In other words, the goal of a clustering method is to find groups of objects that are most similar to each other.

Taxonomy

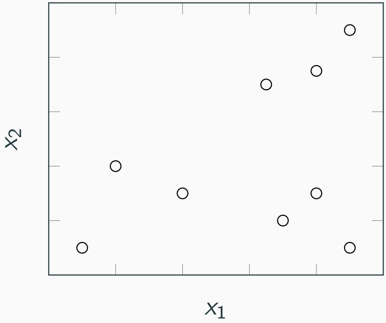
There are three major types of clustering methods:

1. distributed,
2. density-based,
3. hierarchical.

There are more types, but not so popular.

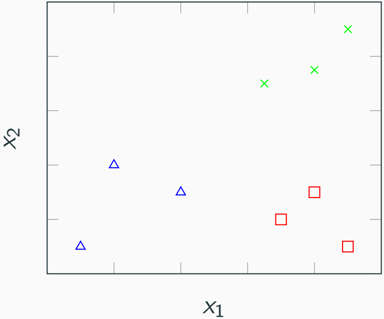
Example

In this figure we have an example data set of elements that we would like to cluster into three groups.



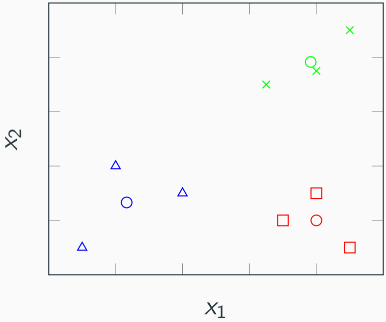
Example

It is obvious that we should have group it as shown in the figure.



Example

The centroids are marked with a circle. Each group has one centroid.



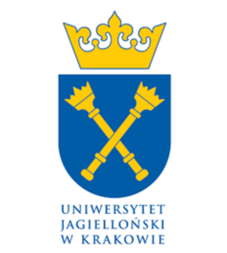
Usage

Clustering methods are widely used. Some application examples are:

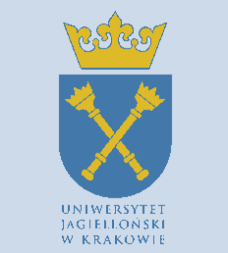
1. customer segmentation,
2. image processing,
3. office localization or city-planning,
4. marketing,
5. and many more.

Image segmentation

Here you can see two examples of how we can use clustering in image segmenatation to distinguish between

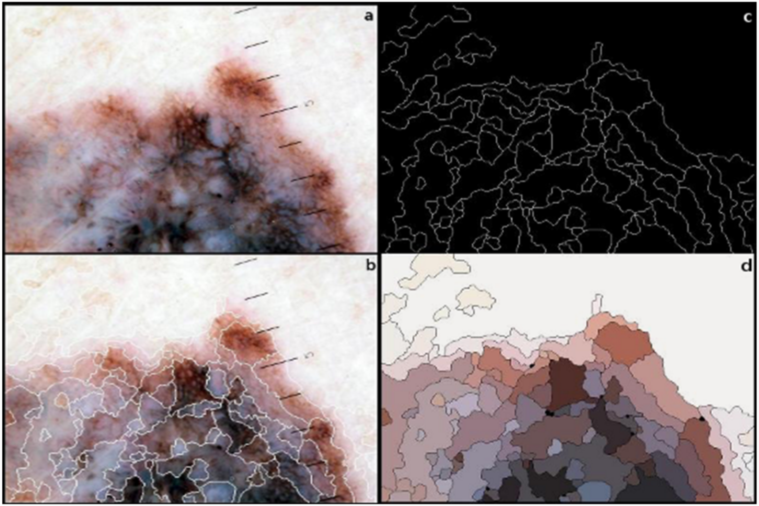


(a)



(b)

More serious about clustering



Distributed clustering

General algorithm

The steps are like following:

1. choose the entrance cluster centroids,
2. calculate the assignation matrix U,
3. calculate new centroids matrix V,
4. calculate the difference between previously assignation matrix U and the new one calculated in current iteration.

General algorithm – details

Choose the entrance cluster centroids

This step is done only once. In most methods the center of each cluster needs to be chosen before the algorithm starts. The two most popular ways to do it, is to set it randomly or set fixed values.

Calculate the assignation matrix U

Assignation matrix calculation step is slightly different in each clustering method. Matrix U consist of c rows and k columns, where c is the number of groups/clusters that we want to have and k is the number of elements in training data set.

General algorithm – details

Calculate new centroids matrix V

Centroids are calculated in most methods in a similar way. The number of groups c_i is the same as the number of centers v_i, where $i = 1, \dots, c$:

$$V = [v_1, v_2, \dots, v_c]. \tag{1}$$

Calculate the change rate

We calculate the assignation matrix from the previous step as well as new centroids in each iteration until the differences between the changes in both are small enough.

k-means

K-means is also known as hard c-means algorithm (hcm). The It is one of the simplest clustering method.

The goal of this algorithm is to assign each element in the training data set into a cluster in a binary way. It means that an element can be assigned to only one cluster fully. It is a strict (hard) way of assignation.

k-means – assignation matrix

In k-means the assignation matrix can look like following:

$$U = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

An element can be assigned to one of two classes in this case.

The assignation is done in a simple way. For each object x_k we measure the distance from it to each center. The closest distance wins:

$$\mu_{ik}^{(t)} = \begin{cases} 1 & \text{if } d(x_k, v_i) < d(x_k, v_j), \text{for each } j \neq i \\ 0 & \text{in other case} \end{cases}. \tag{2}$$

Centroids

Each group center is calculated separately as following:

$$v_i = \frac{\sum_{k=1}^{M} \mu_{ik}^{(t)} x_k}{\sum_{k=1}^{M} \mu_{ik}^{(t)}}. \tag{3}$$

At the end we should get an array of group centroids:

$$V = [v_1, v_2, \ldots, v_c]. \tag{4}$$

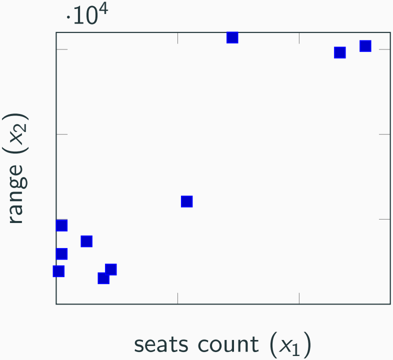
Example – Aircrafts

Let’s take an example of aircrafts

Aircraft name	Distance range (km)	Seats count	Aircraft type
Cesna 510 Mustang	1940	4	private jet
Falcon 10/100	2960	9	private jet
Hawker 900/900XP	4630	9	private jet
ATR 72-600	1528	78	medium size aircraft
Bombardier Dash 8 Q400	2040	90	medium size aircraft
Embraer ERJ145 XR	3700	50	medium size aircraft
Boeing 747-8	14815	467	jet airliner
A380-800	15200	509	jet airliner
Boeing 787-8	15700	290	jet airliner
Boeing 737-900ER	6045	215	jet airliner

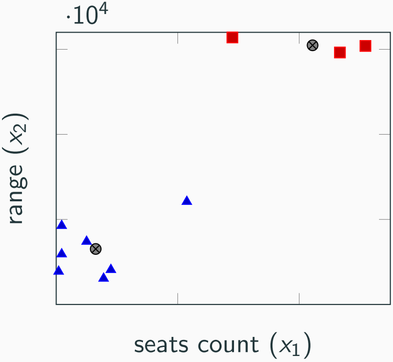
Example – Aircrafts

We can draw the example as following:



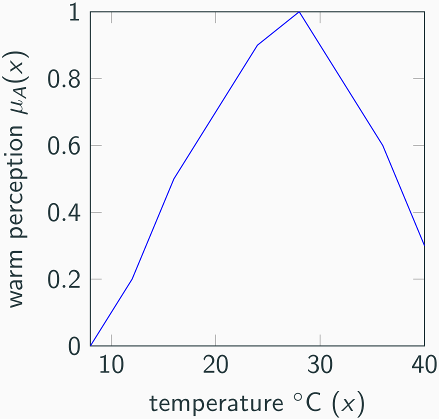
Example

The obvious clustering into two groups would look like in the figure below. The gray circles are the centroids of each group.



Assignation function – fuzzy sets

In fuzzy sets we have an assignation function $\mu_A(x)$ that assigns a value based on the current fuzzy set. If we take the temperature of the weather it could look like following:



Fuzzy

In fuzzy k-means (fcm) the assignation function is different:

$$\mu_{ik} = (\sum_{j=1}^{c} (\frac{d(x_k, v_i)}{d(_k, v_j)})^{\frac{2}{m-1}})^{-1} \tag{5}$$

It assigns a value from 0 to 1 to a group, but the sum of the values cannot be greater than 1.

The centers are also calculated a bit differently:

$$v_i = \frac{\sum_{k=1}^{M} (\mu_{ik}^{(t)})^m x_k}{\sum_{k=1}^{M} (\mu_{ik}^{(t)})^m} \tag{6}$$

. An example assignation matrix can look in case of FCM like following:

$$U = \begin{bmatrix} 0.45 & 0.65 & 0.3 & 0.9 & 0.25 \\ 0.55 & 00.350.7 & 0.1 & 0.75 & \end{bmatrix}.$$

Possibilistic

The possibilistic k-means (PCM) is a bit more complex and the assignation function takes

$$\mu_{ik} = (1 + (\frac{D_{ikA}}{\eta_i})^{\frac{2}{m-1}})^{-1}, \tag{7}$$

where possibilistic distribution measure:

$$\eta_i = \frac{\sum_{k=1}^{M} (\mu_{ik})^m D_{ikA}^2}{\sum_{k=1}^{M} (\mu_{ik})^m}, \tag{8}$$

and

$$D_{ikA}^2 = ||x_k - v_i||_A^2 = (x_k - v_i)^T A(x_k - v_i), \tag{9}$$

where is a diagonal matrix.

An example assignation matrix for PCM can look like following:

$$U = \begin{bmatrix} 0.45 & 0.85 & 0.4 & 0.9 & 0.35 \\ 0.75 & 00.450.8 & 0.15 & 0.75 & \end{bmatrix}.$$

Hierachical clustering

Hierchical clustering types

We have two types of hierchical clustering methods:

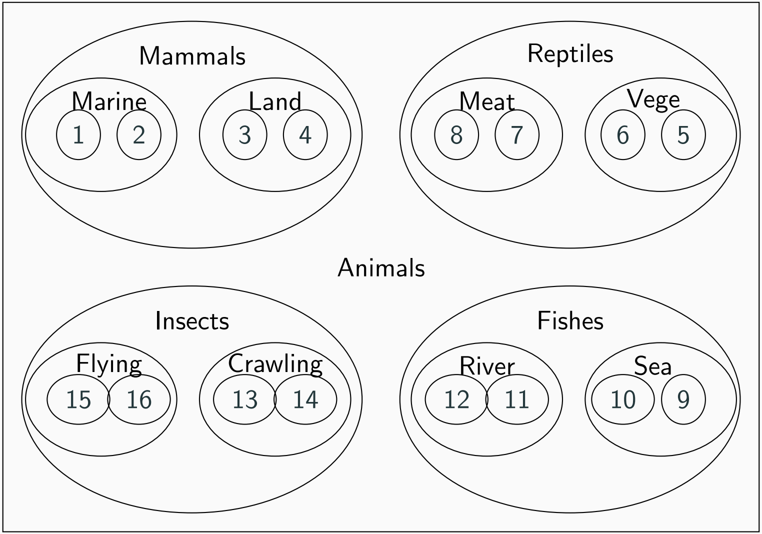
1. agglomerative,
2. divisive.

Both a doing almost the same, but in the opposite direction. Let's take an example to show the differences:

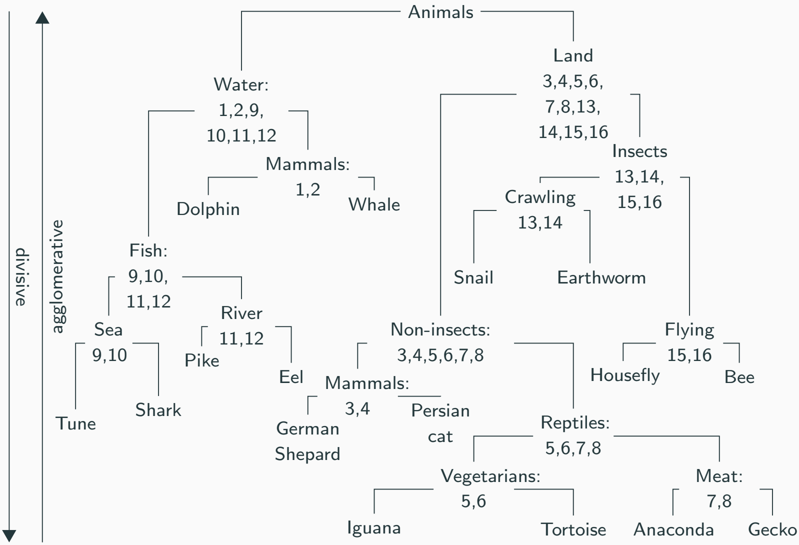
Id	Group	Subgroup	Animal name	Id	Group	Subgroup	Animal name
1	Mammals	Marine	Dolphin	9	Fishes	Sea	Shark
2	Mammals	Marine	Whale	10	Fishes	Sea	Tune
3	Mammals	Land	German Sheppard	11	Fishes	River	Pike
4	Mammals	Land	Persian cat	12	Fishes	River	Eel
5	Reptiles	Vegetarian	Iguana	13	Insects	Crawling	Ladybug
6	Reptiles	Vegetarian	Tortoise	14	Insects	Crawling	Earthworm
7	Reptiles	Meat eaters	Anaconda	15	Insects	Flying	Bee
8	Reptiles	Meat eaters	Gecko	16	Insects	Flying	Housefly

Nested sets

As nested sets the data set would look like following:



Dendrogram



divisive
agglomerative
Animals
Water:
1,2,9,
10,11,12
Land
3,4,5,6,
7,8,13,
14,15,16
Mammals:
1,2
Dolphin
Whale
Insects
13,14,
15,16
Crawling
13,14
Snail
Earthworm
Fish:
9,10,
11,12
Sea
9,10
Tune
Shark
River
11,12
Pike
Eel
Non-insects:
3,4,5,6,7,8
Mammals:
3,4
German
Shepard
Persian
cat
Flying
15,16
Housefly
Bee
Reptiles:
5,6,7,8
Vegetarians:
5,6
Iguana
Tortoise
Meat:
7,8
Anaconda
Gecko

Agglomerative clustering

The agglomerative clustering method is divided into three steps:

1. calculate current dendrogram distance matrix,
2. get lowest distance from matrix,
3. merge clusters/elements into clusters.

It is repeated until we have one cluster or expected clusters number.

Divisive clustering

The divisive clustering method is divided into three steps:

1. calculate distance matrix in each cluster,
2. get highest distance average,
3. split clusters.

It is repeated until we have no cluster to be divided or expected clusters number is reached.

Disimmilarity measure

There are few popular dissimilarity measures that we can use to generate the distance marix.

Measure name	Equation	
Manhattan distance	$\rho_{Man}(x_r, x_s) = \sum_{i=1}^{n} \lvert x_{ri} - x_{si} \rvert$	(10)
Chebyshew distance	$\rho_{Ch}(x_r, x_s) = max_{1 \leq i \leq n} \lvert x_{ri} - x_{si} \rvert$	(11)
Frecht distance	$\rho(x_r, x_s) = \sum_{i=1}^{d} \frac{\lvert x_{ri} - x_{si} \rvert}{1 + \lvert x_{ri} + x_{si}} \frac{1}{2^i}$	(12)
Canberra distance	$\rho(x_r, x_s) = \sum_{i=1}^{d} \frac{\lvert x_{ri} - x_{si} \rvert}{\lvert x_{ri} + x_{si} \rvert}$	(13)
Post office distance	$\rho_{pos}(x_r, x_s) = \begin{cases} \rho_{Min}(x_r, 0) + \rho_{Min}(0, x_s), \text{ for } x_r \neq x_s, \\ 0, \text{ for } x_r = x_s \end{cases}$	(14)
Bray-Curtis distance	$\rho_{bc}(x_r, x_s) = \frac{\sum_{i=1}^{d} \lvert x_{ri} - x_{si} \rvert}{\sum_{i=1}^{d} (x_{ri} - x_{si})}$	(15)

Agglomerative distance measure types

Based on the approach, we can use few methods to measure the distance between clusters:

Method name	Equation	
Single Linkage	$d_{12} = \min_{i,j} d(X_i, Y_j)$	(16)
Complete Linkage	$d_{12} = \max_{i,j} d(X_i, Y_j)$	(17)
Average Linkage	$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{j} d(X_i, Y_j)$	(18)
Centroid Method	$d_{12} = d(\bar{x}, \bar{y})$	(19)

Density-based clustering

Density clustering method steps

DBScan is one of density-based clustering methods. It consist of the following steps:

1. calculate distance matrix,
2. get closest element,
3. merge into a cluster if the distance is small enough.

It can be also used to find noise in the dataset.

Neighborhood

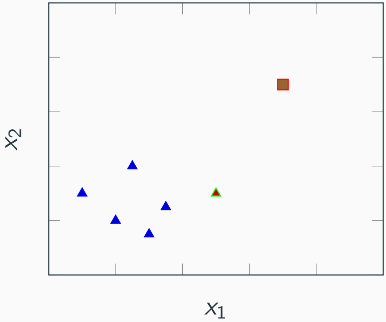
The density is calculated using the neightborhood elements:

$$N_\epsilon : q|d(p, q) \leq \epsilon, \tag{20}$$

where p and q are two elements of the training data set and ϵ is the neighborhood distance.

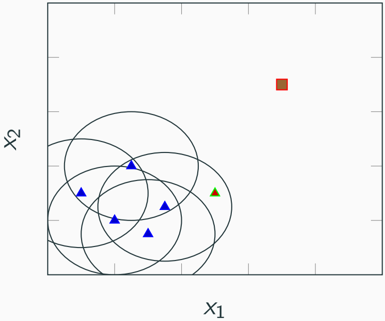
Differences between Density-based and hiearchical

In the figure below we have the core points marked with blue, the border points marked with green and outliers marked with red.



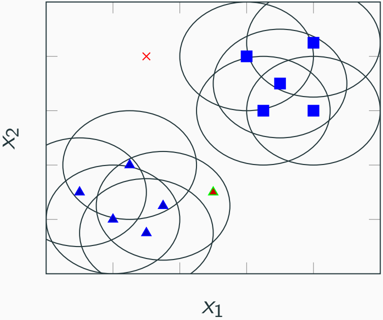
Differences between Density-based and hiearchical

The same figure, but with the neighborhoods:



Differences between Density-based and hiearchical

The same figure, but with the neighborhoods:



Quality metrics

Quality metric types

Finding the best clustering method is not an easy task. To make this task easier we can use multiple validation methods. The most important factors are **homogeneity** and **heterogeneity** of a clustering method.

Other possibility is to use one of a commonly known **validation methods**.

How many clusters should we have? Real world clustering problems can be complex and it can be hard to choose the **best number of clusters**.

Good clustering

We say that the clustering method done well when we have:

- high intra-class similarity,
- low inter-class similarity.

Both values are also known as homogeneity and heterogeneity.

Homogeneity

Two metrics of homogeneity that we explain here are marked as σ_1 and σ_2. Both are related to the differences within each cluster. The differences are known as a dispersion measures within a cluster.

As we do some calculation within a cluster we need to refer to a cluster center. The equation of the average objects dispersion looks like following:

$$\sigma_1(c_i) = \frac{1}{m} \sum_{x_1, x_2 \in c_i} d^2(x_1, x_2), \tag{21}$$

where the m is defined as

$$m = \frac{(n_i - 1)n_i}{2}. \tag{22}$$

The n_i is the count of objects within i cluster. If we have two clusters we calculate two dispersion measures σ_1, one for each cluster.

Homogeneity

A second dispersion measure is marked as σ_2. We calculate here the distance power between each object x within a cluster and the cluster center c_i. We divide the result by the count of objects within the cluster:

$$\sigma_2(c_i) = \frac{1}{n_i} \sum_{x \in c_i} d^2(x, c_i). \tag{23}$$

It looks a bit simpler comparing to σ_1. In both cases the smallest value testifies a better clustering result.

Homogeneity

Similar measures as are the total dispersion measures. Those metrics give a better understanding of recurrence of objects within a cluster and feature space. Both metrics are just sums of dispersion measures σ_1 and σ_2:

$$r(\sigma_1) = \sum_{i=1}^{K} \sigma_1(c_i), \tag{24}$$

$$r(\sigma_2) = \sum_{i=1}^{K} \sigma_2(c_i). \tag{25}$$

Small values of $r(\sigma_1)$ and $r(\sigma_2)$ means high recurrence of objects within the feature space. Higher values mean exactly the opposite.

Heterogeneity

We have four separation measures $s_1(c_i, c_j)$, $s_2(c_i, c_j)$, $s(s_1)$ and $s(s_2)$. The first two separation measures explain how far from each other the clusters are. We measure it for each pair of centroids.

$$s_1(c_i, c_j) = \frac{1}{n_i n_j} \sqrt{\sum_{x_1, \in c_i, x_2 \in c_j} d^2(x_1, x_2)}. \tag{26}$$

We take two objects, each from different cluster and calculate the power distance measure. Next, we sum all the distances from object of two clusters and calculate a square root of it. The value is then divided by the multiplication of the counts of objects in both clusters.

Heterogeneity

The second separation measure is about the distance between two centroids:

$$s_2(c_i, c_j) = d(c_i, c_j). \tag{27}$$

It is the simplest measure so far as it is not even a sum. We need two loops to go through the centroids

Homogeneity and heterogeneity

The third separation measure uses dispersion measure σ_1. It is a simple sum of a division of s_1 for two centroids and σ_1 for centroid c_i:

$$s(s_1) = \sum_{i,j=1;j\neq i}^{K} \frac{s_1(c_i, c_j)}{\sigma_1(c_i)}. \tag{28}$$

This measure is taking the whole feature space into consideration. The sums can be easily calculated if we already have s_1 and σ_1.

The last measure is also simple. It is a sum of measures s_2:

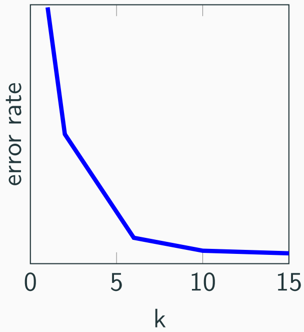
$$s(s_2) = \sum_{i,j=1;j\neq i}^{K} s_2(c_i, c_j) \tag{29}$$

Homogeneity and heterogeneity

We do not need to calculate all measures to get to know if our clustering method is performing well. In most cases we can use just few or one. Especially $r(\sigma_2)$ is used very often.

Clusters number

The number of clusters can be chosen using the elbow method. The goal of this method is to choose many numbers *k* and calculate the error rate. Based on the result of each execution we can plot a chart that can look like



At some point in this chart increasing the number *k* gives a a lower error rate. The error rate is next on a similar level for the next values of *k*.

Internal and external indices

Internal and external indices are other type of measures that shows how good our clustering method is. The difference between internal and external indices depends on the information used to calculate the index.

Internal indices are based only on training data set.

External indices uses the labels and testing data set.

Dunn index

The Dunn index can be easily calculated as a quotient of two distances:

$$C = \frac{d_{min}}{d_{max}}, \tag{30}$$

where the equations of d_{max} and d_{min} are like following:

$$d_{\max} = \max_{1 \leq k \leq K} D_k, \tag{31}$$

$$d_{\min} = \min_{k \neq k'} d_k. \tag{32}$$

Both distances are just the minimum and maximum euclidean distances between objects. The minimum distance is a measure of two object that are in different clusters:

$$d_k = \min_{i,j \in I_k; i \neq j} d(x_i^{(k)} - x_j^{(k')}) \tag{33}$$

.

Dunn index

The clusters are marked with k and k'. The maximum distance takes the distance of two objects within a cluster:

$$D_k = \max_{i,j \in I_k; i \neq j} d(x_i^{(k)} - x_j^{(k)}). \tag{34}$$

D_k and d_k values are calculated for each cluster k, but in Dunn index we take only the highest value of D_k and the lowest value of d_k.

Czekanowski-Dice index

Czekanowski-Dice index is taking labels of the testing data set to measure the quality of a clustering method. It is calculated as following:

$$C = 2\frac{P \times R}{R + R}. \tag{35}$$

Precision R that is also known as Positive Predictive Value (PPV):

$$PPV = \frac{\#TP}{\#TP + \#FP}. \tag{36}$$

TPR is also called sensitivity or recall and is a measure of good predictions within a set of cases:

$$TPR = \frac{\#TP}{\#TP + \#FN}. \tag{37}$$

Questions?