Eredmények kiértékelése

Nagyméretű adathalmazok kezelése (2010/2011/2)

Katus Kristóf, hallgató

Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék

2011. március 18.

Tartalom

- Bevezetés
- Tanítás és tesztelés
- A teljesítmény predikálása
- Kereszt-kiértékelés
- Adatbányászati módszerek összehasonlítása
- Valószínűségek predikálása
- A költségek számolása
- Numerikus predikció kiértékelése

Bevezetés

- A feladat
  - Különböző módokon nyerhetünk ki összefüggéseket
  - Melyiket használjuk az adott problémán?
  - Szisztematikus módon akarjuk a módszereinket összehasonlítani
- Probléma?
  - A tanító halmazon mért teljesítmény nem jó mérőszám
  - Független teszt halmaz kell
- Jó esetben sok adat áll rendelkezésünkre…
  - Ált. szűrni kell őket – a „minőségi” adat ritka
  - Végén kevés használható adatunk lesz tanításhoz
  - A szűréshez emberi (véges) erőforrás is kellhet
- Korlátozott mennyiségű adat esetén…
  - A használt módszerek összehasonlításához statisztikai tesztek kellenek – a véletlen szerepe

Tanítás és tesztelés

- Osztályozási feladatoknál: hiba arány – „error rate”
  - A tanuló halmazon mért teljesítmény nem valószínű, hogy megfelelő mutatója a jövőbeli teljesítménynek
  - Hiba arány a tanuló halmazon: visszahelyettesítési hiba – „resubstitution error”
- Teszt halmaz
  - A tanító halmaztól független adathalmaz
  - Feltételezzük, hogy a tanító és a teszt halmaz is reprezentatív
  - A teszt halmaz jellege eltérő lehet a tanító halmazénál
    - Ld. „credit risk problem”
  - Semmilyen módon nem használható fel az osztályozó létrehozására!

A validáló halmaz

- Néhány tanító eljárás két állomást igényel
- Különböző tanító sémákat szeretnénk összehasonlítani
- Így beszélhetünk:
  - Tanító halmazról – osztályozók felépítése
  - Validáló halmazról – paraméteroptimalizálás vagy egy osztályozó kiválasztása
  - Teszt halmazról – a végső, optimalizált módszer hibaarányának meghatározására
  - Ezek a halmazok egymástól függetlenek kell, legyenek!
- A hibaarány kiszámítása után megtehetjük, hogy a teszt halmazt és a tanító halmazt újra egybeolvasztjuk egy új osztályozó felépítéséhez – ezt fogjuk ténylegesen használni
- A paraméteroptimalizálás után a validáló halmaz is visszaolvasztható

Sok adat, kevés adat

- Sok adat esetén nincs probléma
  - Nagy, független halmazok
  - Ha a tanító és teszt halmaz reprezentatív, jó indikátor lesz a hibaarány a jövőbeli adatokra
  - Minél nagyobb a tanító halmaz, annál jobb az osztályozó – de…
  - Minél nagyobb a teszt halmaz, annál pontosabb a hibabecslés – számszerűsíthető
- Kevés adat esetén probléma van
  - Előállhat, ha pl. a tanító és teszt adatokat kézileg kell osztályoznunk
  - Hogyan hozzuk ki a legtöbbet egy limitált adathalmazból?
  - „Holdout” eljárás – dilemma: tanításhoz, validáláshoz és teszteléshez is kellene külön-külön sok adat

A teljesítmény predikálása

- Teszt halmazon 75%-os sikerességi arányt („success rate”) mérünk
- A jövőbeli adatokon ez mekkora lesz? $\pm 5\%$? $\pm 10\%$?
- Bernoulli folyamat – pl. pénzfeldobás cinkelt érmével
  - Igazi, de ismeretlen sikerességi arány: p
  - N kísérletből S sikeres
  - A megfigyelt sikerességi arány: $f = S/N$
- Konfidencia intervallum
  - Pl. $S = 750$, $N = 1000$ és 80% konfidencia esetén $73{,}2\% \leq p \leq 76.7\%$
  - Pl. $S = 75$, $N = 100$ és 80% konfidencia esetén $69{,}1\% \leq p \leq 80.1\%$
- Egy Bernoulli kísérlet esetén a sikerességi arány
  - várható értéke: p
  - szórásnégyzete: $p(1-p)$
- N kísérlet esetén a várható érték nem módosul, a szórásnégyzete pedig $p(1-p)/N$ lesz

A konfidencia intervallum (1)

- $N \to \infty$ esetén normális eloszlást kapunk
- $\Pr[-z \leq X \leq z] = c$ és normális eloszlás esetén a c-khez tartozó z-k megtalálhatóak táblázatba foglalva (X várható értéke 0)
- $\Pr[X \geq z]$-t szokták megadni („one-tailed” probability) – 0 várható érték és 1 szórás mellett

$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

- Táblázatból kiolvasható: $\Pr[-1{,}65 \leq X \leq 1{,}65] = 90\%$

A konfidencia intervallum (2)

- Standardizált képzése:

$$\Pr\left[-z < \frac{f-p}{\sqrt{p(1-p)/N}} < z\right] = c$$

- Adott c esetén képezzük $(1-c)/2$-t, és megkeressük a hozzá tartozó z-t
- A konfidencia intervallum alsó és felső határának meghatározása:

$$p = \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) \Big/ \left(1 + \frac{z^2}{N}\right)$$

- Feltételeztük, hogy közel normális eloszlással dolgoztunk – a fenti levezetés csak nagy N-ekre igaz (pl. $N > 100$)

Kereszt-kiértékelés

- Kevés az adat, „holdout” módszer – ált. az adat egyharmadát tesztelésre, a maradékot tanításra használjuk
- Ha nincs szerencsénk, a tanításra használt minta nem reprezentatív – ált. nem megállapítható
- „Stratification” és „stratified holdout” – osztályok arányos reprezentálása
- „Repeated holdout” – átlagos hibaarányt számolunk
- Kereszt-kiértékelés
  - Fix számú partíciót használ
  - Pl. „stratified 10-fold cross-validation” – standard módszer
  - Sem a rétegzésnek, sem a 10 részre való felosztásnak nem kell pontosnak lennie!
- Többszöri kereszt-kiértékelés

Leave-one-out kereszt-kiértékelés

- Annyi partícióra osztjuk fel a bemeneti halmazt, amennyi annak számossága (n)
- Pro
  - A lehető legtöbb adatot használjuk tanításra
  - Determinisztikus – nincs véletlen mintavételezés
  - Ismételni nem szükséges – ugyanazt kapjuk
- Kontra
  - n-szer történik végrehajtás
  - A teszt halmazban garantálja a nem rétegzett mintát („nonstratified sample”)
  - Pl. teljesen véletlen bemeneti adathalmazban kétfajta osztály egyenlő mértékben reprezentált…

Bootstrap

- A tanító halmazt visszatevéses mintavételezéssel hozzuk létre
- 0,632 bootstrap
  - n elemű adathalmazt n-szer mintavételezünk visszatevéssel
  - A maradékot a teszt halmaz fogja használni
  - Annak a valószínűsége, hogy egy példányt egyáltalán nem választunk ki:

$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0{,}368$$

  - Ekkor a teszt halmazon mért hiba arány pesszimista becslést adna, ezért:

$$\mathrm{e} = 0{,}632 \cdot \mathrm{e}_{\text{teszt példányok}} + 0{,}368 \cdot \mathrm{e}_{\text{tanító példányok}}$$

  - Többszöri ismétlés után kapjuk az átlagos hiba arányt
- Nagyon kis adathalmazok esetén az egyik legjobb módszer
- De pl. ld. előző példa, a végső hibaarányra a valós 50% helyett a következő optimista becslést kapnánk:

$$0{,}632 \cdot 50\% + 0{,}368 \cdot 0\% = 31{,}6\%$$

Adatbányászati módszerek összehasonlítása

- Használjunk kereszt-kiértékelést, válasszuk azt, amelyiken kisebb a becsült hiba arány. Elég?
- Vajon mennyire pontos a becslés? Jó mindig a kisebb hiba arányút választani?
- Student's t-test: a kereszt-kiértékelések során nyert mintahalmazok hibaarányának átlagát szeretnénk összehasonlítani két különböző tanítási séma esetén – szignifikánsan eltér a két eredmény?
- Paired t-test: ugyanazt a kereszt-kiértékelési kísérletet használva az eredményeket párba tudjuk állítani…
- Jelölések
  - Egymás utáni 10-fold kereszt-kiértékelések után kapott független minták halmaza…
    - …az első tanítási módszer esetén: $x_1, x_2, \ldots, x_k$, átlag: $\bar{x}$
    - …a második tanítási módszer esetén: $y_1, y_2, \ldots, y_k$, átlag: $\bar{y}$
  - Elegendő minta esetén $\bar{x}$ normális eloszlású, μ legyen a valódi középérték

A Student eloszlás

- Ha ismernénk a szórásnégyzetet, standardizálás után meg tudnánk állapítani a konfidencia intervallumot
- Csak becsülni tudjuk a minták alapján: σ_x^2/k, így $\bar{x}$ standardizáltja:

$$\frac{\bar{x}-\mu}{\sqrt{\sigma_x^2/k}}$$

- Ez már nem mutat normális eloszlást – Student eloszlást kaptunk $k-1$ szabadsági fokkal
- A Student eloszlás konfidencia határai 9 szabadsági fok esetén:

$\Pr[X \geq z]$	z
0.1%	4.30
0.5%	3.25
1%	2.82
5%	1.83
10%	1.38
20%	0.88

Párosított t-próba

- Párosított minták esetén: $d_i = x_i - y_i$ és $\bar{d} = \bar{x} - \bar{y}$
- Null hipotézis: ha $\bar{x} = \bar{y}$ akkor $\bar{d} = 0$, vagy legalábbis adott határon belül mozog
- Adott konfidencia szint mellett megnézzük, hogy a különbség meghaladja-e a konfidencia határt (utóbbit 5% vagy 1%-nak szokták választani)
- t-statisztika:

$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2 / k}}$$

- Megj.: 1% esetén a 0,5%-nak megfelelő z értéket olvassuk ki a táblázatból („two-tailed test”)

Nem párosított t-próba

- Mi van akkor, amikor a megfigyelések nem állnak párban?
  - Nem párosított t-próbát használunk
  - Felt. $\bar{x}$ és $\bar{y}$ normális eloszlást mutat $\rightarrow \bar{x} - \bar{y}$ is
  - $\bar{x} - \bar{y}$ szórásnégyzetének legjobb becslője k és l darab minta esetén:

$$\frac{\sigma_x^2}{k} + \frac{\sigma_y^2}{l}$$

  - t-statisztikában ezt használjuk, szabadsági fok a két minta szabadsági fokának minimuma

Problémák véges méretű adathalmazok esetén

- Osszuk fel és az egyes partíciókon végezzünk kereszt-kiértékelést, vagy
- használjuk fel újra az adatokat – nem kapunk független adathalmazokat
- Korrigált újramintavételezett t-próba – heurisztikán alapszik, gyakorlatban működőképesnek bizonyult
  - A „holdout” módszert alkalmazzuk k ismétléssel
  - Különböző véletlen felosztások mellett:
    - n_1 legyen a tanításhoz használt példányok száma
    - n_2 legyen a teszteléshez használt példányok száma
    - $d_i$különbségeket a teszt halmazon számoljuk

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\sigma_d^2}}$$

- 10-fold kereszt-kiértékelés esetén 10 ismétlés mellett: $k = 100, n_2/n_1 = 0{,}1/0{,}9$ és σ_d^2 100 különbségen alapszik

Valószínűségek predikálása

- Eddig csak helyes és nem helyes osztályozásról beszéltünk
- Érdemes lehet megmondani, hogy mennyire vagyunk biztosak egy példány besorolásában

A négyzetes veszteség-függvény

- Egy példány esetén k különböző kimenetelünk van
- Adott példányhoz: $p_1, p_2, \ldots, p_k$
- Egyetlen kimenetel: $a_1, a_2, \ldots, a_k$ (i. komponens 1)

$$\text{Egy példányra:} \quad \sum_j (p_j - a_j)^2 = 1 - 2p_i + \sum_j p_j^2$$

- Minimalizálás esetén a legjobb választás: $p_i = \Pr[\text{osztály} = i]$, vagy becslés
  - Legyenek a valódi valószínűségek $p_1^*, p_2^*, \ldots, p_k^*$, így $p_i^* = \Pr[\text{osztály} = i]$
  - Ekkor:

$$E\left[\sum_j (p_j - a_j)^2\right] = \sum_j \left(E[p_j^2] - 2E[p_j a_j] + E[a_j^2]\right)$$

$$= \sum_j \left(p_j^2 - 2p_j p_j^* + p_j^*\right) = \sum_j \left(\left(p_j - p_j^*\right)^2 + p_j^*\left(1 - p_j^*\right)\right)$$

Az információ veszteség-függvény

$$-\log_2 p_i$$

- Várható értéke: $-p_1^* \log_2 p_1 - p_2^* \log_2 p_2 - \cdots - p_k^* \log_2 p_k$
- Minimalizálható $p_j = p_j^*$ választással
- Szerencsejáték hasonlat
- „Zero-frequency problem”

Melyik veszteség-függvényt használjuk?

- A négyzetes veszteség-függvény
  - Nemcsak p_i-t, hanem a p_j-ket is figyelembe veszi
- Az információ veszteség-függvény
  - Csak p_i-től függ
  - Bünteti, ha egy osztályhoz kis valószínűséget rendelünk

A rossz osztályozás költsége

- Néhány példa
- A különböző kimenetelek két osztályos predikció esetén:

		Predicted class: yes	Predicted class: no
Actual class	yes	true positive	false negative
	no	false positive	true negative

- True Positive Rate $= \frac{TP}{TP+FN}$ | False Positive Rate $= \frac{FP}{FP+TN}$
- Overall Success Rate | Error Rate

A konfúziós mátrix

- Több osztály esetén konfúziós mátrixot használunk
- Egy három osztályos predikció különböző kimenetelei (aktuális és várható):

		Predicted class			
		a	b	c	Total
Actual class	a	88	10	2	100
	b	14	40	6	60
	c	18	10	12	40
	Total	120	60	20	

(a)

		Predicted class			
		a	b	c	Total
Actual class	a	60	30	10	100
	b	36	18	6	60
	c	24	12	4	40
	Total	120	60	20	

(b)

- Kappa statisztika, itt pl.

$$\frac{140 - 82}{200 - 82} = 49{,}2\%$$

Költség-érzékeny osztályozás

- Alapértelmezett költség-mátrixok

		Predicted class	
		yes	no
Actual class	yes	0	1
	no	1	0

(a)

		Predicted class		
		a	b	c
Actual class	a	0	1	1
	b	1	0	1
	c	1	1	0

(b)

- Összköltséget számolhatunk adott teszt halmazon adott modell mellett
- a, b, c osztályok esetén legyenek p_a, p_b, p_c, költség mátrix pedig (b) táblázat
- a predikálása esetén a predikció költségének várható értéke:

$$[0, 1, 1] \cdot [p_a, p_b, p_c]^T = p_b + p_c = 1 - p_a$$

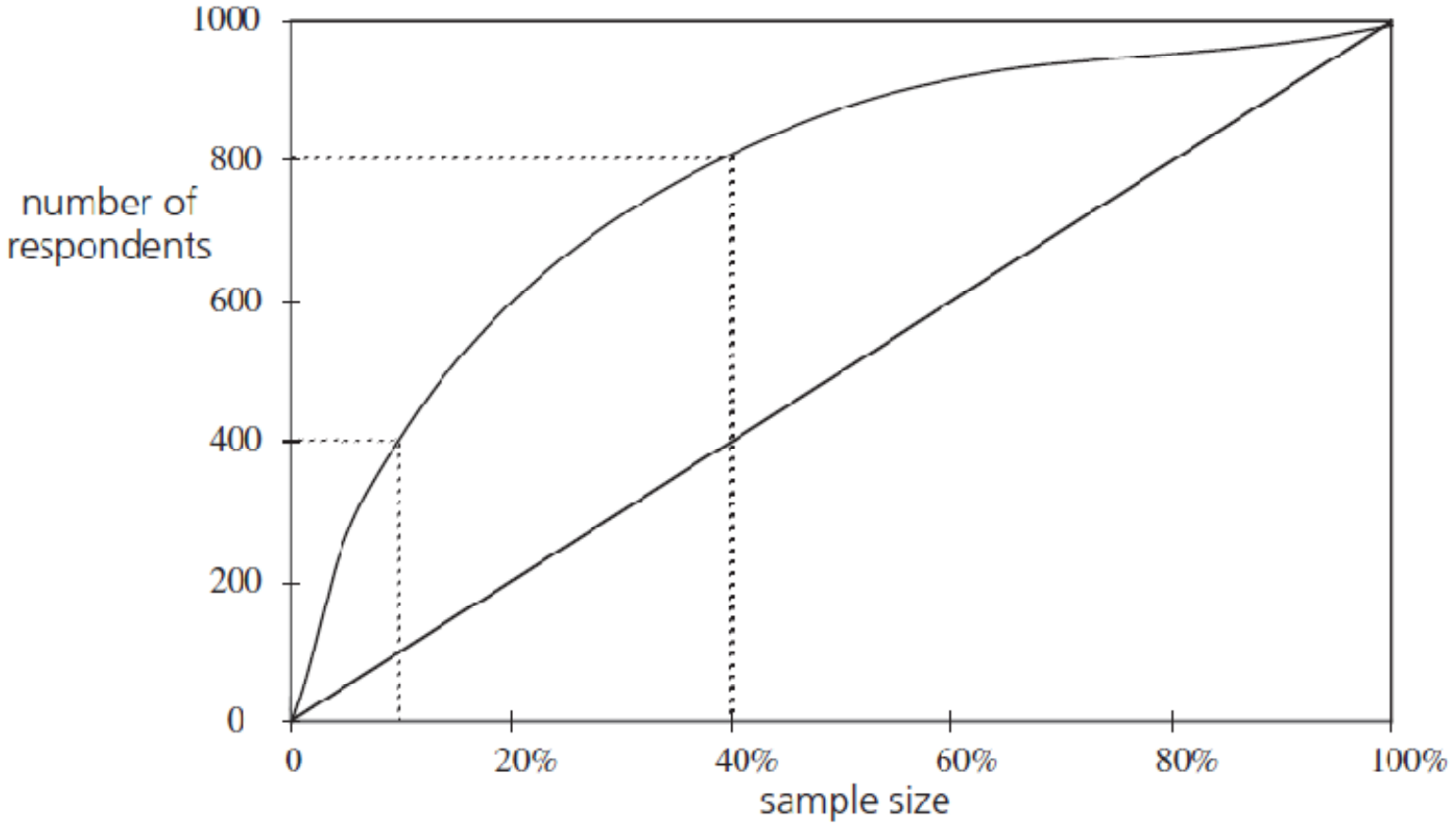
Költség-érzékeny tanulás

- Szeretnénk a költség mátrixot a tanulási folyamatba bevonni
- Két-osztályos predikció esetén egyszerűen a példányok arányát változtatjuk a tanuló halmazban

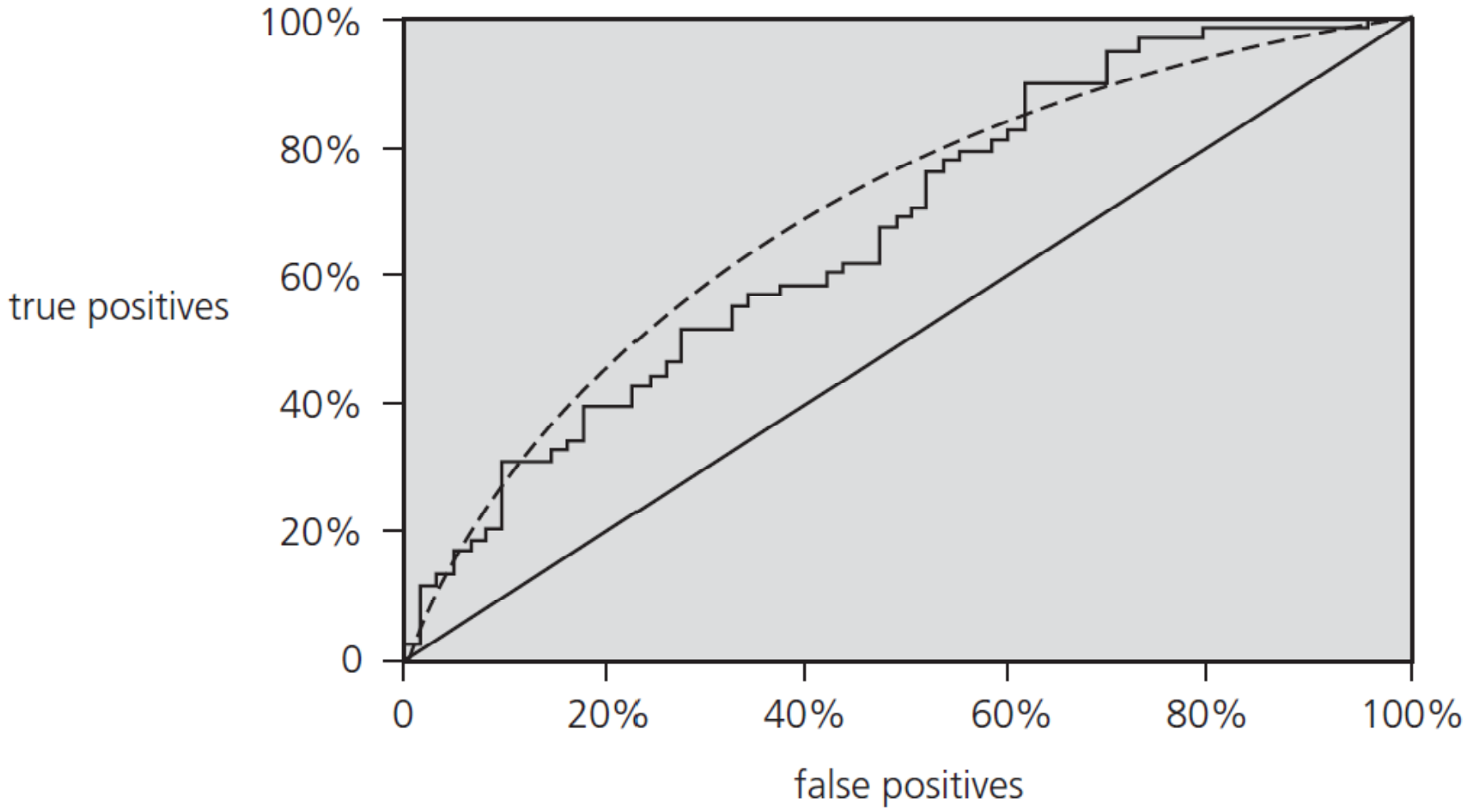
A lift diagram (1)

Rank	Predicted probability	Actual class	Rank	Predicted probability	Actual class
1	0.95	*yes*	11	0.77	*no*
2	0.93	*yes*	12	0.76	*yes*
3	0.93	*no*	13	0.73	*yes*
4	0.88	*yes*	14	0.65	*no*
5	0.86	*yes*	15	0.63	*yes*
6	0.85	*yes*	16	0.58	*no*
7	0.82	*yes*	17	0.56	*yes*
8	0.80	*yes*	18	0.49	*no*
9	0.80	*no*	19	0.48	*yes*
10	0.79	*yes*	. . .	. . .	. . .

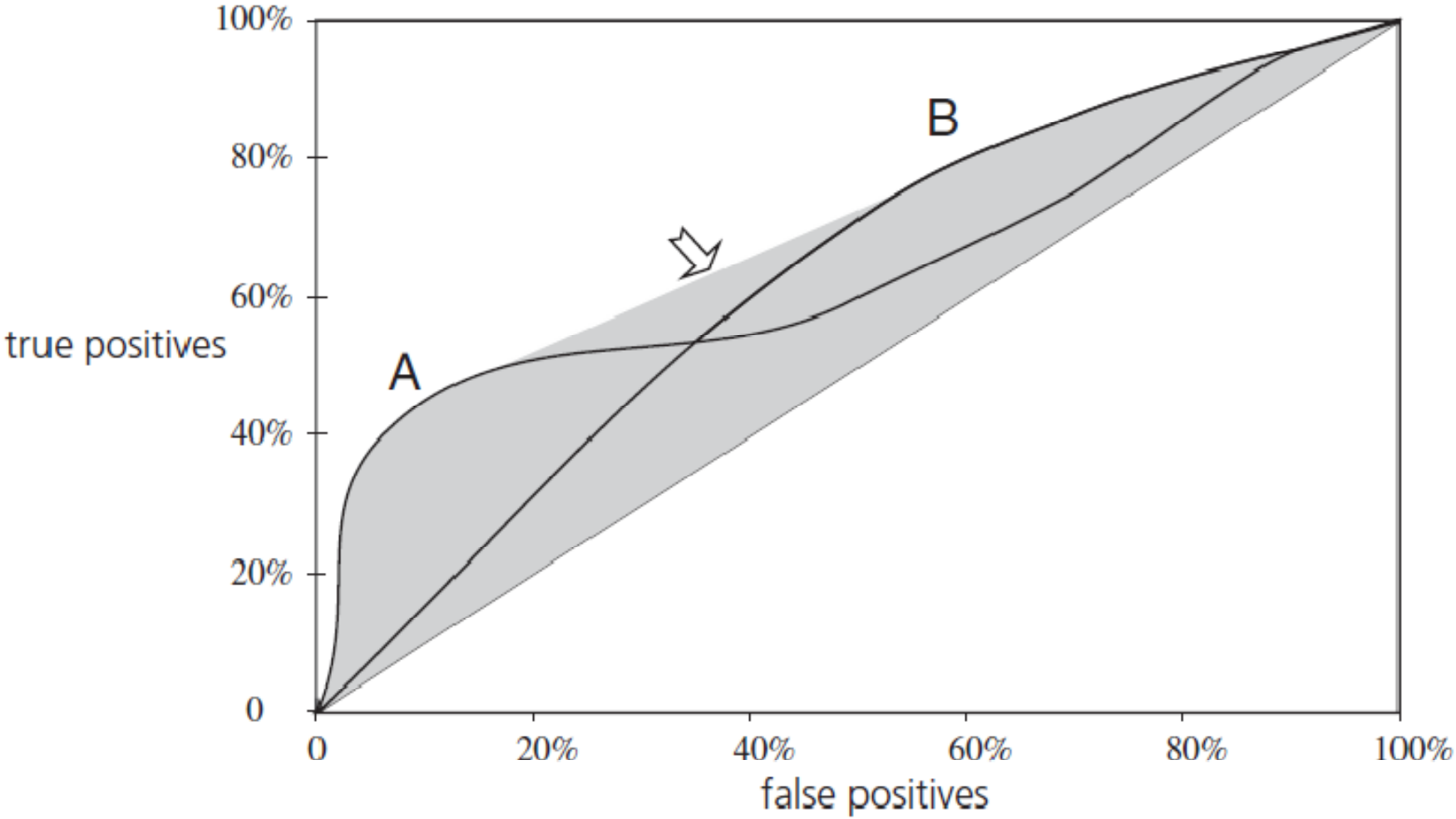
A lift diagram (2)



A ROC görbe


100%
80%
60%
40%
20%
0
true positives
0
20%
40%
60%
80%
100%
false positives

ROC görbe két tanuló módszer esetén



Recall-precision görbék

$$\text{Recall} = \frac{\text{az eredményül kapott releváns dokumentumok száma}}{\text{az összes releváns dokumentum száma}}$$

$$\text{Precision} = \frac{\text{az eredményül kapott releváns dokumentumok száma}}{\text{az eredményül kapott összes dokumentum}}$$

- 3-point average recall
- 11-point average recall
- F-measure:

$$\frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

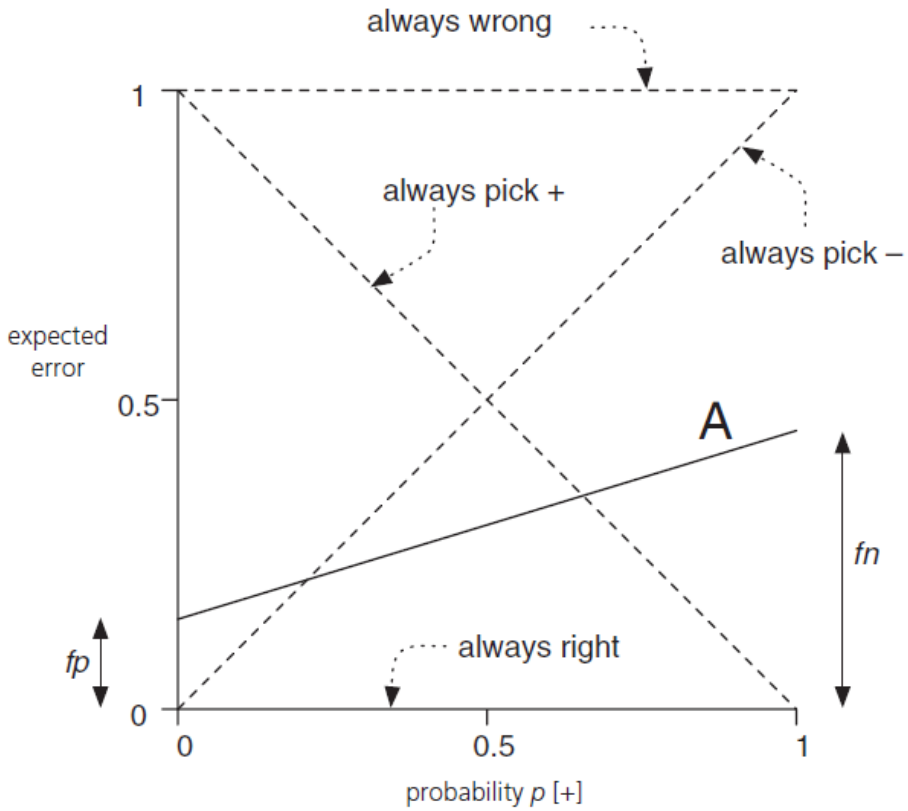
A különböző mérési módszerek összehasonlítása

	Domain	Plot	Axes	Explanation of axes
lift chart	marketing	TP vs. subset size	TP	number of true positives
			subset size	$\frac{TP+FP}{TP+FP+TN+FN} \times 100\%$
ROC curve	communications	TP rate vs. FP rate	TP rate	$tp = \frac{TP}{TP+FN} \times 100\%$
			FP rate	$fp = \frac{FP}{FP+TN} \times 100\%$
recall–precision curve	information retrieval	recall vs. precision	recall	same as TP rate *tp*
			precision	$\frac{TP}{TP+FP} \times 100\%$

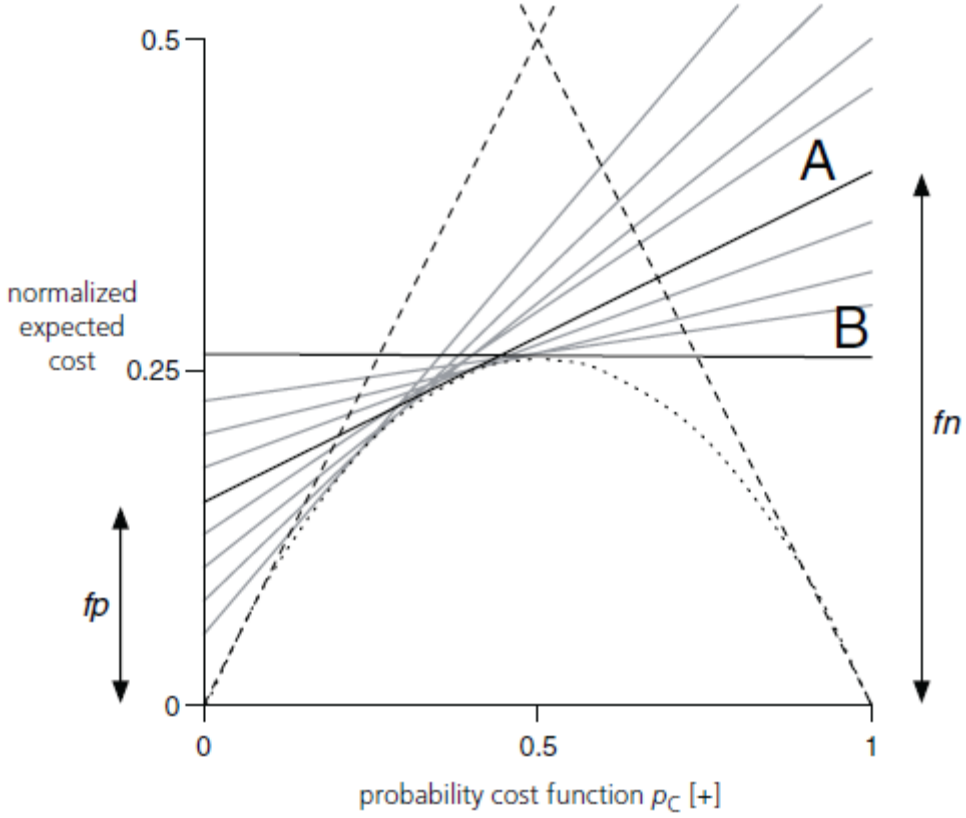
- Szenzitivitás, specificitás, szorzatuk:

$$\text{sensitivity} \cdot \text{specificity} = tp(1-fp) = \frac{TP \cdot TN}{(TP+FN) \cdot (FP+TN)}$$

Költség görbék



A hiba görbe



A költség görbe

- Normalizált várható költség $= fn \cdot p_c[+] + fp \cdot (1 - p_c[+])$
- Valószínűségi költségfüggvény $p_c[+] = \frac{p[+]C[+|-]}{p[+]C[+|-]+p[-]C[-|+]}$

Numerikus predikció kiértékelése

Performance measure	Formula
mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
relative squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i (p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_p=\frac{\sum_i (p_i-\bar{p})^2}{n-1}$, and $S_A=\frac{\sum_i (a_i-\bar{a})^2}{n-1}$

* p are predicted values and a are actual values.

Felhasznált irodalom

- Ian H. Witten & Eibe Frank: Data Mining – Practical Machine Learning Tools and Techniques (Morgan Kaufmann, 2005)