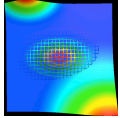


Nagy adathalmazok labor

2018-2019 ősz

2018.09.04

1. Alapok
2. Klasszifikáció vs. regresszió



Mindenféle

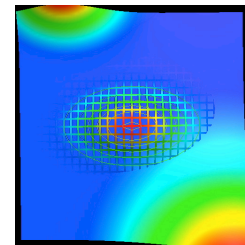
Bálint Daróczy

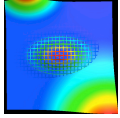
daroczyb@ilab.sztaki.hu

Alapvetően: MTA SZTAKI, Lágymányosi u. 11

Honlap:

<http://cs.bme.hu/~daroczyb/dm/>





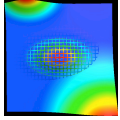
Követelmények

szerdánként 8:00-9:30 SZIT labor
kéthetente 8:00-9:30 gyakorlat

Zárthelyi: november vége

Projekt: önálló vagy kiscsoport

Megajánlott jegy

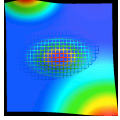


Ajánlott irodalom

1. Tan, Steinbach, Kumar (TSK): Introduction to Data Mining
Addison-Wesley, 2006, Cloth; 769 pp, ISBN-10: 0321321367, ISBN-13:
9780321321367
<http://www-users.cs.umn.edu/~kumar/dmbook/index.php>
2. Leskovic, Rajraman, Ullmann: Mining of Massive Datasets
<http://infolab.stanford.edu/~ullman/mmds.html>
3. Devroye, Györfi, Lugosi: A Probabilistic Theory of Pattern Recognition, 1996
4. Rojas: Neural Networks, Springer-Verlag, Berlin, 1996
5. Hopcroft, Kannan: Computer Science Theory for the Information Age
<http://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/hopcroft-kannan-feb2012.pdf>

+ cikkek

Frameworks



Scikit

Pandas

Chainer

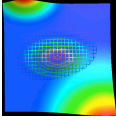
Tensorflow

Keras

Spark

Underlying: python





Projekt

Szövegbányászat:

- osztályozás
- további meta adatok
- külső adatforrások

Ajánló rendszerek:

- tartalom alapon
- collaborative filtering

Képkeresés és osztályozás

Query:

<top>

<num> C301 </num>

<HU-title> Nestlé márkák </HU-title>

<HU-desc> Mely márkákat értékesít a Nestlé világszerte? </

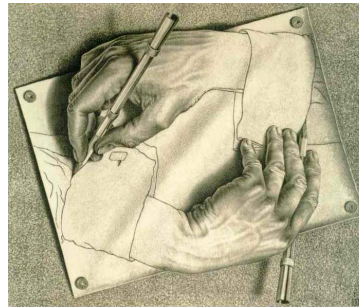
HU-desc> <HU-narr> A releváns dokumentumok olyan

árucikkek neveit tartalmazzák, amelyeket a Nestlé ill.

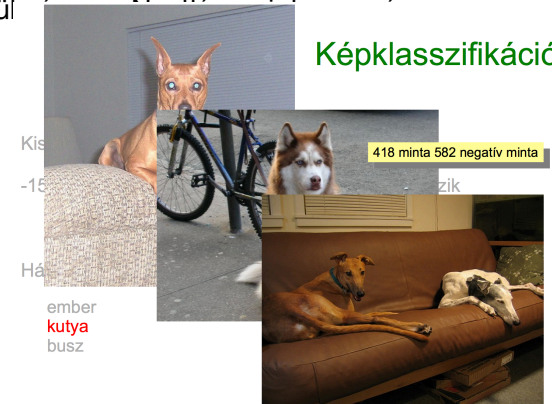
stlé-csoporthoz tartozó vállalat értékesít. Utóbbi

os utalás szül

/top>



Képklasszifikáció



Felhasználó /film	Napoleon Dynamite	Szörny RT.	Cindarell a	Life on Earth
Dávid	1	?	?	3
Dóri	5	3	5	5
Péter	?	4	3	?

Reprezentáció

- Adathalmaz **objektumok halmaza**, mely tartalmazza az **attribútumokat** is
- Reményeink szerint az attribútumok **megfelelően reprezentálják** az adott objektumot
- Attribútumok **típusai** lehetnek:
 - pl. bináris
 - nominális
 - numerikus

Attribútumok, “feature”, változók

Adat objektumok, “record”

<i>Tid</i>	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

Reprezentáció

Attribútumok, “feature”, változók

Egy adathalmaz lehet:

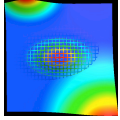
- rendezhető
 - időbeli vagy soros
 - térbeli
- ritka vagy sűrű

Pár kikötés:

- fix attribútumhalmaz
- előre meghatározott attribútumok

Adat objektumok, “
record”

<i>Tid</i>	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



Gépi tanulás

Legyen adott egy N elemű ismert halmaz, X és minden eleméhez egy y érték.

Keressük az olyan $f(x)$ függvényeket melyek jól közelítik y -t.

Sok a kérdés:

Hogyan határozzuk meg a közelítés minőségét?

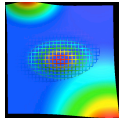
Milyen függvényosztályok között keresünk?

Hogyan találunk egy adott függvényosztályon belül megfelelő példányt?

Komplexitás?

Generalizáció?

stb.



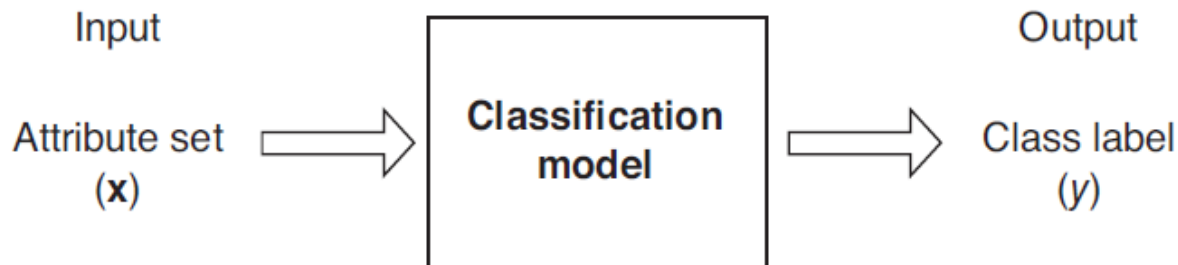
Osztályozás vs. generatív modellek

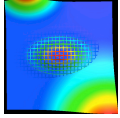
Gyakorlatban, ha a célváltozónk bináris vagy egy véges halmaza eleme, osztályozásról beszélünk.

Nominális $\rightarrow$ Bináris?

Megj.: A tanulás fontos része a visszajelzés.

Generatív modellek?

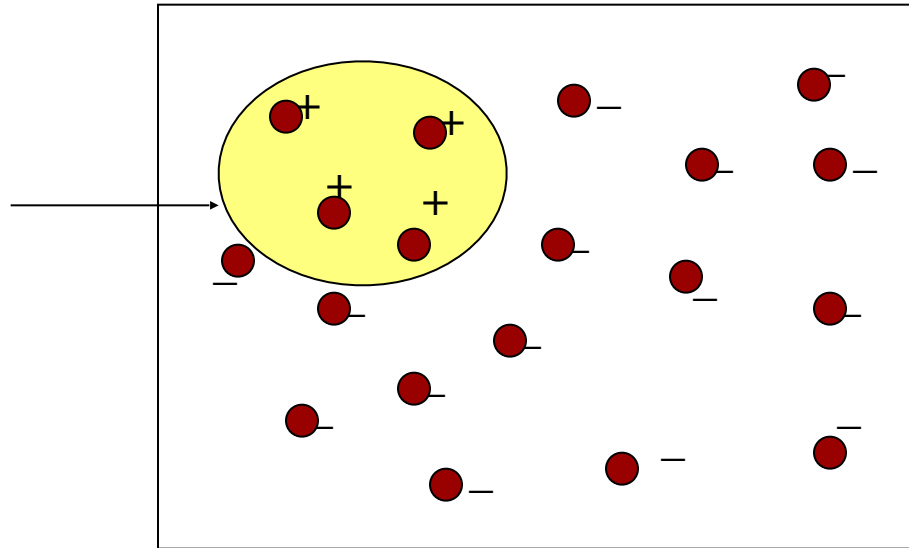




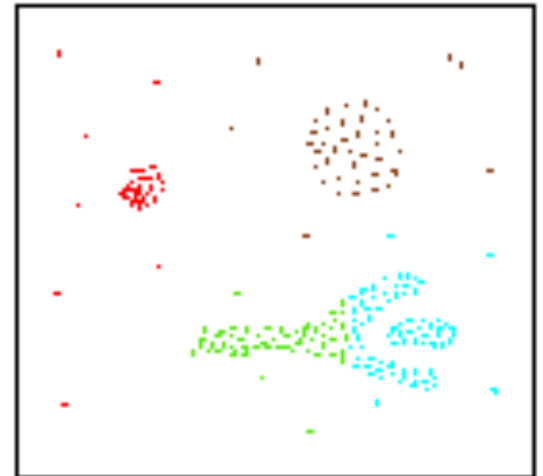
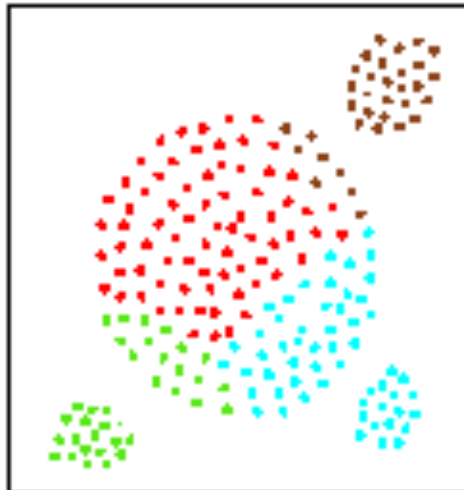
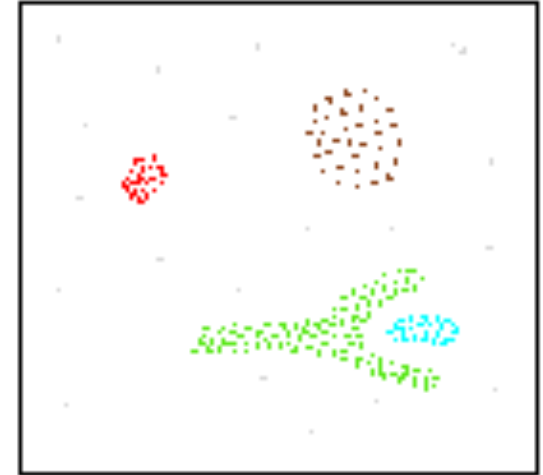
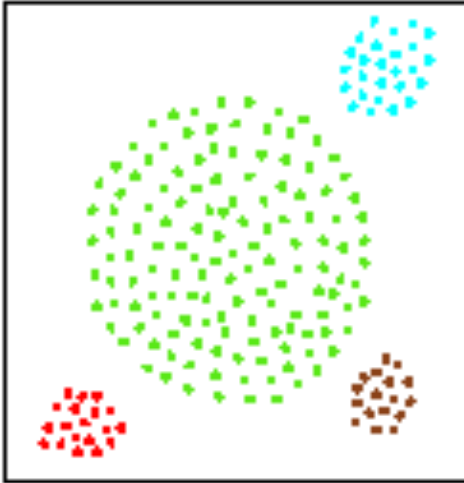
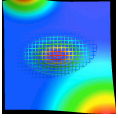
Osztályozás

Mintahalmaz

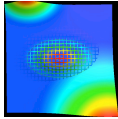
cimke



Klaszterezés



Ábra: TSK



K-Means

Feltesszük, hogy az adatpontjaink egy valós vektortérben helyezkednek el.

A klasztereket a középpontjuk határozza meg.

K-means ($D; k$)

Adott egy $C_1, C_2, \dots, C_k$ centroidok tetszőleges k elemű kezdeti halmaza

Amíg változik a centroidok halmaza:

minden d mintát soroljuk a hozzá legközelebbi centroid klaszterébe

Módosítsuk a meglévő centroidokat

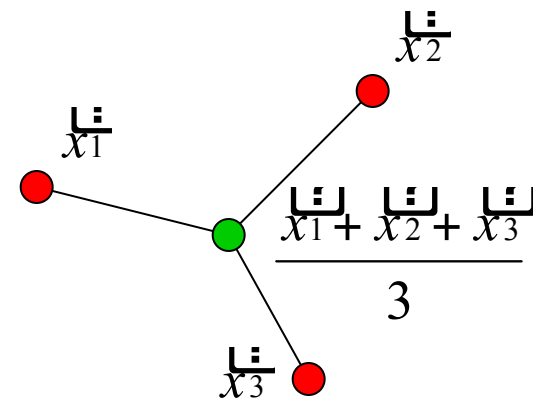
C_i legyen C_i -be tartozó elemek középértéke

A kezdőpontok meghatározása lehet:

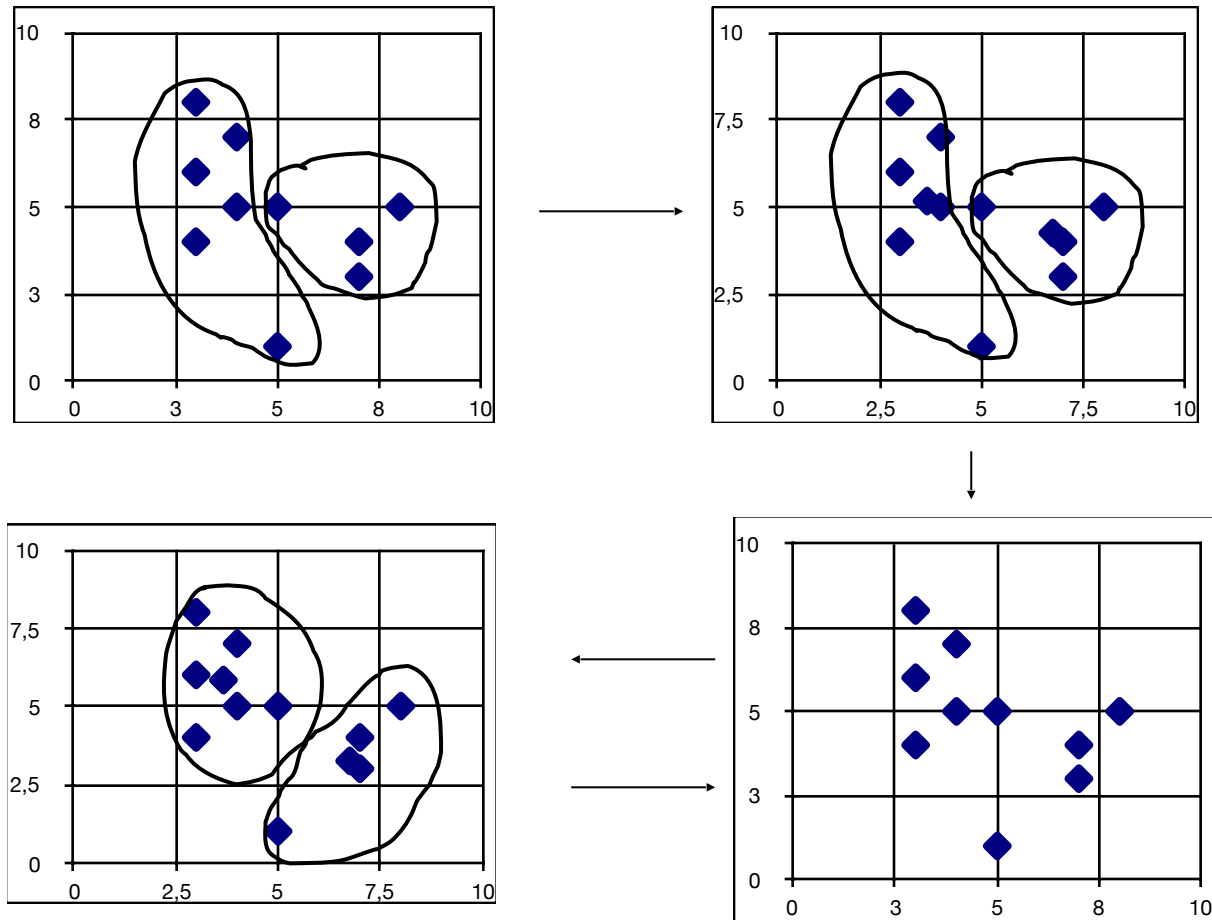
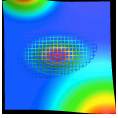
- véletlen pontokból
- véletlen választott tanulópontokból

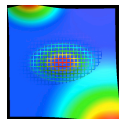
Az algoritmust megállítjuk, ha:

- a klaszterezés nem változik
- a küszöbhibánál kisebb az approx. hiba
- elértük a maximális iterációk számát

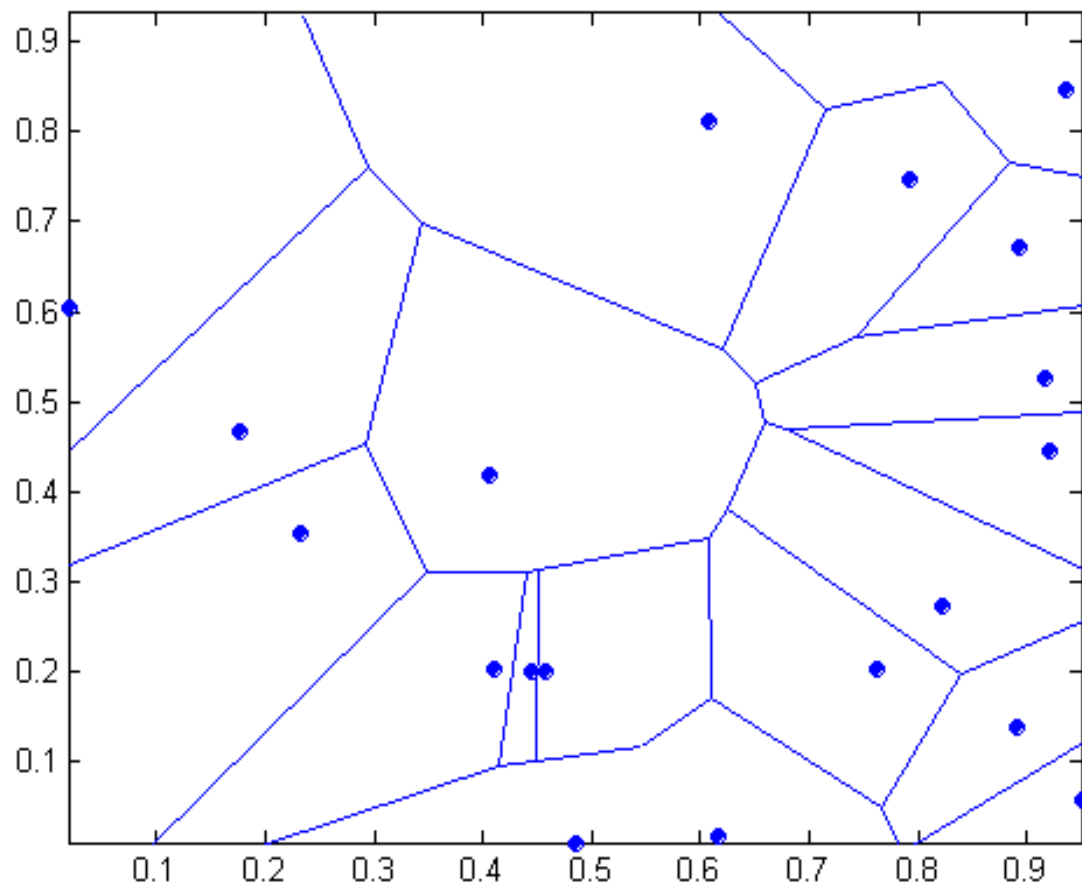


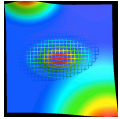
K-means





K-means





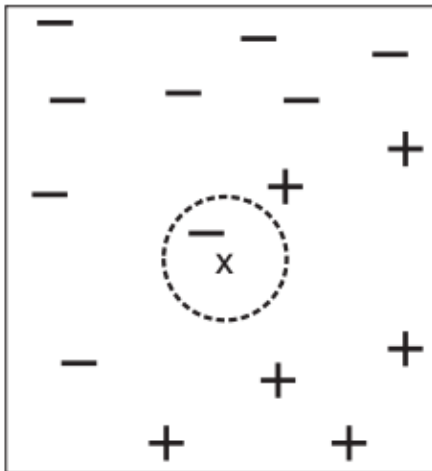
K- nearest neighbor (K-NN)

Alapelv:

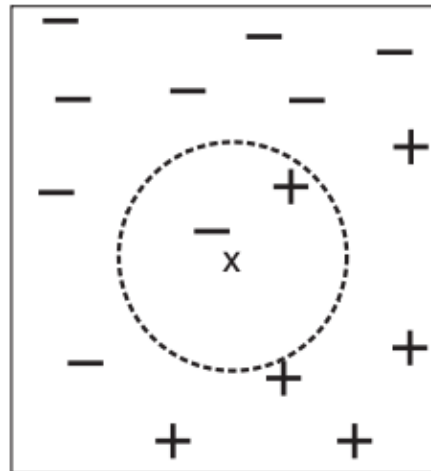
“Úgy megy mint egy kacska, úgy úszik mint egy kacska, úgy eszik mint egy kacska, tehát kacska!”

1. Minden tesztadathoz keressük meg a K legközelebbi elemet.
2. A legtöbbet reprezentált osztályt jósoljuk az adott tesztadatnak

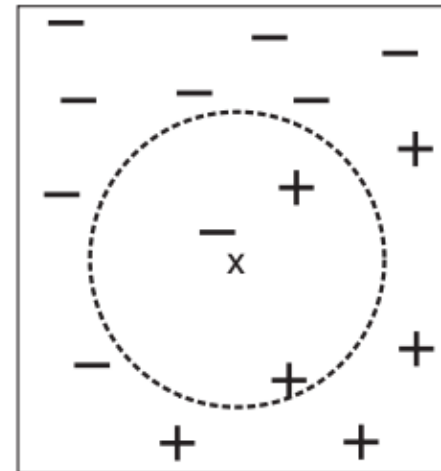
Példa:



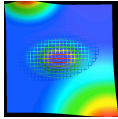
(a) 1-nearest neighbor



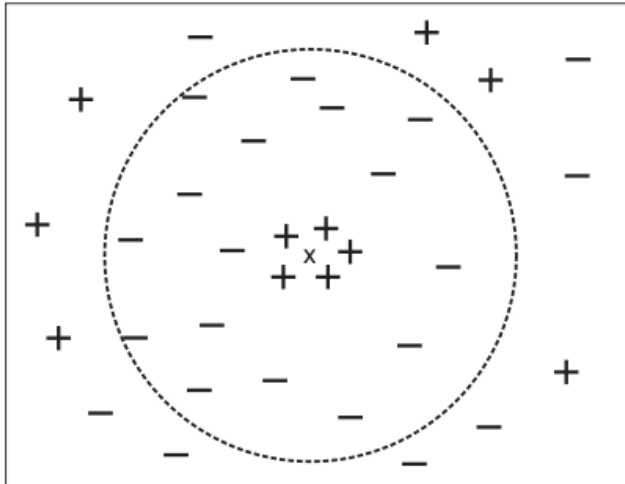
(b) 2-nearest neighbor



(c) 3-nearest neighbor



K- nearest neighbor (K-NN)



Mi a gond?

- mohó (eager) algoritmus: kész modell készítése amint a tanuló adat létezik
→ többé nincs szükség a tanulóhalmazra
- lusta (lazy) algoritmus: kész jóslatot a tanulóhalmaz segítségével készít → szükséges a tanulóhalmaz a jósláshoz

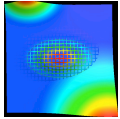
A K-NN lusta algoritmus :

Méret?

Bonyolultság?

Skálázhatóság?

Ábra: TSK



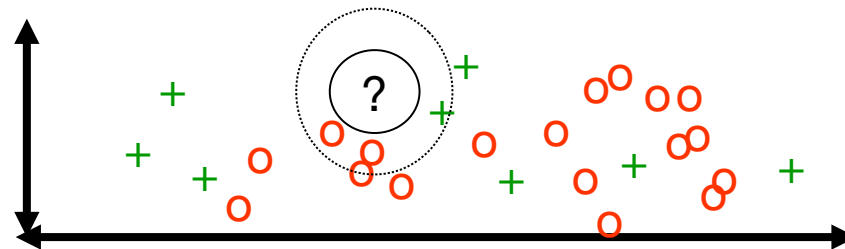
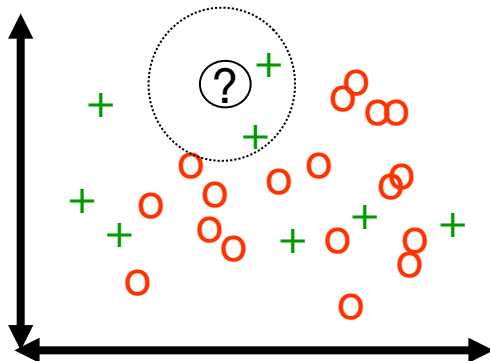
K- nearest neighbor (K-NN)

Távolság mértékek (pl.)

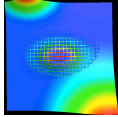
- Minkowski távolság
- Mahalanobis $D_M(x) = \sqrt{(x - \mu)^T S^{-1}(x - \mu)}$.
- Cosine, Jaccard, Kullback-Leibler, Jensen-Shannon stb.

Problémák:

- skála
- automatikus távolság választás
- normalizáció



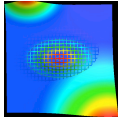
Ábra: TSK



Kiértékelés

Confusion matrix:

Alapigazság/ predikció	p	n	Total
p	True Positive (TP)	False Negative (FN)	TP+FN
n	False Positive (FP)	True Negative (TN)	FP+TN
Total	TP+FN	FP+TN	



Kiértékelés

Accuracy: a helyesen klasszifikálás valószínűsége

$$TP+TN/(TP+FP+TN+FN)$$

Precision (p): egy releváns dokumentum helyes klasszifikálásának valószínűsége

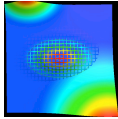
$$TP/(TP+FP)$$

Recall (r): annak a valószínűsége, hogy egy releváns dokumentumot helyesen klasszifikálunk

$$TP/(TP+FN)$$

F-measure: a precision és a recall harmónikus közepe

$$(2 \cdot p \cdot r / (p + r))$$



Kiértékelés

False-Positive Rate (FPR) =
FP/(FP+TN)

True-Positive Rate (TPR) =
TP/(TP+FN)

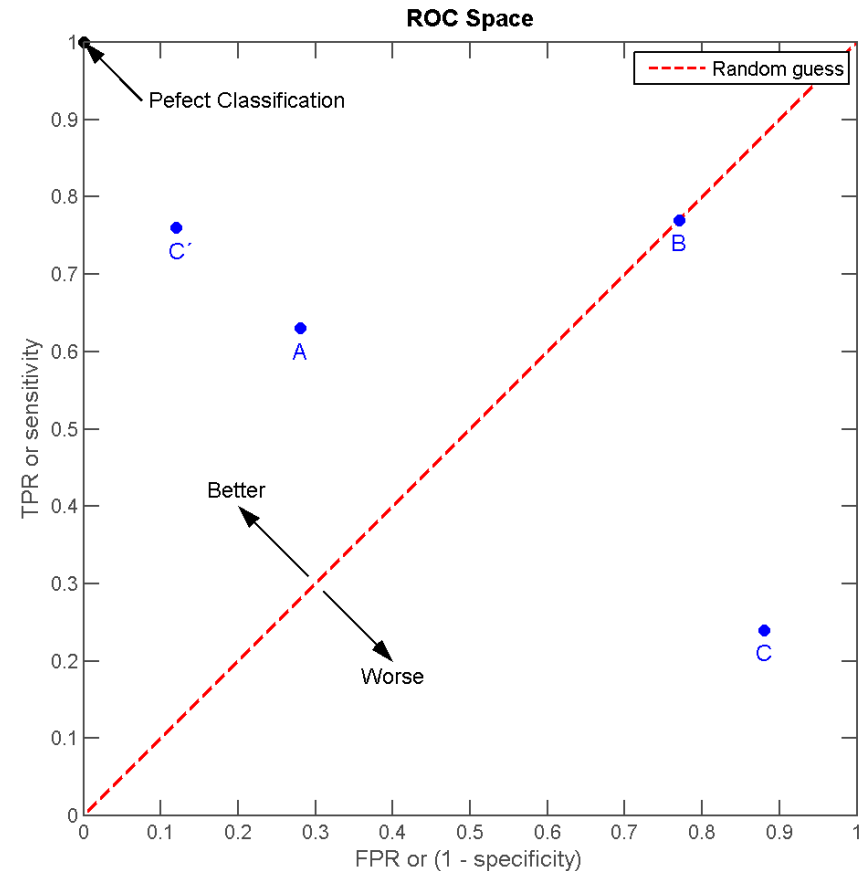
ROC: Receiver Operating
Characteristic

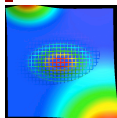
MAP: Mean Average
Precision

$$\text{AveP} = \frac{\sum_{r=1}^N (P(r) \times \text{rel}(r))}{\text{number of relevant documents}}$$

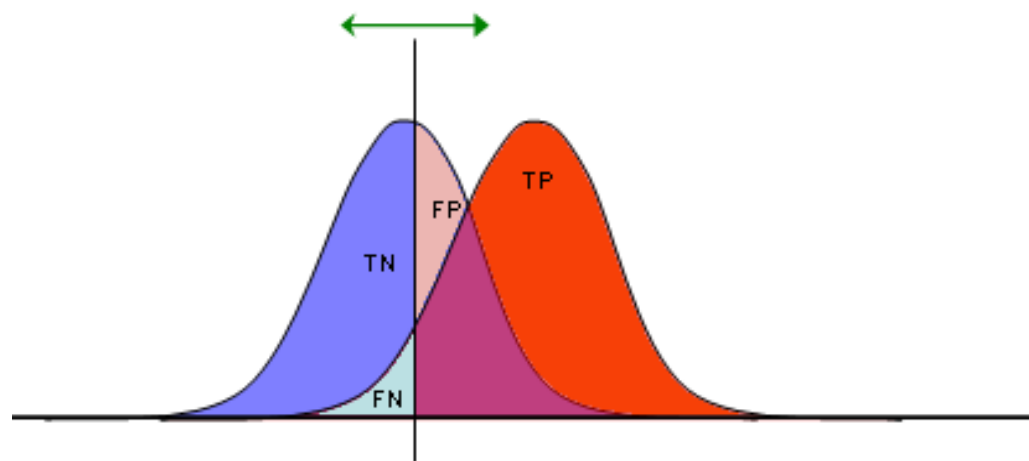
nDCG: normalized
Discriminative
Cumulative Gain

$$\text{DCG}_p = \text{rel}_1 + \sum_{i=2}^p \frac{\text{rel}_i}{\log_2(i)}$$

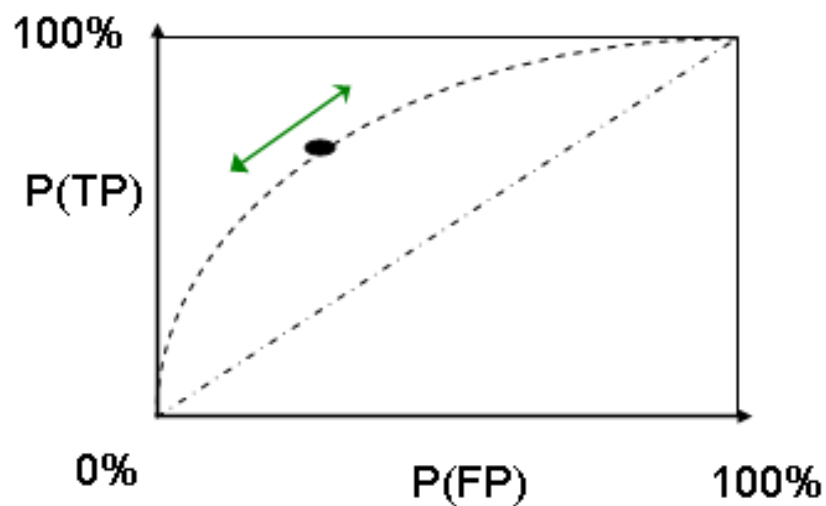


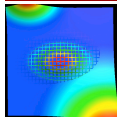


Kiértékelés



TP	FP
FN	TN
1	1





ROC: Receiver Operating Characteristic

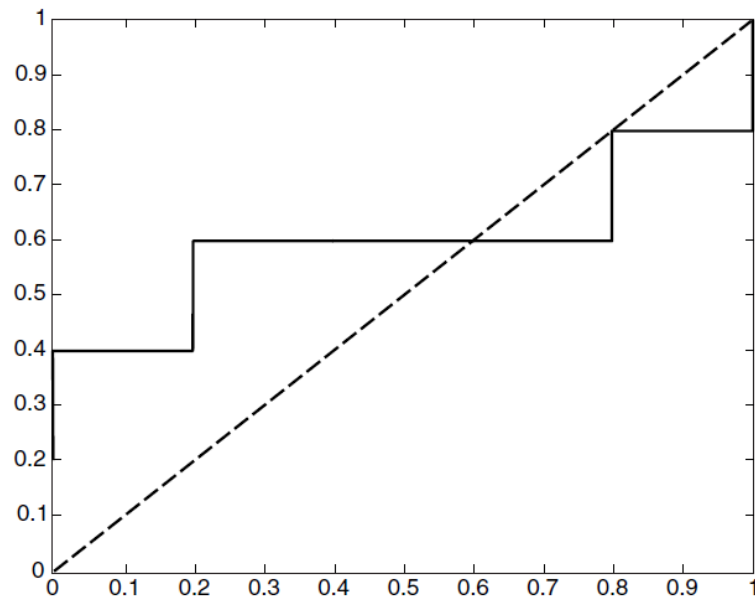
- csak **bináris** osztályozásnál használható (i.e. egy osztályra)
- **Area Under Curve:** AUC annak a valószínűsége, hogy véletlen pozitív elemeket előrébb sorol mint véletlen negatívakat
- mivel az jóslatok sorrendjéből számítjuk, a klasszifikálónak nem csak bináris jóslatokat kell visszaadnia
- előnye, hogy **nem függ a vágási ponttól**
- **Milyen szempontból lehetne továbbfejleszteni?**

Class	+	–	+	–	–	–	+	–	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC: Receiver Operating Characteristic

24

Class	+	-	+	-	-	-	+	-	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0



AUC=?

