

Egyenlőtlenségi mérőszámok alkalmazása az adatbányászatban

Hajdu Ottó

BCE: Statisztika Tanszék

BME: Pénzügyek tanszék

Budapest, 2011

Adatbányászati feladatok

1. Ismert mintákon, példákön való tanulás (extracting Knowledge)
2. Exploratív, prediktív céllal,
3. A kutatási folyamat fő mozzanatai:
 1. Adatelőkészítés: Tisztítás, transzformálás, redukció,
 2. Asszociációs szabályok feltárása: (Mi-mivel gyakran?)
 3. Klasszifikálás: (Ki milyen lesz?)
 1. Döntési fák,
 2. Bayes – módszer,
 3. Neurális hálózatok,
 4. k – legközelebbi szomszédok,
 4. Predikció, regresszió: (Mi mennyi lesz?)
 5. Klaszterezés: Partíció, hierarchikus: (Ki-kivel lesz együtt?)

Sajátos Data Mining terminológia

1. Változó: Feature, Attribute,
2. Független változó: Input,
3. Függő változó: Target value, Training value, Class variable,
4. Előrejelzett érték: Output,
5. Reziduális érték: Error,
6. Megfigyelések: Samples, Patterns, Instances, Examples,
7. Becslés: Training, Learning, Adaptation,
8. Paraméterbecslések (koefficiensek): Synaptic Weights,
9. Tengelymetszet (intercept, constant): Bias,
10. Becslési kritérium: Error function, Cost function,
11. Interakció: Higher-order neurons,
12. Transzformáció: Functional link
13. Függő változó modellezése: Supervised learning,
14. Függő változó hiánya: Unsupervised learning,
15. Kategória kimenetű változó: Concept Hierarchy.

Klasszifikálás

1. **Kategória**-kimenetek egyikének az előrejelzése
2. **Bináris** eset: Igen(1), Nem(0): pl. Csőd(1), Működés(0)
3. Jóslás, predikció:
 1. Klasszifikációs szabály: „**Rule**”
 2. → **Döntés**: Igen(1), Nem(0)
 3. → **Korrekt** (**True**) v.s. **Inkorrekt** (**False**) klasszifikációk:

Klasszifikációs mátrix: Observed v.s. Predicted kategóriák gyakoriságai

<i>Observed</i>	<i>Predicted</i>	<i>Predicted</i>	<i>Totális</i>	<i>Korrekt</i>	<i>Korrekt %</i>
	<i>0</i>	<i>1</i>		<i>%</i>	<i>mutató</i>
<i>0</i>	<i>T0</i>	<i>F1</i>	<i>N0</i>	<i>T0 / N0</i>	<i>Specificity</i>
<i>1</i>	<i>F0</i>	<i>T1</i>	<i>N1</i>	<i>T1 / N1</i>	<i>Sensitivity</i>
<i>Totális</i>			<i>N</i>		

A klasszifikációs *Cut* - szabály

Observed: 5db 1 és 4db 0,

C: cut-off érték

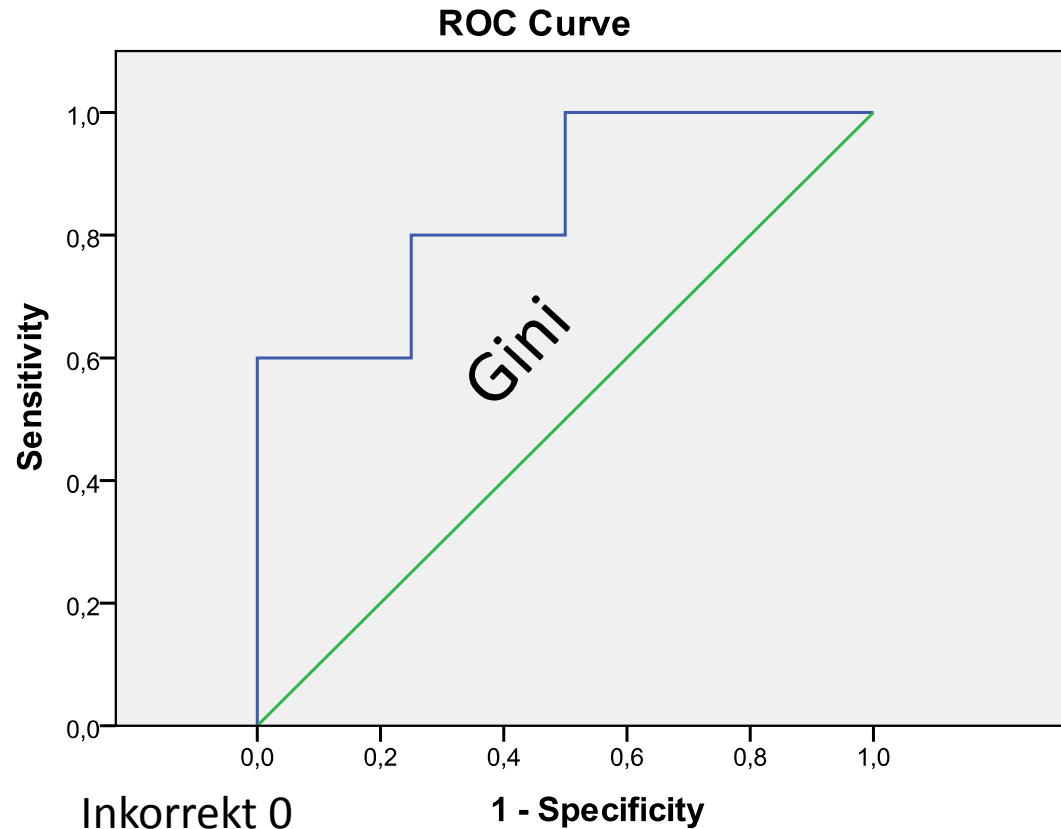
V
á
l
l
a
l
k
o
z
á
s

		<i>C =</i> 1 .85 .75 .65 .55 .45 .35 .25 .15 0										
<i>Pr(1:Csőd)</i>	<i>Obs</i>	<i>Előrejelzés</i>										
.1	0	0	0	0	0	0	0	0	0	0	0	1
.2	0	0	0	0	0	0	0	0	0	1	1	1
.3	1	0	0	0	0	0	0	0	1	1	1	1
.4	0	0	0	0	0	0	0	1	1	1	1	1
.5	1	0	0	0	0	0	1	1	1	1	1	1
.6	0	0	0	0	0	1	1	1	1	1	1	1
.7	1	0	0	0	1	1	1	1	1	1	1	1
.8	1	0	0	1	1	1	1	1	1	1	1	1
.9	1	0	1	1	1	1	1	1	1	1	1	1
Korrekt	1	0	1	2	3	3	4	4	5	5	5	5
Inkorrekt	0	0	0	0	0	1	1	2	2	3	4	4

Receiver Operating Curve (ROC)

Cut-off-érték szelektálás

Korrekt 1



Gini = 0.85

Cut	Koordináták	
	1-Specificity	Sensitivity
1.00	0.00	0.0
0.85	0.00	0.2
0.75	0.00	0.4
0.65	0.00	0.6
0.55	0.25	0.6
0.45	0.25	0.8
0.35	0.50	0.8
0.25	0.50	1.0
0.15	0.75	1.0
0.00	1.00	1.0

Konkordancia - analízis

monoton asszociációs mértékek

Minél több a ***konkordáns (0_1) páros***, annál hűbb a klasszifikációs szabály.

Konkordáns: "+", Diszkordáns: "-", Tight: "="						
Pr(Csőd)	0.3	0.5	0.7	0.8	0.9	Csődös
0.1	+	+	+	+	+	
0.2	+	+	+	+	+	
0.4	-	+	+	+	+	
0.6	-	-	+	+	+	
Működő						N_x=20

N₊ : konkordánsok,
N₋ : diszkordánsok,
N₌ : kötések
száma,
Összesen: N_x.

$$\textbf{\underline{Concordance Index}} = (N_+ + N_- / 2) / N_x = 17 / 20 = 0.85 = \textbf{\underline{Gini}}$$

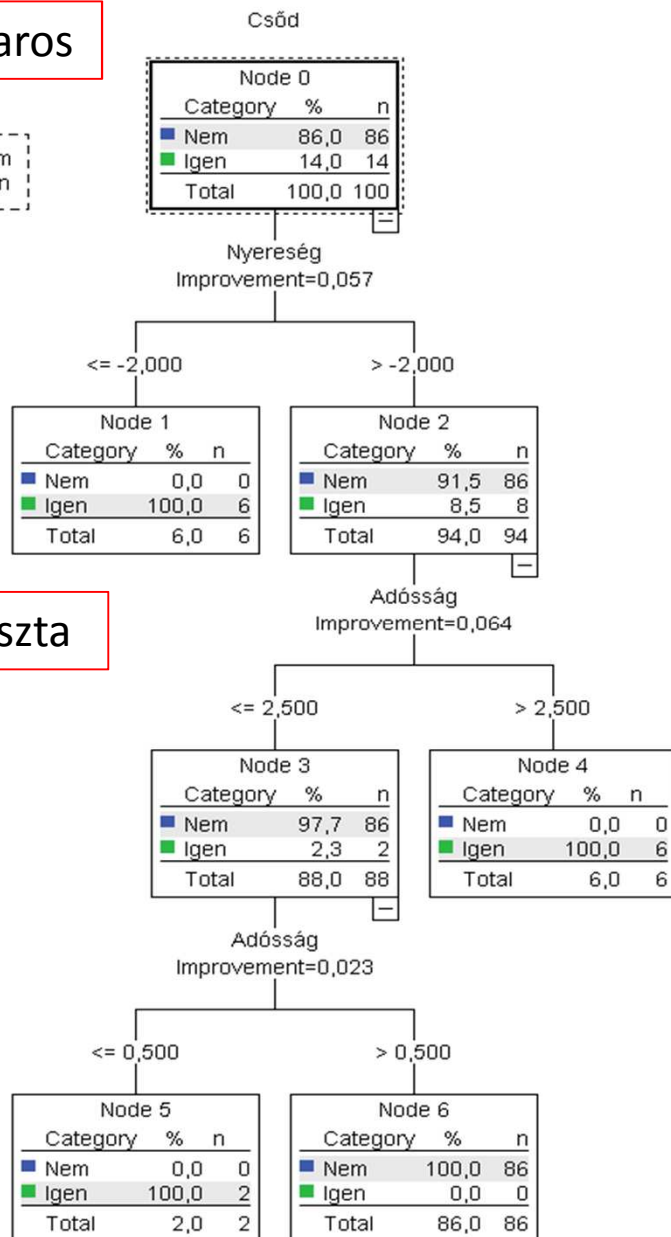
$$\text{Sommer's } D = (N_+ - N_-) / N_x, \quad = (17-3) / 20 = 0.70$$

$$\text{Goodman-Kruskal's Gamma} = (N_+ - N_-) / (N_+ + N_-), \quad = (17-3) / (17+3) = 0.70$$

$$\text{Kendall's Tau-a} = (N_+ - N_-) / (n(n-1)/2) \quad = (17-3) / (9*8/2) = 0.39$$

Zavaros

■ Nem
■ Igen



Tiszta

Klasszifikációs fák

Gini zavarossága

Node 0 _ Gini		$Nem_{Predicted}$	$Igen_{Predicted}$
	p	0.86	0.14
$Nem_{Observed}$	0.86	$Cost_{NN} = 0$	$Cost_{NI} = 0.2$
$Igen_{Observed}$	0.14	$Cost_{IN} = 1$	$Cost_{II} = 0$

Gini Node 0 =

minden párosításban:

Sum of $(p_{Obs} * p_{Pred} * Cost) =$

$$0.86 * 0.14 * 0.2 + 0.14 * 0.86 * 1 =$$

0.144.

Improvement node # =

$Prior\# * (Gini\# - \text{Átlag}(\text{Left Gini}, \text{Right Gini}))$

$Prior0 = 1$

Gain – chart

Gains (Csődök) for Nodes					
Node	Node-by-Node		Response	Cumulative	
	Node	Csőd	Csődarány	Node	Gain
	N	N		%	%
1	6	6	100	6	43
4	6	6	100	12	86
5	2	2	100	14	100
6	86	0	0	100	100

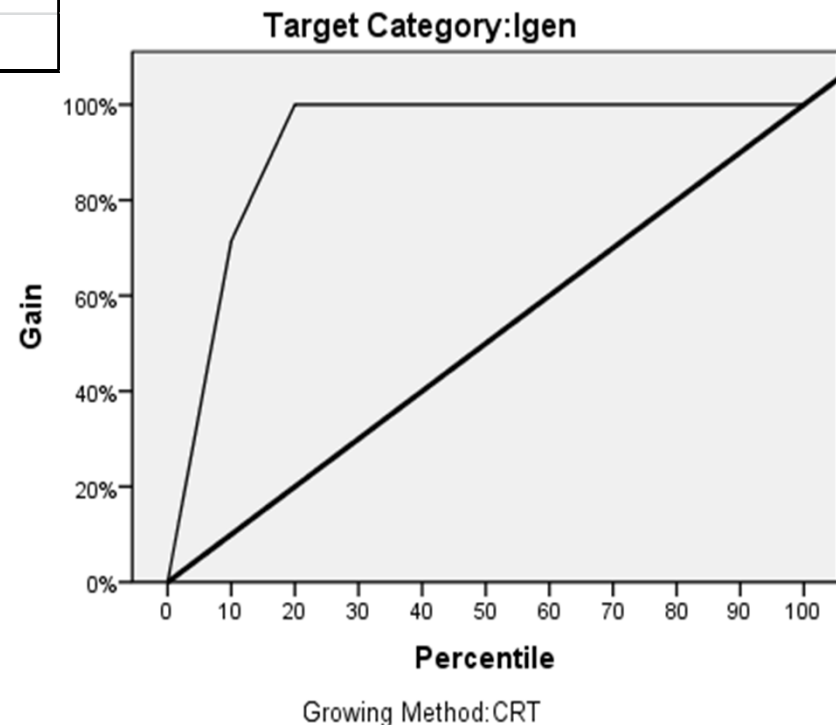
Node elrendezés:

A csődarány szerint, csökkenőleg:

Fontossági rangsor

Lorenz – görbe:

A csődök koncentrációja
a fontos terminál (levél)
Node-ok birtokában



Entrópia hasítási kritérium

Településtípus alapján

Prop (%)		0.202	0.212	0.257	0.329	1.000
		$p =$				
Munkanélküli		Budapest	Nagyváros	Többi város	Községek	Total
0		0.968	0.926	0.905	0.879	0.914
1		0.029	0.070	0.085	0.103	0.077
2		0.003	0.004	0.009	0.016	0.009
3		0.0001	0.0002	0.001	0.002	0.0009
Total		1.00	1.00	1.00	1.00	1.00
		$- p * \log(p) =$				
		Budapest	Nagyváros	Többi város	Községek	Total
0		0.031	0.071	0.090	0.114	0.083
1		0.103	0.187	0.210	0.235	0.197
2		0.016	0.021	0.041	0.066	0.042
3		0.001	0.001	0.007	0.012	0.007
Entrópia		0.15	0.28	0.35	0.43	0.33
Átlag						0.32
Javulás						0.01

**Háztartások
szegmentálása:**

Célváltozó (tisztá node):
A munkanélküliek száma
tekintetében

Nem bináris vágás:
Településtípus alapján,

A javulás az átlagos
Entrópia csökkenése: **0.01**

Optimális vágás: a legnagyobb javulást adó ismév (pl. iskola) mentén

Entrópia alapú diszkretizálás

Háztartásfő - korcsoportok kialakítása a munkanélküliség tekintetében

1. $\text{Életkor} \leq 40 \rightarrow \text{Alsó}\%$,
2. $\text{Életkor} > 40 \rightarrow \text{Felső}\%$
3. Információs nyereség: $I(40)$:
4.
$$I(40) = (A\% * Ent(A) + F\% * Ent(F)) / 100$$

$= \underline{\text{átlag}} \rightarrow \text{max}$
5. *Rekurzív módon ismétlendő, a leállási kritérium teljesüléséig*

Log – likelihood klasztertávolság

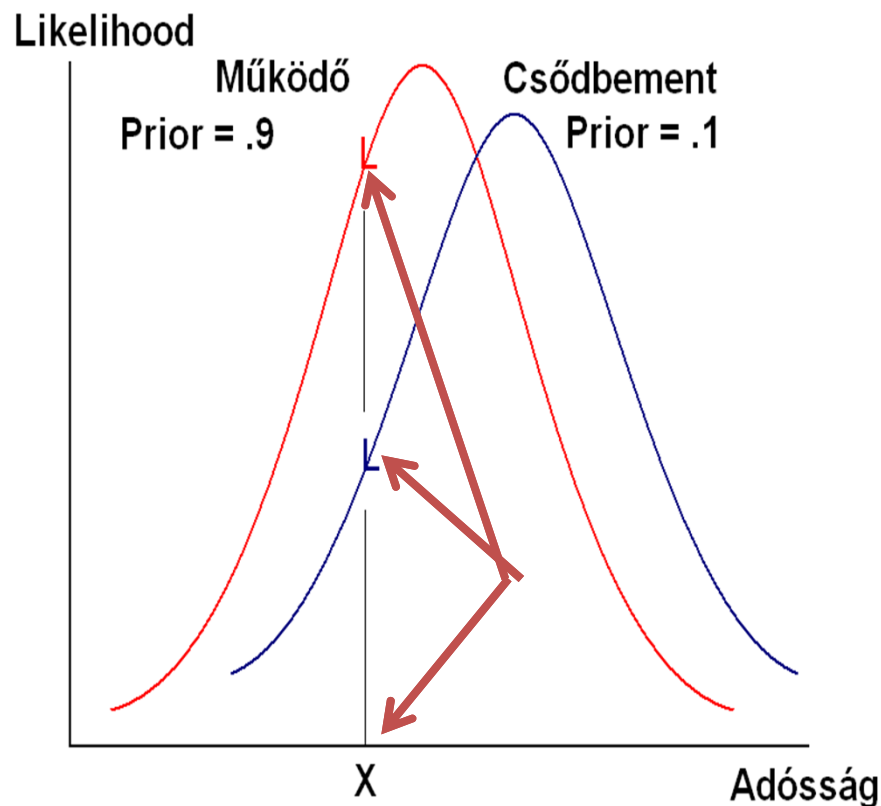
1. **Induló klaszterek**: Budapest, Nagyváros, Többi város, Községek
2. **Láncmódszer**:
3. Hierarchikusan, **agglomeratív** módon:
 1. A két legközelebbi klaszter összevonandó
 2. Az „A” és „B” klaszterek távolsága a munkanélküliek száma {0,1,2,3} tekintetében:
 1. **Metrika**: Entrópia
 2. **Linkage Rule**: $\text{Entrópia}(A) + \text{Entrópia}(B) - \text{Entrópia}(A*B)$

Entrópia	Távolságmátrix			
	Budapest	Nagyváros	Többi város	Községek
Budapest	0			
Nagyváros	$Bp+Nv - Bp*Nv$	0		
Többi város	$Bp+Tv - Bp*Tv$	$Nv+Tv - Nv*Tv$	0	
Községek	$Bp+Kö - Bp*Kö$	$Nv+Kö - Nv*Kö$	$Tv+Kö - Tv*Kö$	0

A közvetett hír információ tartalma: Cross-Entropy

Naive-Bayes metódus

TEÁOR = ####



A Hír (a posteriori):

- 1) A klasszifikálandó X adósság érték
- 2) A csődindikátor gyakorisági eloszlásai:
i) Működő, ii) Csődbement körben

A Jóslás:

Posterior (Csőd | X) =

Csődbementek %-os aránya
az X vállalkozások körében.

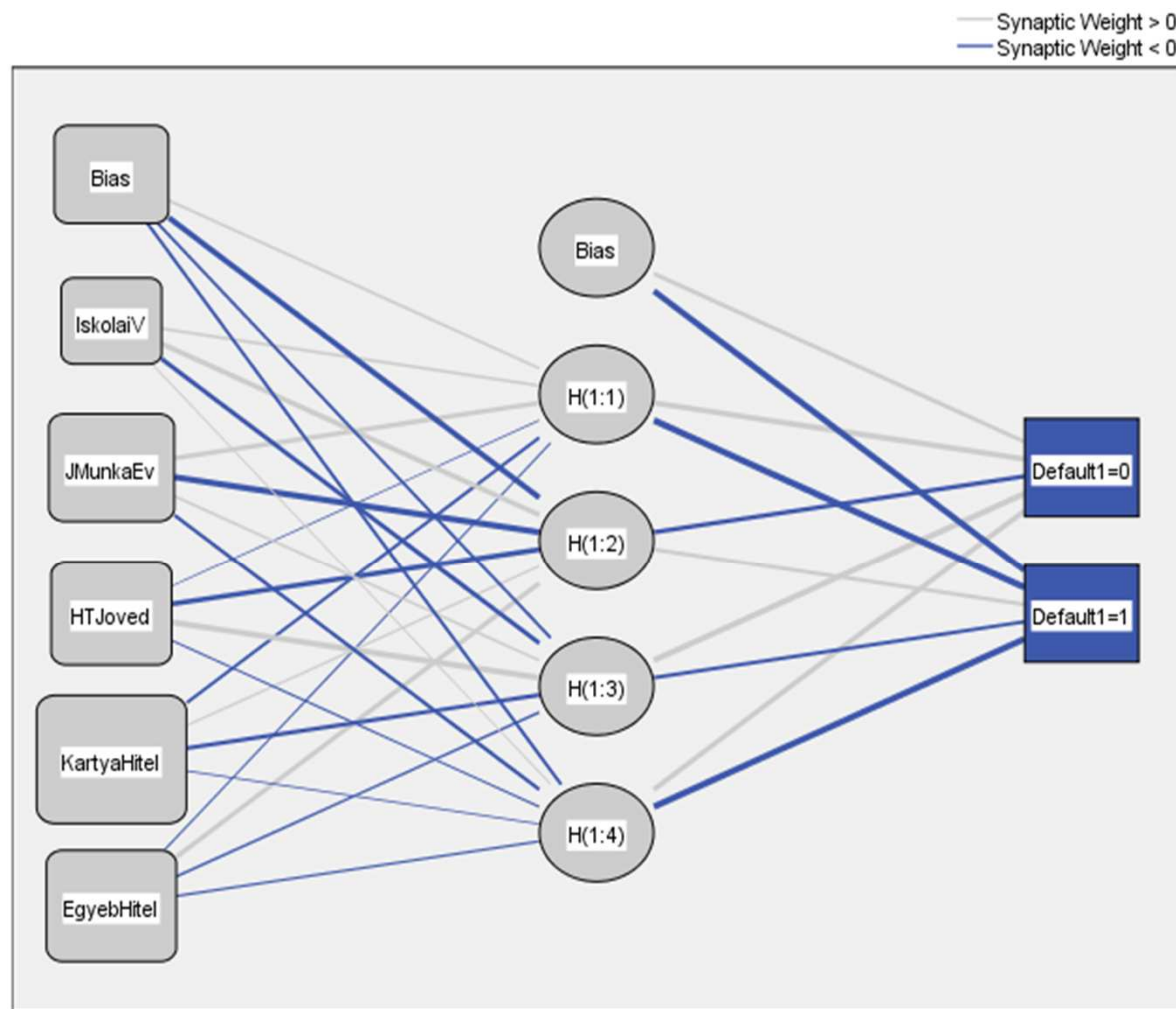
A Jóslás információja:

Cross_Entropy =

Sum [-Post * log(Prior / Post)]

Magasabb Sum érték preferált

A közvetett hír pontatlansága



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Softmax

Cross - Entropy
Error Function
Neurális hálózatok

BAR ügyfél adatok, Y:
Default1 = 0: Jó adós,
Default1 = 1: Rossz adós

A Jóslás: $Pred.Y$ | X hír

A Jóslás pontatlansága:

Error (.) =
(Sample, Pattern, Instance)
=

$Sum (-Y * \log(Pred.Y / Y))$
Alacsonyabb érték preferált