Data Mining
Introduction

Prof. Dr. T. Nouri
Nouri@Nouri.ch

181120

Decision Tree classification , supervised
Entscheidungsbaum

Baum erstellen

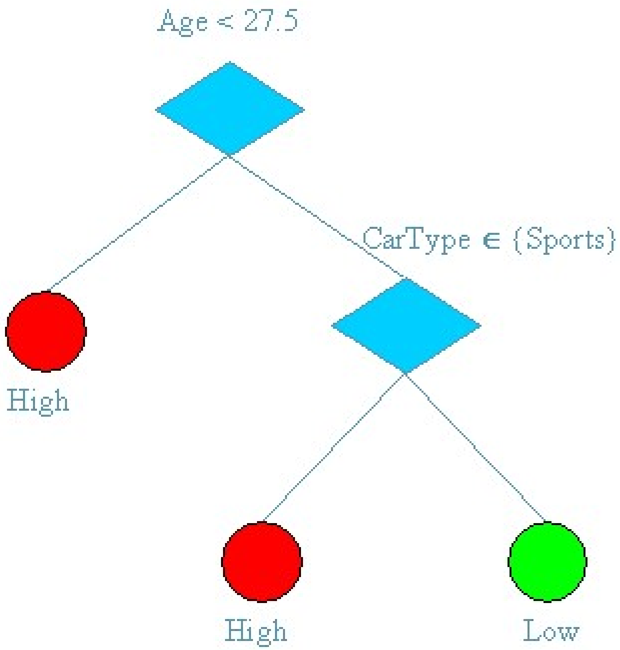
$$\text{"Entropie"} = -p_+ \ln\left({p_+}/{n} \right) - p_- \ln\left({p_-}/{n} \right)$$

p_+ = Anz. richtig klassifizierte
p_- = Anz. falsch klassifizierte
n = Anzahl Objekte
ln=Logarithmus basis 2

Example1: Decision Tree classification

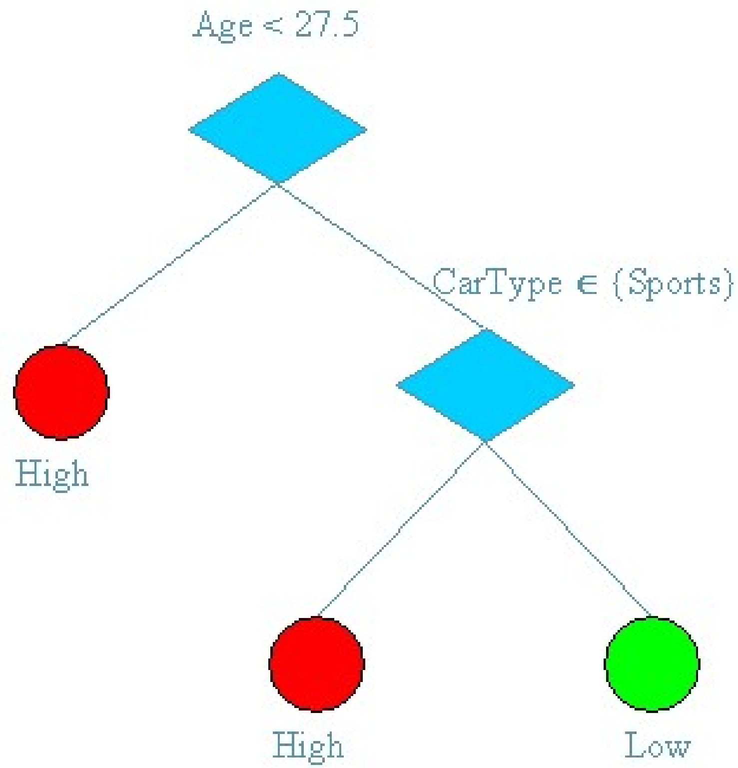
Tid	Age	Car Type	Class
0	23	Family	High
1	17	Sports	High
2	43	Sports	High
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

Numeric Categorical



Age=40, CarType=Family ⇒ Class=Low

From Tree to Rules:



1) $Age < 27.5 \Rightarrow High$

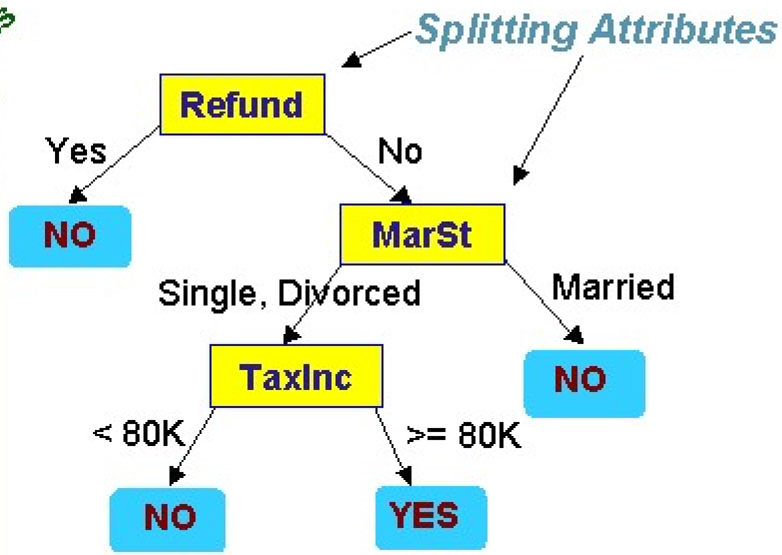
2) $Age >= 27.5$ and
$CarType = Sports \Rightarrow High$

3) $Age >= 27.5$ and
$CarType \neq Sports \Rightarrow High$

Example2: Decision Tree classification

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



The splitting attribute at a node is determined based on the Gini index.

1. Extrahieren Sie eine Rule von diesem Baum!
2. Ryan hat NO refund, Married, Income 120K Cheatet er oder nicht?

Association Rules, unsupervised

Warenkorbanalyse (Beispiel für das Auffinden von Assoziationsregeln): Unter einem Warenkorb versteht man dabei eine Sammlung von Dingen, die ein Kunde etwa in einem Supermarkt in einer Transaktion erworben hat.
Lieferanten oder Ladeninhaber sowie Supermarktbetreiber möchten nun herausfinden, welche Dinge zusammen gekauft werden, etwa um deren Platzierung im Regal oder in der Werbung zu verbessern.

Idee der Assoziationsregel:
<u>informal erkennen des Zusammenhangs zwischen verschiedenen Teilen:</u>
=> z.B. gemeinsam in Kundentransaktionen erscheinende Teil: "Füller => Tinte"
<u>Ziel:</u>
einen gewissen Rahmen zu schaffen, in welchem sich derartige Aussagen (oder Vermutungen) einerseits erhärten und andererseits sogar systematisch ermitteln lassen.

Assoziationsregel: LS => RS

(Warenkorb-) Tabelle

TID	KundenID	Datum	Teil	Preis	Qty
134	201	02.12.97	Füller	35	2
134	201	02.12.97	Tinte	2	1
134	201	02.12.97	Heft	5	3
134	201	02.12.97	Seife	1	6
107	83	13.11.97	Füller	35	1
107	83	13.11.97	Tinte	2	1
107	83	13.11.97	Heft	5	1
110	135	13.11.97	Füller	35	1
110	135	13.11.97	Heft	5	1
103	201	26.08.97	Füller	35	2
103	201	26.08.97	Tinte	2	2
103	201	26.08.97	Seife	1	4

wobei **LS** (linke Seite) und **RS** (rechte Seite) disjunkte Mengen von Dingen („Itemsets11) sind und die Bedeutung analog zum Beispiel lautet: Wird jedes Teil in der linken Seite **LS** gekauft, so wird (wahrscheinlich) auch jedes Teil in der rechten Seite **RS** gekauft.

Wir gehen damit also aus von einer Menge $I = \{i_i \ldots i_m\}$ von Dingen oder Items und bezeichnen Mengen von Dingen, also Teilmengen $T \subseteq I$, als Transaktionen. Der Gegenstand der Analyse ist eine "Datenbank" $D = \{T_i \ldots T_k\}$ von Transaktionen.

formale Beschreibung von Warenkorb:

Die betrachtete Menge D umfasst vier Transaktionen, die jeweils durch einen Identifikator eindeutig gekennzeichnet sind.

Die Transaktionen sind über einer Menge

    I = {Füller, Tinte, Heft, Seife, ...}

von Teilen gebildet.

Assoziationsregeln schreiben wir dann auch in der Form **R**:

    LS => RS

Support einer Menge von Dingen: Die „Wichtigkeit" oder Bedeutung einer Menge von Dingen. Je höher der durch das Mass zugeordnete Wert, desto wichtiger die betreffende Menge.

Confidence: Die „Stärke" einer Regel.
Die Confidence einer Regel **R**: **LS => RS**: bezeichnet man der Prozentsatz der Transaktionen, die RS umfassen, falls sie auch alle Elemente von LS enthalten

Die Confidence einer Regel deutet damit den Grad der Korrelation zwischen Verkäufen von Mengen von Dingen (in der Datenbank) an. Diese Definition von Confidence benutzt den Support , d.h. sie sind aus mathematischer Sicht nicht unabhängig voneinander.
Man kann dies durchaus als eine gewisse Kritik an diesen beiden Massen ansehen

Betrachten wir hierzu einige Beispiele:
Die oben bereits betrachtete Regel R : „Füller => Tinte" lässt sich wie folgt bewerten:
Da die Teile Füller und Tinte in drei der vier in Transaktionen gemeinsam vorkommen, gilt
supp(R) = 3/4 = 0,75
Weiter gilt supp(Füller) = 4/4 = 1, also erhält man
conf(R) = 0,75/1 = 0,75
Es enthalten also 3/4 der Transaktionen Tinte, sofern sie bereits Füller enthalten.
Die Regel laute „Bier =>Chips":
Ein Support von 0,8 bedeutet dann, dass in 80 % der Transaktionen Bier und Chips gemeinsam vorkommen; unabhängig davon bedeutet eine Confidence von 0,5, dass die Hälfte der Leute, die Bier gekauft haben, auch Chips (dazu) gekauft haben.
Ist der Support gering, kann es sich um einen zufälligen Zusammenhang handeln (z.B. "Heft => Seife")
Iist die Confidence einer Regel gering , so ist die linke Seite nicht stark mit der rechten korreliert (z.B. bei "Heft => Seife").
In realen Anwendungen wird man meistens so vorgehen, dass man einen Mindest-Support sowie eine Minimal-Confidence vorgibt und sich dann nur für Regeln interessiert, welche beides enthalten.

Beispiel Warenkorbtabelle mit $\sigma = 0.7$ und $\gamma = 0.8$:

I = {Füller, Tinte, Heft, Seife}

- **häufige Einermenge:**

{{Füller}, {Tinte}, {Heft}}

- **häufige Zweiermenge:**

{{Füller, Tinte}, {{Füller, Heft}}

- **potenziellen Regeln:**

Füller => Tinte

Tinte => Füller

Füller => Heft

Heft => Füller

- **Überprüfen mit Confidence:**

conf(1) = 0.75

conf(2) = 1

conf(1) = 0.75

conf(4) = 1

Lösungsansätze:

LS	RS	supp	conf
Füller	Tinte	0.7	0.75
Tinte	Füller	0.7	1
Füller	Heft	0.7	0.75
Heft	Füller	0.7	1

Beispiel

TID	*Items*
1	**Bread, Milk**
2	**Beer, Diaper, Bread, Eggs**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Bread, Diaper, Milk**

Association Rule: $X \Rightarrow_{s,\alpha} y$

Support: $s = \frac{\sigma(X \cup y)}{|T|} \; (s = P(X, y))$

Confidence: $\alpha = \frac{\sigma(X \cup y)}{\sigma(X)|} \; (\alpha = P(y \mid X))$

Example:

$\{\text{Diaper, Milk}\} \Rightarrow_{s,\alpha} \text{Beer}$

$$s = \frac{\sigma(\text{Diaper, Milk, Beer})}{\text{Total Number of Transactions}} = \frac{2}{5} = 0.4$$

Exercice

TransaktionID	PassagierID	Ziel
431	102	New York
431	102	London
431	102	Cairo
431	102	Paris
701	*38*	*New York*
701	*38*	*London*
701	*38*	*Cairo*
11	531	New York
11	531	Cairo
301	*102*	*New York*
301	*102*	*London*
301	*102*	*Paris*

Having the following Rule: **Rule:** *Who visit New York, visit London too.<==>New York=>London.*

Calculate the support and the Confidence of this Rule?

Clustering: Unsupervised, Descriptif

What is clustering?

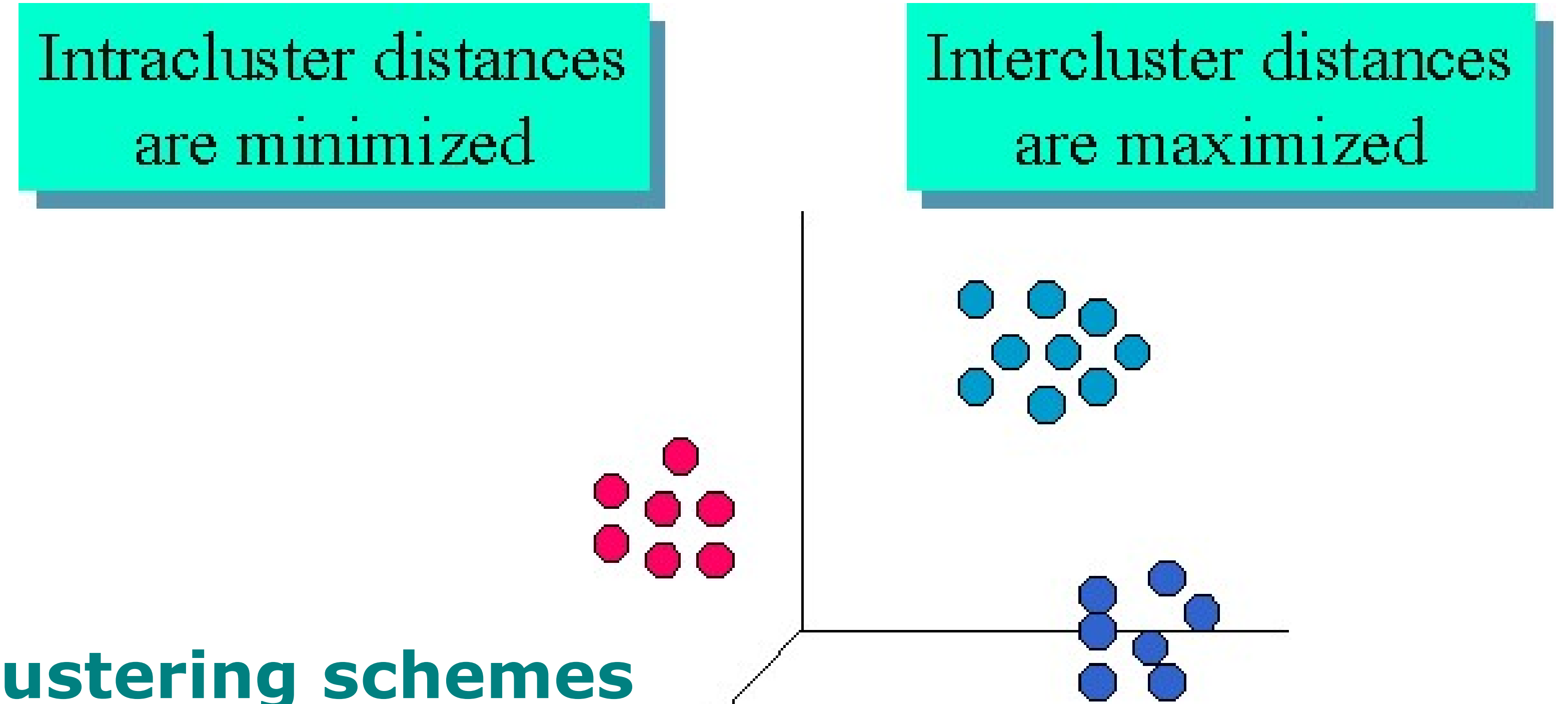
Clustering is a non supervised technique!!!(Decision tree is a supervised algorithm).

Clustering involves grouping data into several new classes. It is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. For example, we may want to cluster houses to find distribution patterns.

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering analysis helps construct meaningful partitioning of a large set of objects.

The task of clustering is to maximize the intra-class similarity and minimize the interclass similarity.

Euclidean Distance Based Clustering in 3-D space.

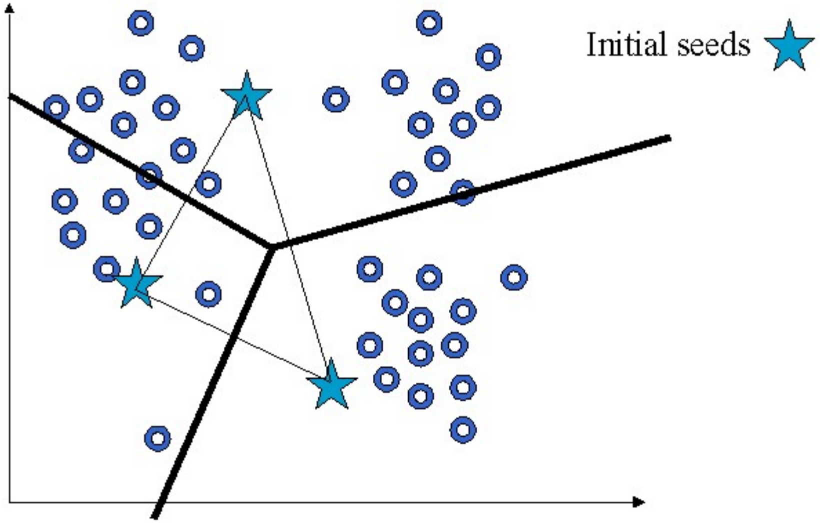


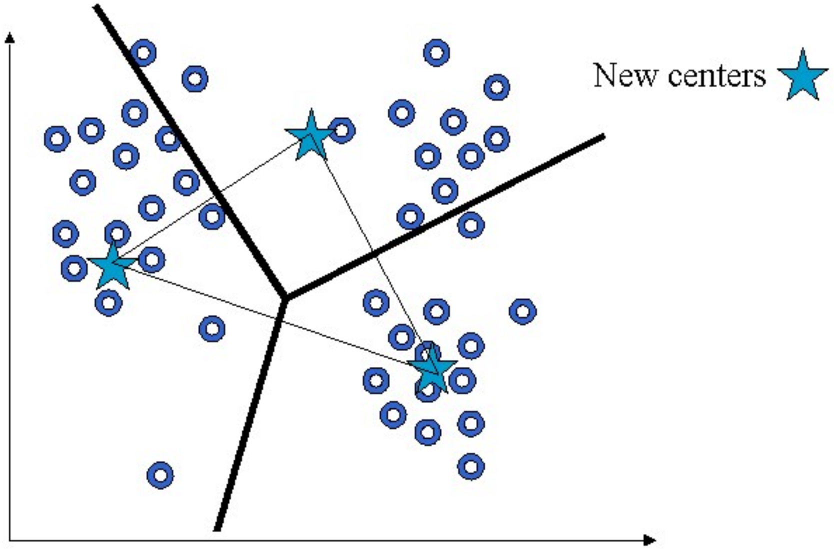
Clustering schemes

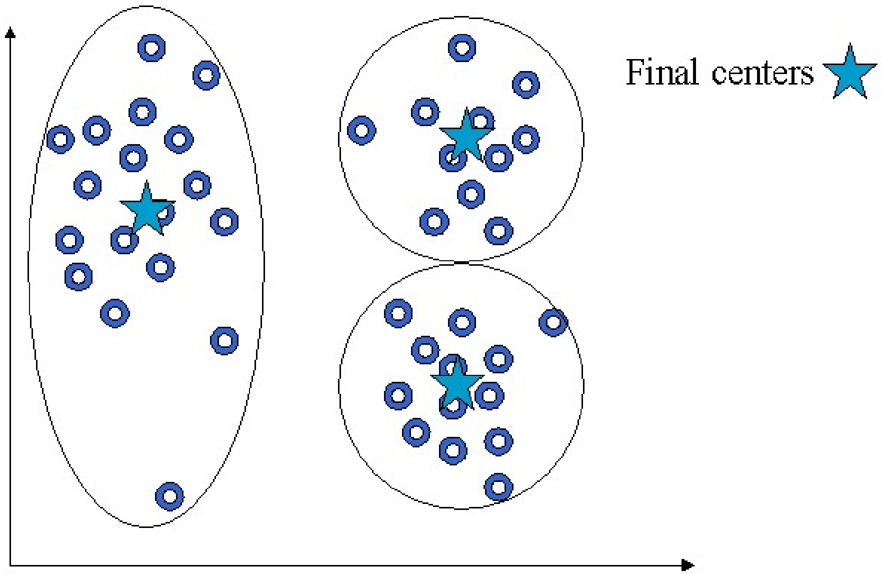
•Distance-based (Numeric: Euclidean distance (root of sum of squared differences along each dimension or Angle between two vectors).
•Categorical (Number of common features (categorical))
•Partition-based (Enumerate partitions and score each)
•Model-based
•Estimate a density (e.g., a mixture of gaussians)
•Compute P(Feature Vector i | Cluster j)
•Finds overlapping clusters too

The k-means algorithm

1.Specify ‘k’, the number of clusters
2.Guess ‘k’ seed cluster centers
3.Look at each example and assign it to the center that is closest
4.Recalculate the center
Iterate on steps 3 and 4 till centers converge or for a fixed number of times

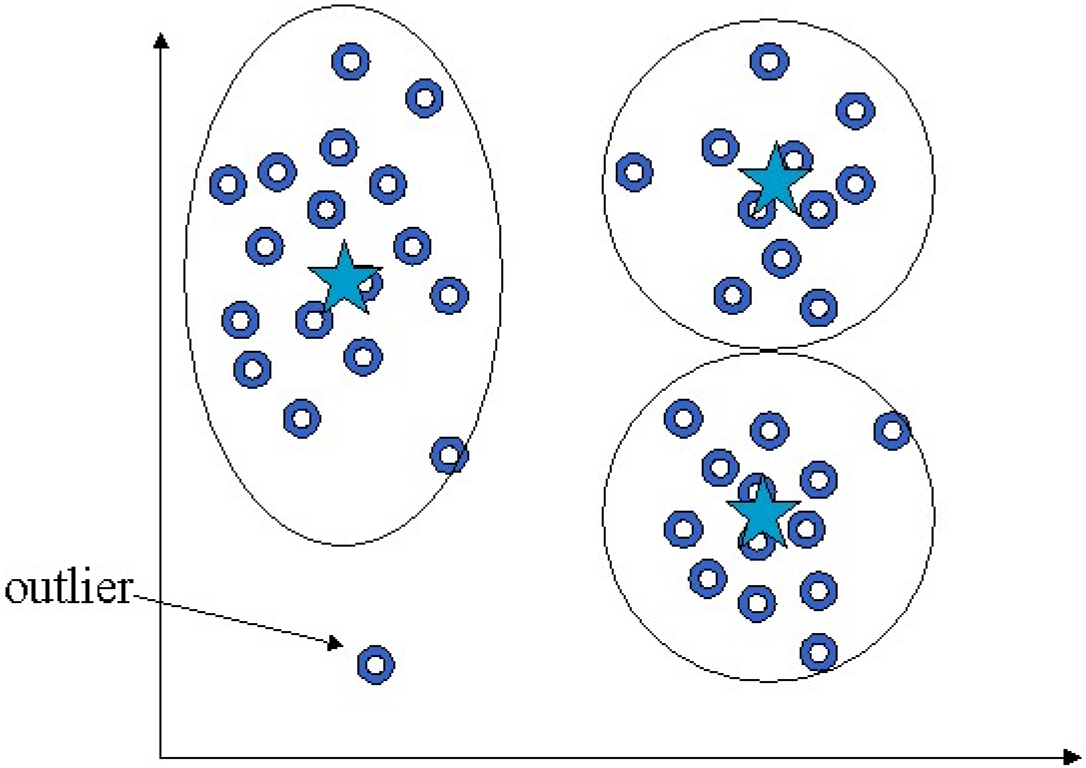






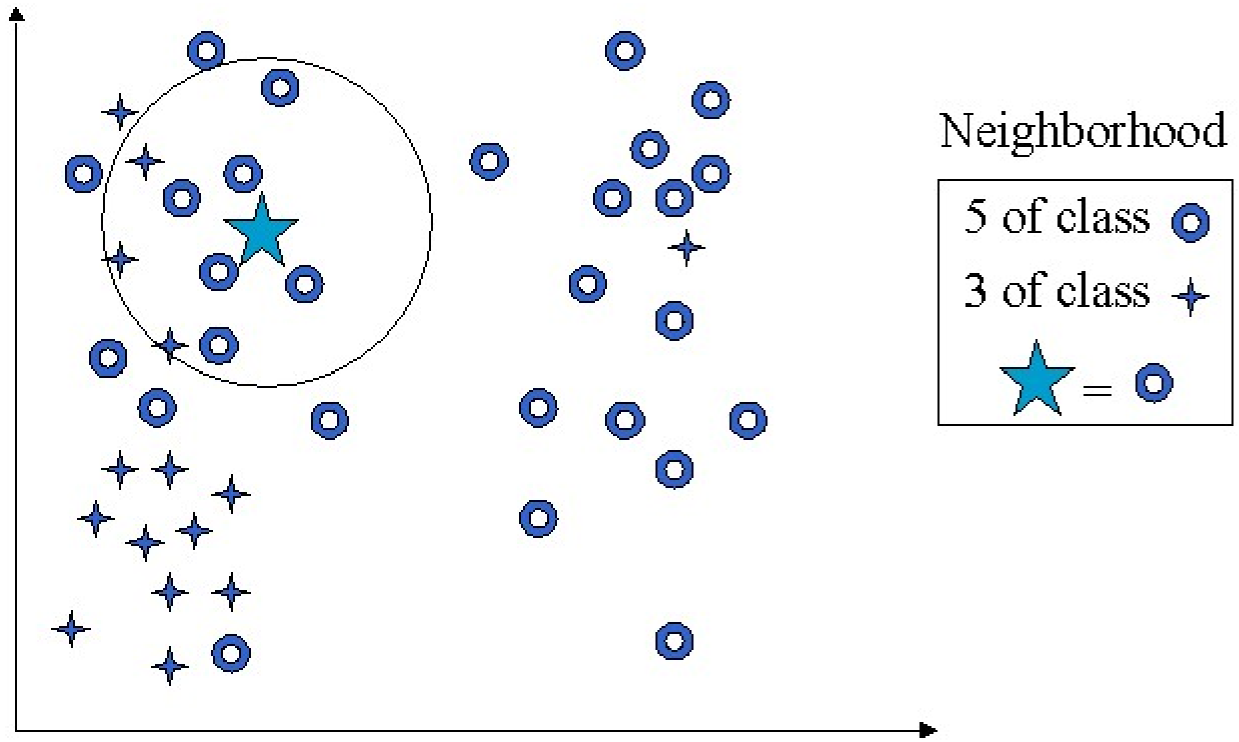
Deviation/outlier detection

•Find points that are very different from the other points in the dataset
•Could be "noise", that causes problems for classification or clustering
•Could be the really "interesting" points, for example, in fraud detection, we are mainly interested in finding the deviations from the norm



K-nearest neighbors

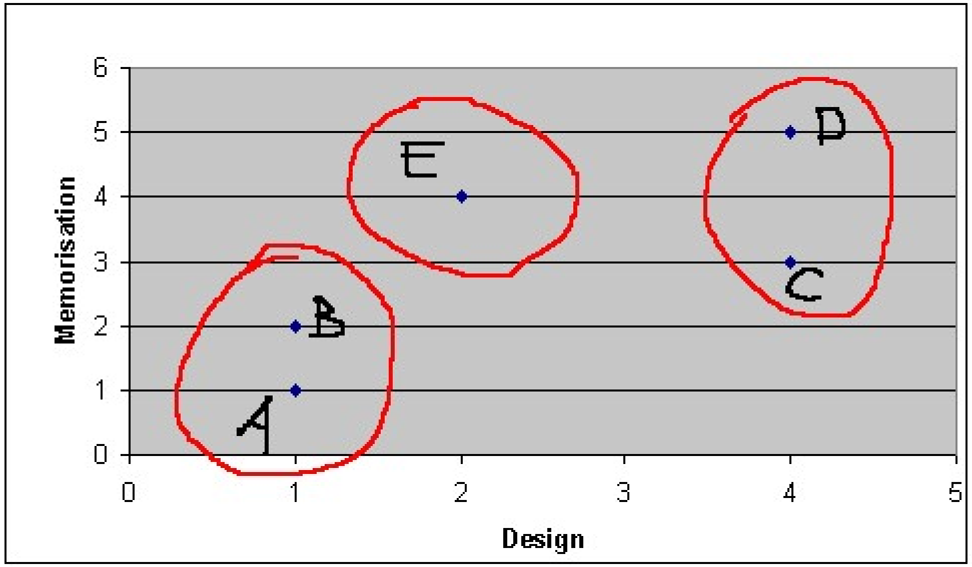
- Classification technique to assign a class to a new example
- Find k-nearest neighbors, i.e., most similar points in the dataset (compare against all points!)
- Assign the new case to the same class to which most of its neighbors belong



Clustering Example

There is many way to build cluster and to calculate distances. We take the most commun technique: eucledian distance.

	Design	Memorisation
Product A	1	1
Product B	1	2
Product C	4	3
Product D	4	5
Product E	2	4



The distance between A and B is 1 (2-1). The distance between B and E can be calculated using the following rule: $d(B,E)^2 = d(B,F)^2 + d(F,E)^2 = (4-2)^2 + (2-1)^2 = 5 \Rightarrow d(B,E) = 2.24$.

Also, we are ready to calculate the other distances:

Off course this matrix is symetric. d(A,B)=d(B,A).

We start to group the nearest to each other points. The first group AB is created. The matrix will look like this:

The way to calculate the distance C, D, E to AB is important. Of course there are many calculation way. One of them is to consider the mean distance between AB and C or to consider the distance between C and the gravity center of AB. Other way is to take the shortest distance AB and C, that means B to C.

The choose of the calculation's algorithm make the difference between different classification tools. It has a big influence of the calculation in the next iteration.

	A	B	C	D	E
A	-	1	3.61	5	3.16
B		-	3.16	4.24	2.24
C			-	2	2.24
D				-	2.24
E					-

	AB	C	D	E
AB	-	3.61	5	3.16
C		-	2	2.24
D			-	2.24
E				-

To continue our example, we consider the highest distance. The highest distance AB to C is d(A,C) = 3.61. d(B,C) = 3.16. We re
We regroup C and D, they have the shorts distance 2.
The matrix look like this:
Now we regroup CD and E, they have the shortest distance 2.24, and again the matrix look like this:

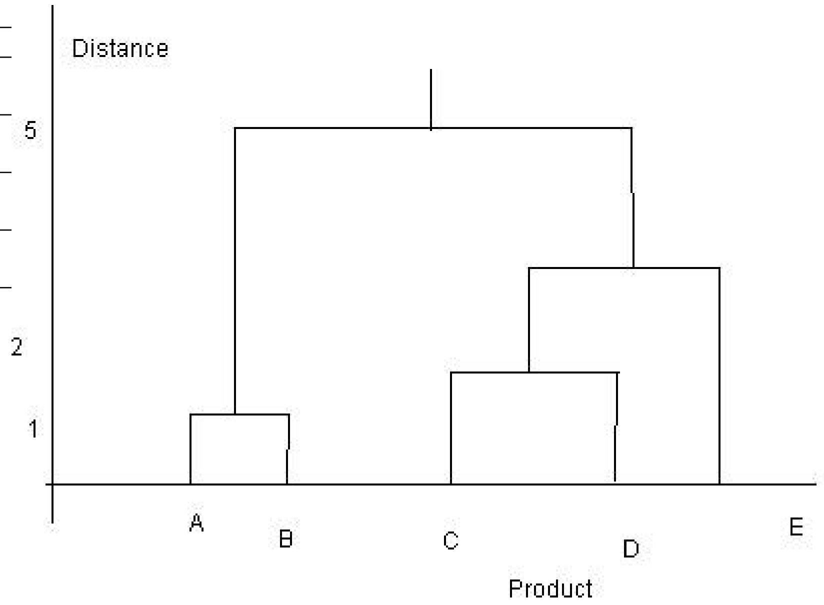
The grouping work is finished, now we are ready to build the classification tree based on the calculate distance.
In the following graphic (called dendogramm) the x-axis are the product and y-axis are the distances.

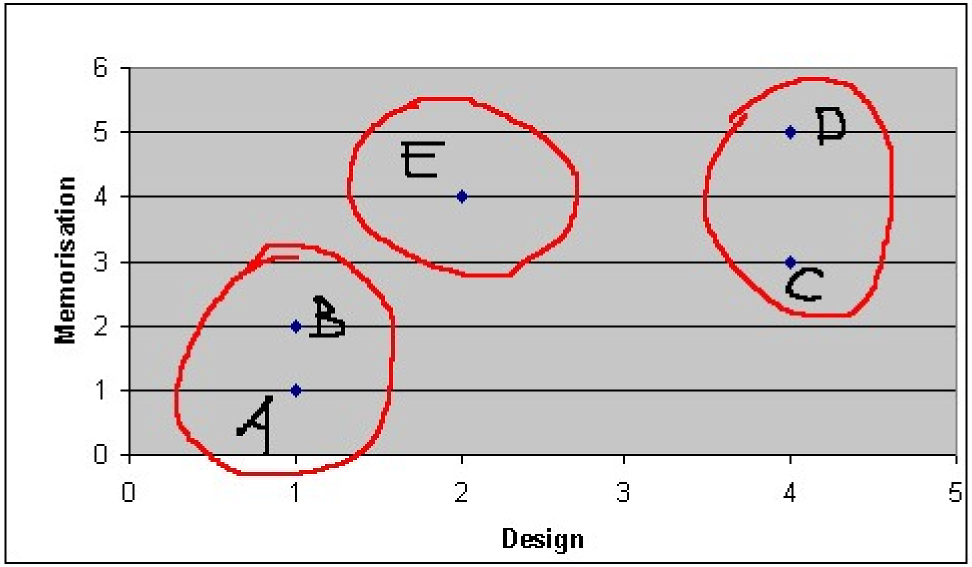
Dendogramm based on the minimal euclide-distance.

	AB	C	D	E
AB	-	3.61	5	3.16
C		-	2	2.24
D			-	2.24
E				-

	AB	CD	E
AB	-	5	3.16
CD		-	2.24
E			-

	AB	CDE
AB	-	5
CDE		-





If there is more than two variable, the distance can be calculate according to the following rule: $\sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$

This is an extension of the Pythagore theorem.
The distance is used as grouping factor of the population. If the distance is short, the population is considered to be homogen.