ANALÝZA A KLASIFIKACE DAT

RNDr. Eva Janoušová

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

HODNOCENÍ ÚSPĚŠNOSTI KLASIFIKACE A SROVNÁNÍ KLASIFIKÁTORŮ

ÚVOD

Vstupní data

Subjekt	Objem hipokampu	Objem komor	Skutečnost (správná třída)
1	2	12	pacient
2	4	10	pacient
3	3	8	pacient
4	5	7	kontrola
5	3	9	kontrola
6	4	5	kontrola

Výsledek klasifikace

pacient
pacient
kontrola
kontrola
pacient
kontrola

Jak dobrá je klasifikační metoda, kterou jsme použili?

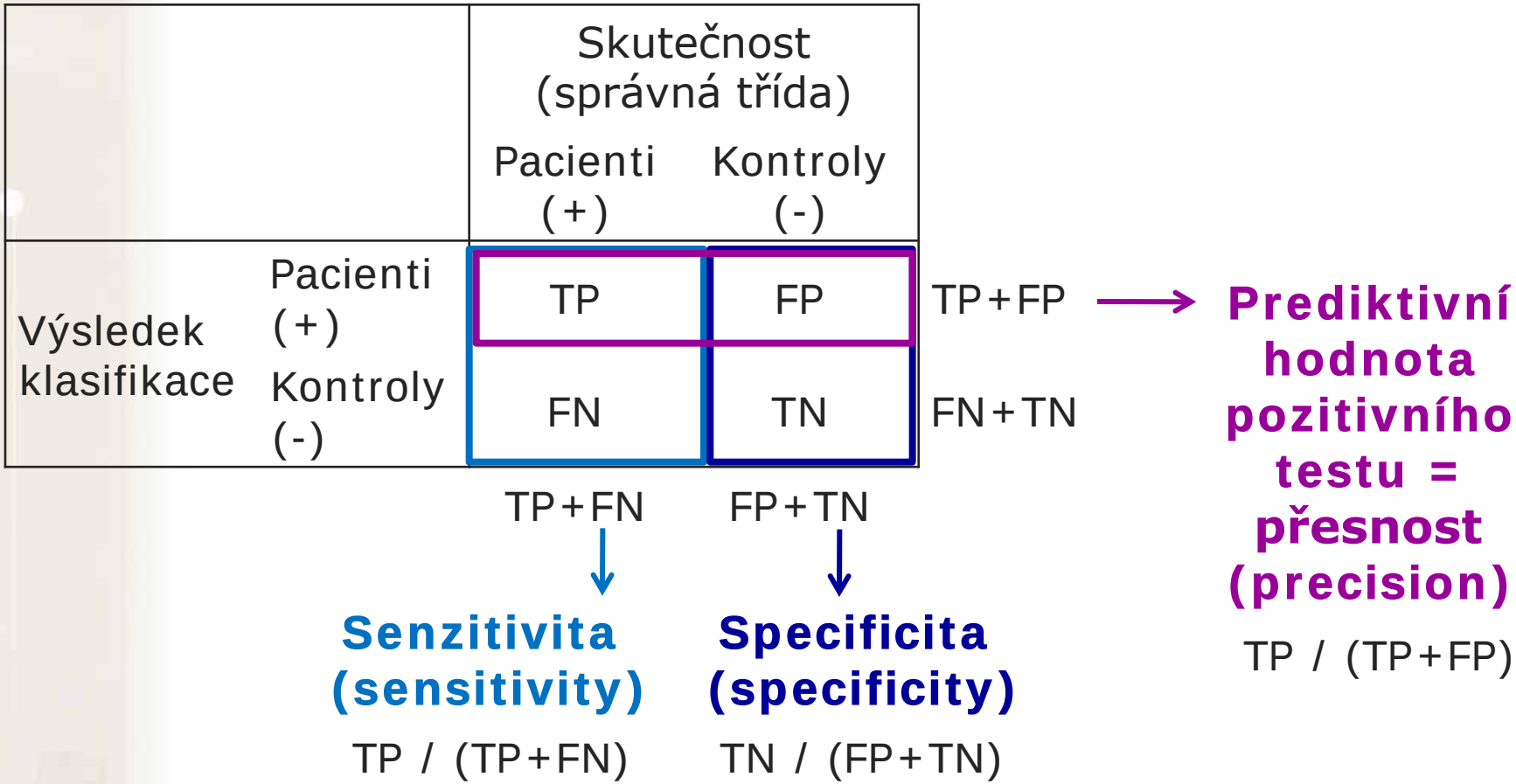
HODNOCENÍ ÚSPĚŠNOSTI KLASIFIKACE

Matice záměn (konfusní matice, confusion matrix):

		Skutečnost (správná třída)	
		Pacienti (+)	Kontroly (-)
Výsledek klasifikace	Pacienti (+)	TP	FP
	Kontroly (-)	FN	TN

- **TP** („true positive“) – kolik výsledků bylo skutečně pozitivních (tzn. kolik pacientů bylo správně diagnostikováno jako pacienti).
- **FP** („false positive“) – kolik výsledků bylo falešně pozitivních (tzn. kolik zdravých lidí bylo chybně diagnostikováno jako pacienti).
- **FN** („false negative“) – kolik výsledků bylo falešně negativních (tzn. kolik pacientů bylo chybně diagnostikováno jako zdraví).
- **TN** („true negative“) – kolik výsledků bylo skutečně negativních (tzn. kolik zdravých lidí bylo správně diagnostikováno jako zdraví).

HODNOCENÍ ÚSPĚŠNOSTI KLASIFIKACE



Celková správnost (accuracy)
(TP+TN)/(TP+FP+FN+TN)

Chyba (error)
(FP+FN)/(TP+FP+FN+TN)

PŘÍKLAD – KLASIFIKACE POMOCÍ LDA

Subjekt	Skuteč-nost	Výsledek LDA
1	P	P
2	P	P
3	P	K
4	K	K
5	K	P
6	K	K

Výsledek klasifikace	Skutečnost (správná třída)	
	Pacienti (+)	Kontroly (-)
Pacienti (+)	TP=2	FP=1
Kontroly (-)	FN=1	TN=2

Senzitivita: $TP/(TP+FN)=2/(2+1)=0.67$

Specificita: $TN/(FP+TN)=2/(1+2)=0.67$

Přesnost: $TP/(TP+FP)=2/(2+1)=0.67$

Správnost: $(TP+TN)/(TP+FP+FN+TN)=(2+2)/(2+1+1+2)=0.67$

Chyba: $(FP+FN)/(TP+FP+FN+TN)=(1+1)/(2+1+1+2)=0.33$

INTERVALY SPOLEHLIVOSTI PRO CHYBU

- chyba: $\frac{FP+FN}{TP+FP+FN+TN} = \frac{N_{error}}{N}$
- ze statistického hlediska – odhad pravděpodobnosti chyby: $\hat{P}_E = \frac{N_{error}}{N}$ (vychází z: $N_{error} \sim Bi(N, P_E)$)
- za splnění předpokladů, že $\hat{P}_E \cdot N > 5$, $\left(1 - \hat{P}_E\right) \cdot N > 5$ a $N > 30$, lze spočítat 95% interval spolehlivosti pro chybu pomocí aproximace na normální rozdělení:

$$\left\langle \hat{P}_E - 1.96 \times \sqrt{\frac{\hat{P}_E\left(1 - \hat{P}_E\right)}{N}}, \hat{P}_E + 1.96 \times \sqrt{\frac{\hat{P}_E\left(1 - \hat{P}_E\right)}{N}} \right\rangle$$

INTERVALY SPOLEHLIVOSTI PRO CELKOVOU SPRÁVNOST

- celková správnost: $\frac{TP+TN}{TP+FP+FN+TN} = 1 - \frac{N_{error}}{N}$
- z toho plyne: $\hat{P}_A = 1 - \hat{P}_E = \frac{N_{cor}}{N}$ (tedy $N_{cor} \sim Bi(N, P_A)$)
- za splnění předpokladů, že $\hat{P}_A \cdot N > 5$, $\left(1 - \hat{P}_A\right) \cdot N > 5$ a $N > 30$, lze spočítat 95% interval spolehlivosti pro správnost pomocí aproximace na normální rozdělení:

$$\left\langle \hat{P}_A - 1.96 \times \sqrt{\frac{\hat{P}_A\left(1 - \hat{P}_A\right)}{N}}, \hat{P}_A + 1.96 \times \sqrt{\frac{\hat{P}_A\left(1 - \hat{P}_A\right)}{N}} \right\rangle$$

PŘÍKLAD – POKRAČOVÁNÍ

Chyba: $\hat{P}_E = (FP+FN)/(TP+FP+FN+TN) = 0.33$

IS pro chybu:

$$\left[\hat{P}_E - 1.96\times\sqrt{\frac{\hat{P}_E\left(1-\hat{P}_E\right)}{N}}, \hat{P}_E + 1.96\times\sqrt{\frac{\hat{P}_E\left(1-\hat{P}_E\right)}{N}}\right]$$

$$\left[0.33 - 1.96\times\sqrt{\frac{0.33(1-0.33)}{6}}, 0.33 + 1.96\times\sqrt{\frac{0.33(1-0.33)}{6}}\right]$$

$$[0, 0.71]$$

Správnost: $\hat{P}_A = (TP+TN)/(TP+FP+FN+TN) = 0.67$

IS pro správnost:

$$\left[\hat{P}_A - 1.96\times\sqrt{\frac{\hat{P}_A\left(1-\hat{P}_A\right)}{N}}, \hat{P}_A + 1.96\times\sqrt{\frac{\hat{P}_A\left(1-\hat{P}_A\right)}{N}}\right]$$

$$\left[0.66 - 1.96\times\sqrt{\frac{0.66(1-0.66)}{6}}, 0.66 + 1.96\times\sqrt{\frac{0.66(1-0.66)}{6}}\right]$$

$$[0.29, 1.00]$$

TRÉNOVACÍ A TESTOVACÍ DATA

1. resubstituce
2. náhodný výběr s opakováním (bootstrap)
3. predikční testování externí validací (hold-out)
4. křížová validace (cross validation)
   - q k-násobná (k-fold)
   - q „odlož-jeden-mimo“ (leave-one-out, jackknife)

1. RESUBSTITUCE

- stejná trénovací a testovací množina

- **výhody:**
  - \+ jednoduchá
  - \+ rychlá

- **nevýhody:**
  - \- příliš optimistické výsledky

2. NÁHODNÝ VÝBĚR S OPAKOVÁNÍM (BOOTSTRAP)

- náhodně vybereme *N* subjektů s opakováním jako trénovací data (tzn. subjekty se v trénovací sadě mohou opakovat) a zbylé subjekty (ani jednou nevybrané) použijeme jako testovací data
- pro rozumně velká data se vybere zhruba 63,2% subjektů pro učení a 36,8% subjektů pro testování
- trénování a testování se provede jen jednou
- **výhody:**
  + velká trénovací sada
  + rychlé
- **nevýhody:**
  - data se v trénovací sadě opakují
  - výsledek vcelku závislý na výběru trénovacích dat

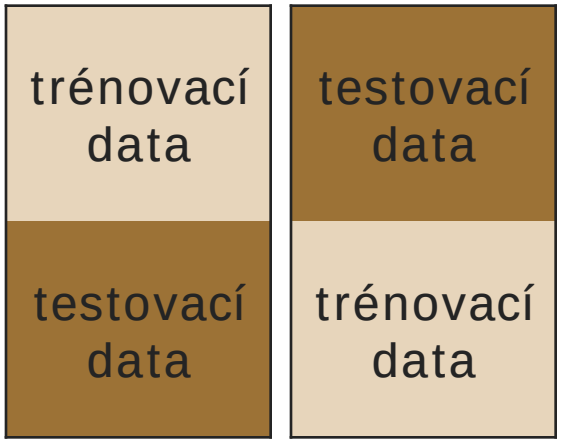
3. PREDIKČNÍ TESTOVÁNÍ EXTERNÍ VALIDACÍ (HOLD-OUT)

- použití části dat (většinou dvou třetin) na trénování a zbytku dat (třetiny) na testování

- **výhody:**
  + nezávislá trénovací a testovací sada
- **nevýhody:**
  - méně dat pro trénování i testování
  - výsledek velmi závislý na výběru trénovacích dat

3. PREDIKČNÍ TESTOVÁNÍ EXTERNÍ VALIDACÍ (HOLD-OUT) – MODIFIKACE 1

- použití části dat (obvykle poloviny) pro trénování a zbytku (poloviny) pro testování a následné přehození testovací a trénovací sady → zprů-měrování 2 výsledků klasifikace

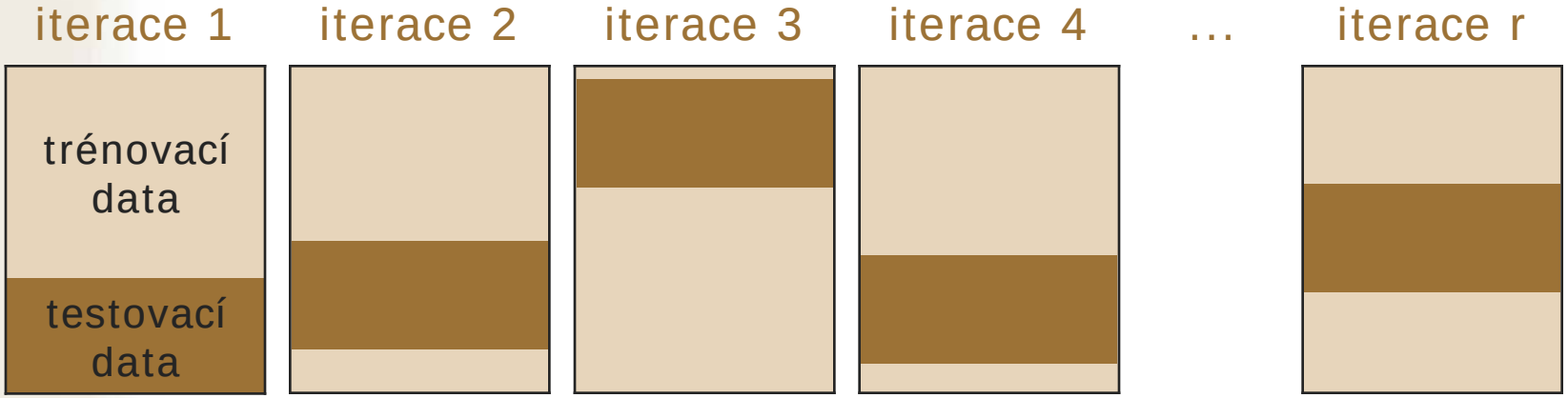


- **výhody:**
  + nezávislá trénovací a testovací sada

- **nevýhody:**
  - výsledek stále hodně závislý na výběru trénovacích dat (i když trochu méně než předtím)
  - při malých souborech může být polovina dat pro trénování příliš málo

3. PREDIKČNÍ TESTOVÁNÍ EXTERNÍ VALIDACÍ (HOLD-OUT) – MODIFIKACE 2

- *r*-krát náhodně rozdělíme soubor na trénovací a testovací data (většinou dvě třetiny pro trénování a třetinu pro testování) a *r* výsledků zprůměrujeme



- **výhody:**
  + poměrně přesný odhad úspěšnosti klasifikace
- **nevýhody:**
  - trénovací i testovací sady se překrývají
  - časově náročné

4. KŘÍŽOVÁ VALIDACE (CROSS VALIDATION, CV)

- používán též název příčná validace
- obecně: *k*-násobná (k-fold) křížová validace
- speciální případ: „odlož-jeden-mimo“ (leave-one-out = jackknife) křížová validace

4A. k-NÁSOBNÁ KŘÍŽOVÁ VALIDACE (k-FOLD CROSS VALIDATION)

- rozdělení souboru na *k* částí, 1 část použita na testování a zbylých *k*-1 částí na trénování → postup se opakuje (všechny části 1x použity pro testování)

např. pro *k*=5:

iterace 1	iterace 2	iterace 3	iterace 4	iterace 5
testování	trénování	trénování	trénování	trénování
trénování	testování	trénování	trénování	trénování
trénování	trénování	testování	trénování	trénování
trénování	trénování	trénování	testování	trénování
trénování	trénování	trénování	trénování	testování

- **výhody:**
  - \+ testovací sady se nepřekrývají
  - \+ poměrně přesný odhad úspěšnosti klasifikace
- **nevýhody:**
  - \- trénovací sady se překrývají
  - \- časově náročné

4B. „ODLOŽ-JEDEN-MIMO“ KŘÍŽOVÁ VALIDACE

- anglický překlad: leave-one-out (nebo jackknife)
- pro $k=N$ (tzn. v každé z N iterací je jeden subjekt použit na testování a zbylých N-1 subjektů na trénování)
- platí výhody a nevýhody zmíněné u k-násobné křížové validace se třemi komentáři:
  - časově nejnáročnější ze všech možných k
  - velmi vhodná pro malé soubory dat
  - v některých článcích se uvádí, že lehce nadhodnocuje úspěšnost → doporučuje se 10-násobná křížová validace

PŘÍKLAD – "ODLOŽ-JEDEN-MIMO" KŘÍŽOVÁ VALIDACE

Iterace:	iter. 1	iter. 2	iter. 3	iter. 4	iter. 5	iter. 6
	1	1	1	1	1	1
	2	2	2	2	2	2
	3	3	3	3	3	3
	4	4	4	4	4	4
	5	5	5	5	5	5
	6	6	6	6	6	6
Skutečnost:	pacient	pacient	pacient	kontrola	kontrola	kontrola
Výsledek klasifikace:	pacient	kontrola	kontrola	kontrola	pacient	kontrola

Výsledek klasifikace	Skutečnost pac.	Skutečnost kont.
pacient	TP=1	FP=1
kontrola	FN=2	TN=2

Senzitivita: $1/(1+2)=0.33$

Specificita: $2/(1+2)=0.67$

Přesnost: $1/(1+1)=0.50$

Správnost: $(1+2)/(1+1+2+2)=0.50$

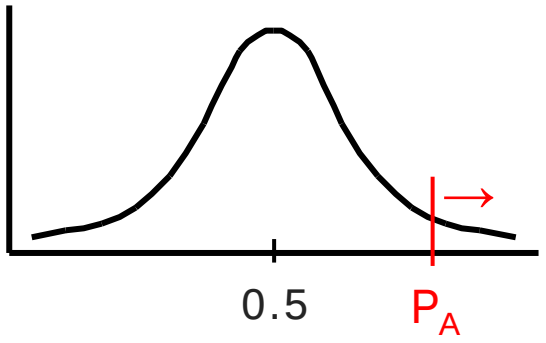
Chyba: $(1+2)/(1+1+2+2)=0.50$

JE KLASIFIKACE LEPŠÍ NEŽ NÁHODNÁ KLASIFIKACE?

- permutační testování
- jednovýběrový binomický test

PERMUTAČNÍ TESTOVÁNÍ

- r-krát náhodně přeházíme identifikátory příslušnosti do skupin u subjektů a provedeme klasifikaci (se stejným nastavením jako při použití originálních dat)
- p-hodnota se vypočte jako: $^{n}/_{r}$, kde *n* je počet iterací, v nichž byla úspěšnost klasifikace (např. celková správnost) vyšší nebo rovna úspěšnosti klasifikace originálních dat (P_A)
- pozn. pokud histogram z r celkových správností získaných permutacemi neleží kolem 0.5, máme v algoritmu zřejmě někde chybu!



JEDNOVÝBĚROVÝ BINOMICKÝ TEST

- testujeme, zda se liší celková správnost (což je podíl správně zařazených subjektů) od správnosti získané náhodnou klasifikací
- správnost u náhodné klasifikace: $P_{A_0} = {}^{N_i}/_{N}$, kde N_i je počet subjektů nejpočetnější skupiny
- $Z = \frac{P_A - P_{A_0}}{\sqrt{\left(P_{A_0}\left(1 - P_{A_0}\right)\right)/N}}$
- Pokud $|z| > 1.96$, zamítáme nulovou hypotézu o shodnosti správnosti naší klasifikace a správnosti náhodné klasifikace

PŘÍKLAD – JEDNOVÝBĚROVÝ BINOMICKÝ TEST

- uvažujme např. výsledek klasifikace pacientů a kontrol pomocí LDA (pomocí resubstituce):
  $P_A = 0.67,\ N = 6,\ P_{A_0} = {}^{N_i}/_{N} = 0.5$
- $z = \frac{P_A - P_{A_0}}{\sqrt{\left(P_{A_0}\left(1-P_{A_0}\right)\right)/N}} = \frac{0.67-0.5}{\sqrt{\left(0.5(1-0.5)\right)/6}} = 0.83$
- Protože $|z| < 1.96$, nezamítáme nulovou hypotézu o shodnosti správnosti naší klasifikace a správnosti náhodné klasifikace (tzn. neprokázali jsme, že by naše klasifikace byla lepší než náhodná klasifikace)

SROVNÁNÍ ÚSPĚŠNOSTI KLASIFIKACE

- Srovnání 2 klasifikátorů
- Srovnání 3 a více klasifikátorů

SROVNÁNÍ 2 KLASIFIKÁTORŮ

Klasifikátor 1	Klasifikátor 2	
	Správně (1)	Chybně (0)
Správně (1)	N_{11}	N_{10}
Chybně (0)	N_{01}	N_{00}

Celkem:

$$N_{11} + N_{10} + N_{01} + N_{00} = N_{ts}$$

McNemarův test: $x^2 = \frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}}$

Pokud $\chi^2 > 3.841$, zamítáme nulovou hypotézu H_0 o shodnosti celkové správnosti klasifikace pomocí dvou klasifikátorů

Dvouvýběrový binomický test:

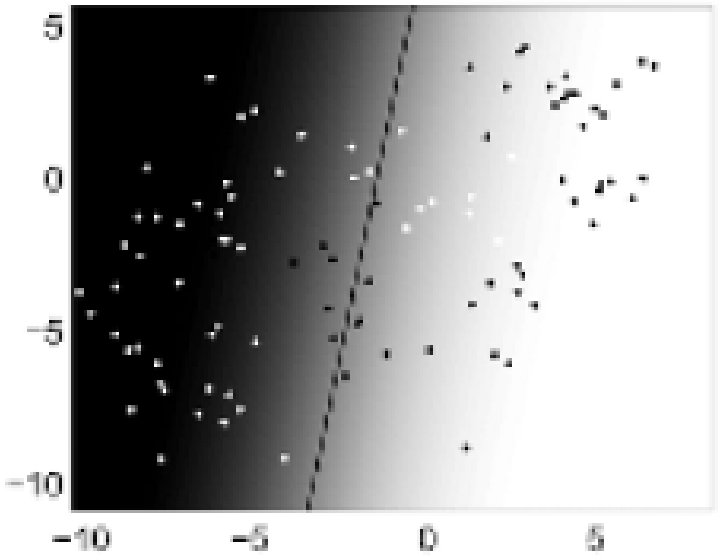
$$z = \frac{p_1 - p_2}{\sqrt{(2p(1-p))/(N_{ts})}} \qquad p_1 = \frac{N_{11} + N_{10}}{N_{ts}}; \qquad p_2 = \frac{N_{11} + N_{01}}{N_{ts}} \qquad p = \tfrac{1}{2}(p_1 + p_2)$$

Pokud $|z| > 1.96$, zamítáme nulovou hypotézu H_0 o shodnosti podílu správně klasifikovaných subjektů dvou klasifikátorů

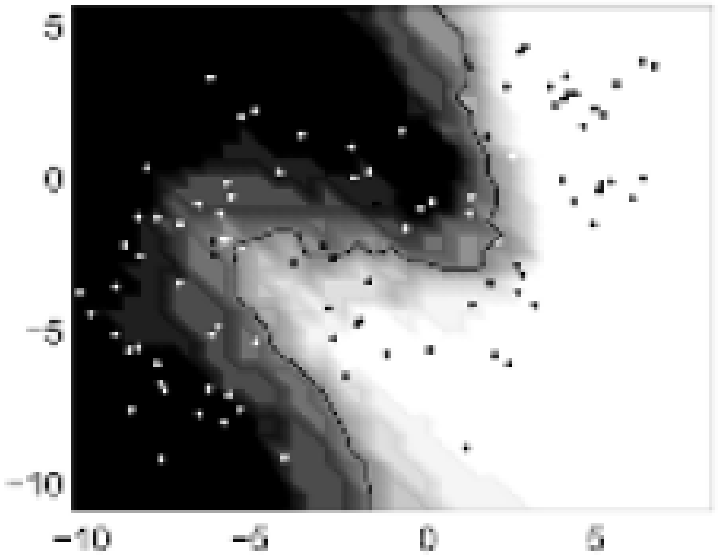
Dvouvýb. binomický test předpokládá nezávislost (tzn. že každý klasifikátor byl testován na jiném testovacím souboru) → raději používat McNemarův test

SROVNÁNÍ 2 KLASIFIKÁTORŮ - PŘÍKLAD

Lineární diskriminační analýza (LDA)



Metoda 9 nejbližších sousedů (9-nn)



SROVNÁNÍ 2 KLASIFIKÁTORŮ - PŘÍKLAD

Matice záměn:

LDA		9-nn	
42	8	44	6
8	42	2	48
84% správnost		92% správnost	

Shody u klasifikátorů:

Klasifikátor 1: LDA	Klasifikátor 2: 9-nn	
	Správně (1)	Chybně (0)
Správně (1)	$N_{11} = 82$	$N_{10} = 2$
Chybně (0)	$N_{01} = 10$	$N_{00} = 6$

McNemarův test:

$$x^2 = \frac{(|10-2|-1)^2}{10+2} = \frac{49}{12} \approx \mathbf{4.0833}$$

Protože $\chi^2 > 3.841$, zamítáme H_0.

Dvouvýb. binomický test:

$$z = \frac{0.84 - 0.92}{\sqrt{(2 \times 0.88 \times 0.12)/(100)}} \approx \mathbf{-1.7408}$$

Protože $|z| < 1.96$, nezamítáme H_0.

SROVNÁNÍ 3 A VÍCE KLASIFIKÁTORŮ

Testuje se, zda jsou statisticky významně odlišné správnosti klasifikátorů měřené na stejných testovacích datech – tzn. H_0: $p_1 = p_2 = \cdots = p_L$, kde p_L je správnost L-tého klasifikátoru. Poté je možno srovnávat správnosti klasifikátorů vždy po dvou, aby se zjistilo, které klasifikátory se od sebe liší.

Cochranův Q test:

$$Q_C = (L-1)\frac{L\sum_{i=1}^{L} G_i^2 - T^2}{LT - \sum_{j=1}^{N_{ts}} (L_j)^2}$$

Pokud $Q_C > \chi^2\,(L-1)$, zamítáme H_0.

F-test:

$$F_{\text{cal}} = \frac{MSA}{MSAB}$$

Pokud $F_{cal} > F(L-1, (L-1) \times (N_{ts}-1))$, zamítáme H_0.

Looney doporučuje F-test, protože je méně konzervativní.

S. W. Looney. A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, 8:5–9, 1988.

SROVNÁNÍ 3 A VÍCE KLASIFIKÁTORŮ - PŘÍKLAD

Matice záměn:

LDA		9-nn		Parzen	
42	8	44	6	47	3
8	42	2	48	5	45
84% správnost		92% správnost		92% správnost	

Cochranův Q test:

$$Q_C = 2 \times \frac{3 \times (84^2 + 92^2 + 92^2) - 268^2}{3 \times 268 - (80 \times 9 + 11 \times 4 + 6 \times 1)} \approx \mathbf{3.7647}$$

Protože $Q_C < \chi^2\,(L-1) = 5.991$, nezamítáme H_0.

F-test:

$$F_{\text{cal}} = \frac{0.2223}{0.0549} \approx \mathbf{4.0492}$$

Protože $F_{cal} > F(2{,}198) = 3.09$, zamítáme H_0.

SHRNUTÍ

- výpočet úspěšnosti klasifikace (správnosti, chyby, senzitivity, specificity a přesnosti) pomocí matice záměn
- výpočet intervalu spolehlivosti pro správnost a chybu
- volba trénovacího a testovacího souboru:
  - resubstituce
  - náhodný výběr s opakováním (bootstrap)
  - predikční testování externí validací (hold-out)
  - křížová validace (cross validation): k-násobná, „odlož-jeden-mimo“
- srovnání úspěšnosti klasifikace s náhodnou klasifikací
  - permutační testování
  - jednovýběrový binomický test
- srovnání úspěšnosti klasifikace 2 klasifikátorů:
  - McNemarův test
  - dvouvýběrový binomický test
- srovnání úspěšnosti klasifikace 3 a více klasifikátorů:
  - Cochranův Q test
  - F-test

Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ