Analýza a klasifikace dat – přednáška 6

RNDr. Eva Janoušová

Podzim 2015

Sekvenční klasifikace

Typy klasifikátorů

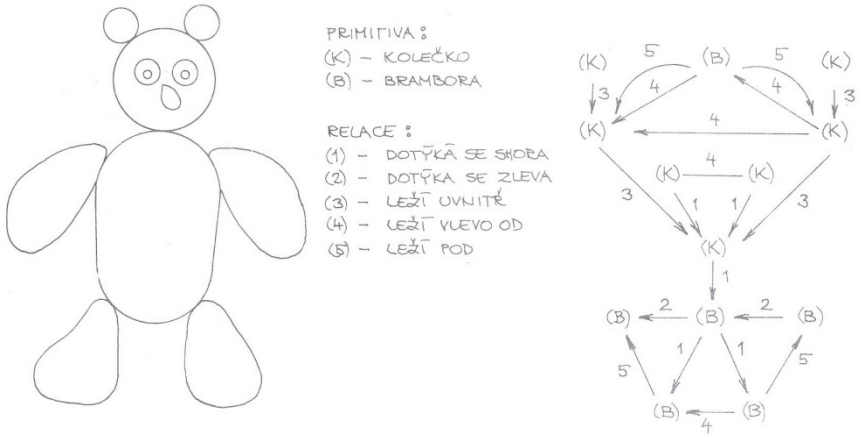
1. **Podle reprezentace vstupních dat:**
   - **příznakové klasifikátory:** paralelní x sekvenční
   - **strukturální (syntaktické) klasifikátory**
   - **kombinované klasifikátory**
2. **Podle jednoznačnosti zařazení do skupin:**
   - **deterministické klasifikátory**
   - **pravděpodobnostní klasifikátory**
3. **Podle typů klasifikačních a učících algoritmů:**
   - **parametrické klasifikátory**
   - **neparametrické klasifikátory**
4. **Podle způsobu učení:**
   - **učení s učitelem:** dokonalým x nedokonalým
   - **učení bez učitele**
5. **Podle podle principu klasifikace:**
   - **klasifikace pomocí diskriminačních funkcí**
   - **klasifikace pomocí vzdálenosti od etalonů klasifikačních tříd**
   - **klasifikace pomocí hranic v obrazovém prostoru**

Typy klasifikátorů – podle reprezentace vstupních dat

- **příznakové** – vstupní data vyjádřena vektorem hodnot jednotlivých proměnných (příznaků):
  - paralelní – zpracování vektoru jako celku (např. Bayesův klasifikátor)
  - sekvenční – zpracování (občas i měření) proměnných postupně (např. klasifikační stromy)

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

- **strukturální (syntaktické)** – vstupní data popsána relačními strukturami



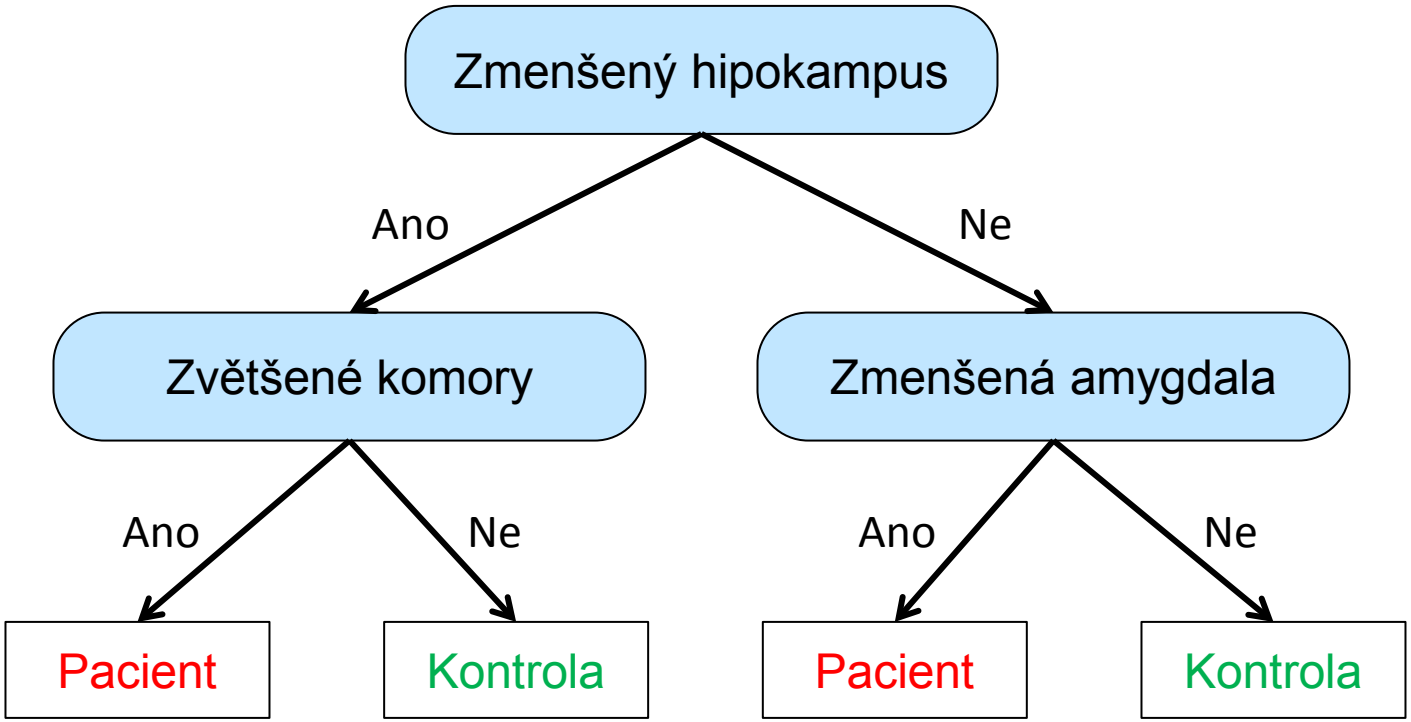
- **kombinované** – jednotlivá primitiva doplněna příznakovým popisem

Sekvenční klasifikace - motivace

- až dosud (bayesovské klasifikátory, klasifikátory s diskriminační hranicí, s minimální vzdáleností, …) – pevný konstantní počet příznaků
- kolik a jaké proměnné?
  - málo proměnných – možná chyba klasifikace
  - moc proměnných – možná nepřiměřená pracnost, vysoké náklady
  - použít proměnné, které nesou co nejvíce informace o klasifikační úloze
- **sekvenční klasifikace** – kompromis mezi velikostí klasifikační chyby a cenou určení příznaků
  - klasifikace na základě klasifikačního stromu
  - klasifikace s rostoucím počtem proměnných, přičemž okamžik ukončení klasifikační procedury stanoví klasifikátor sám podle předem daného kritéria pro kvalitu rozhodnutí (tj. na základě vlastností klasifikačních tříd, resp. objektů v nich)

Klasifikační (rozhodovací) stromy a lesy

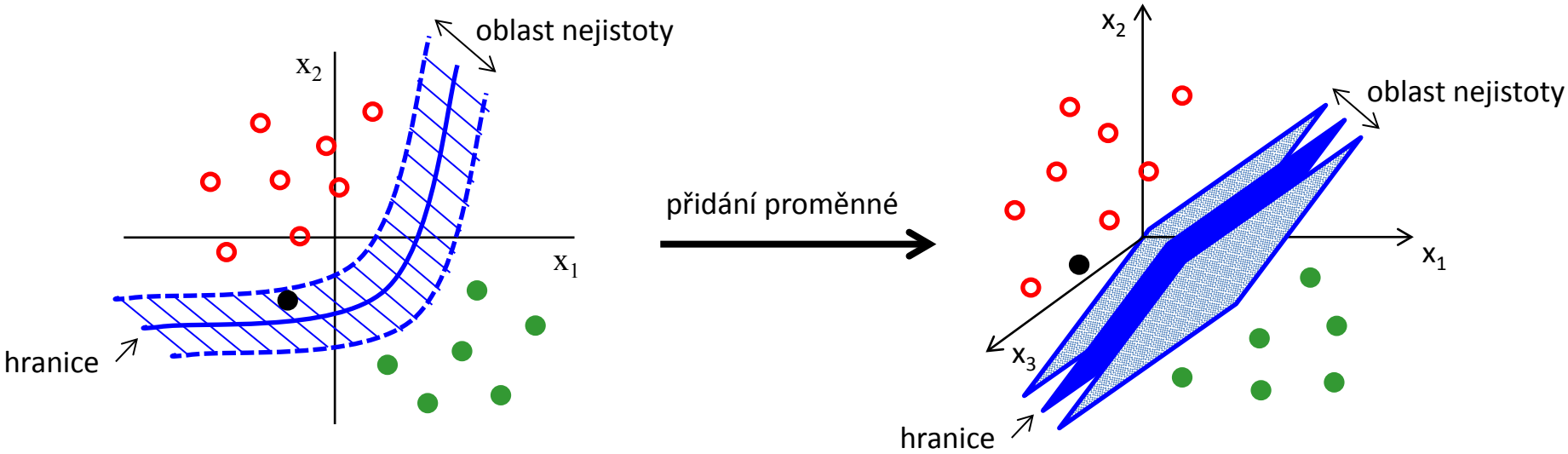
Princip: Postupné rozdělování datasetu do skupin podle hodnot jednotlivých proměnných.



Klasifikační lesy – použití více klasifikačních stromů ke klasifikaci.

Klasifikace s rostoucím počtem proměnných

Princip: Seřadíme proměnné podle množství informace, které nesou, a pak opakovaně provádíme klasifikaci objektu (subjektu) s postupně se zvyšujícím počtem proměnných, dokud objekt nejsme schopni jednoznačně zařadit



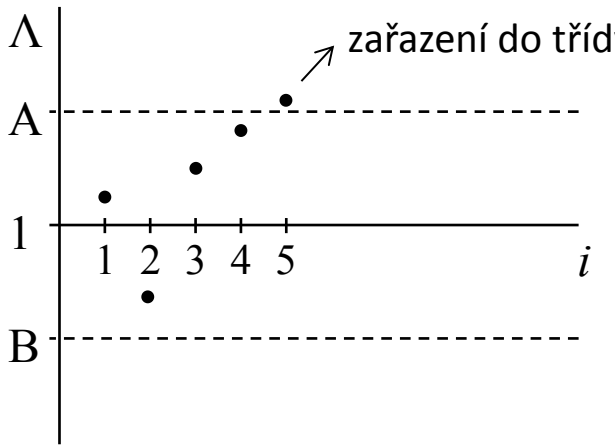
Kritéria pro řízení sekvenčního klasifikátoru:

- Waldovo kritérium
- Reedovo kritérium
- Modifikované Waldovo kritérium
- Modifikované Reedovo kritérium

Waldovo kritérium

- objekt **x** popsán množinou hodnot proměnných $\{x_1, x_2, \ldots\}$
- mějme $p(x_1, x_2, \ldots, x_i|\omega_1)$ a $p(x_1, x_2, \ldots, x_i|\omega_2)$, což jsou *i*-rozměrné hustoty pravděpodobnosti (tzn. dané prvními *i* proměnnými) výskytu objektu $\mathbf{x} = (x_1, x_2, \ldots, x_i)$ v *i*-tém klasifikačním kroku v třídách ω_1 a ω_2
- A a B jsou konstanty ($0<B<1<A<\infty$)
- spočítáme věrohodnostní poměr: $\Lambda_i = \frac{p(x_1, x_2, \ldots, x_i \mid \omega_1)}{p(x_1, x_2, \ldots, x_i \mid \omega_2)}$

1. pokud je $\Lambda_i \leq B$, pak se objekt **x** zařadí do třídy ω_2 a proces se ukončí
2. pokud je $\Lambda_i \geq A$, pak se objekt **x** zařadí do třídy ω_1 a proces se ukončí
3. pokud je $\Lambda_i \in (B,A)$, přidáme další proměnnou (příznak) x_{i+1} a proces se opakuje

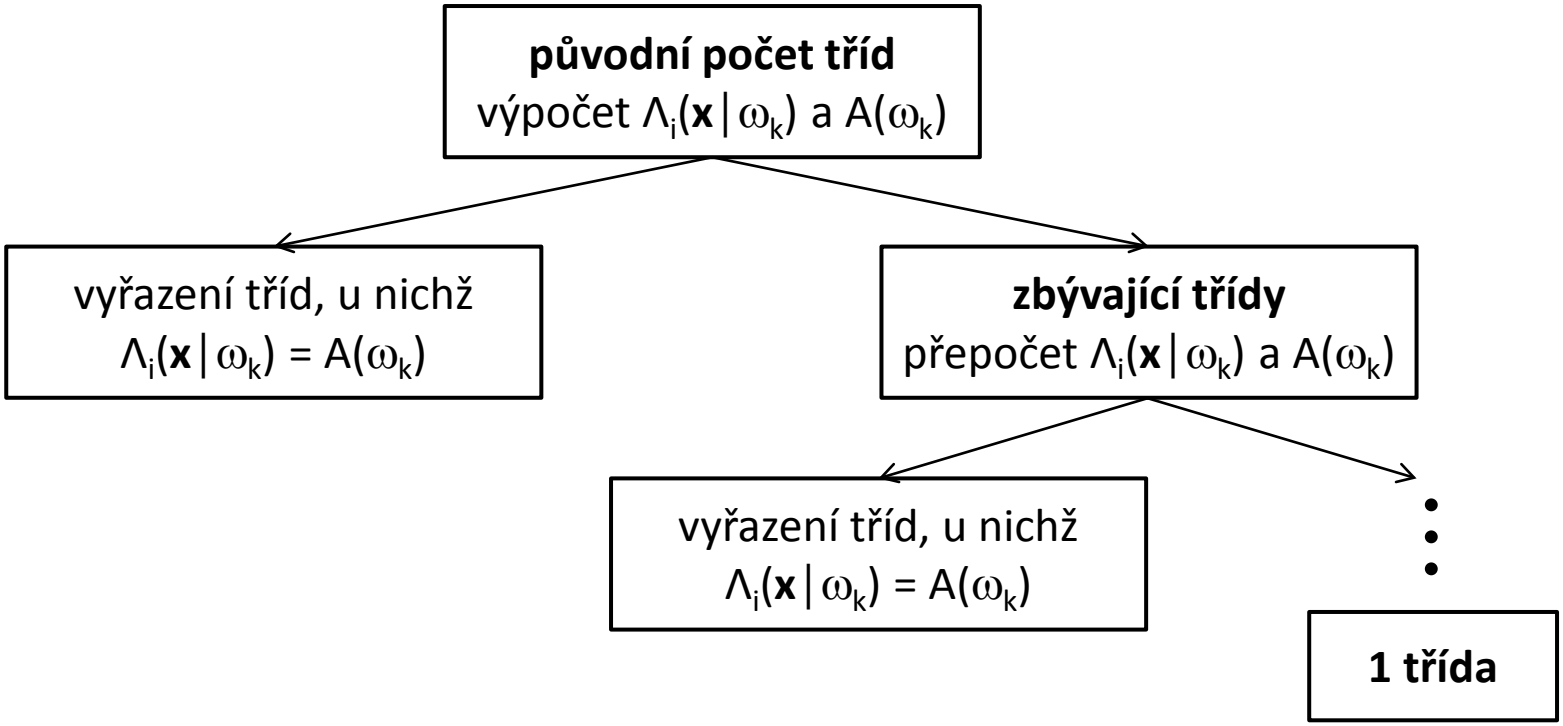


Optimální vlastnosti Waldova kritéria, protože:

- průměrný počet proměnných je menší nebo stejný jako u kritérií s pevným počtem proměnných
- průměrný počet kroků je menší než u jiných sekvenčních kritérií

Reedovo kritérium

- u více než 2 klasifikačních tříd
- založeno na výpočtu zobecněného věrohodnostního poměru pro *k*-tou třídu $\Lambda_i(\mathbf{x}\,|\,\omega_k)$ a mezní hodnoty *k*-té třídy $A(\omega_k)$
- postup:



- pokud není v některém kroku možné vyloučit žádnou třídu, zvýší se počet proměnných o 1 a proces pokračuje od začátku

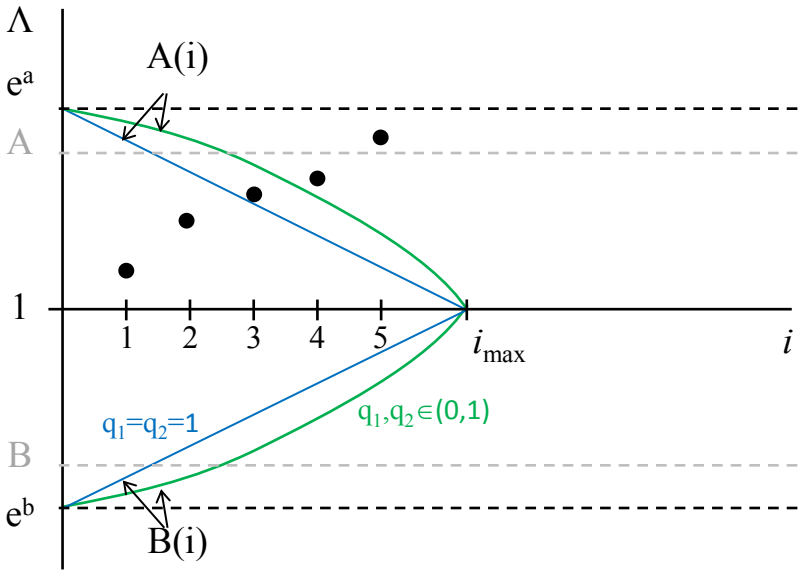
Modifikované Waldovo kritérium

- přes optimální vlastnosti Waldova kritéria může nastat:
  - počet kroků pro některé objekty velký, i když střední hodnota nízká
  - střední hodnota počtu kroků velká, pokud chceme malé pravděpodobnosti chybných rozhodnutí
- 2 možnosti řešení:

a) po určitém počtu kroků se sekvenční výpočet přeruší a dokončí se na základě nějakého rozhodnutí vycházejícího z nějakého kritéria založeného na pevném počtu příznaků

b) zavedení proměnných hranic A(i) a B(i)

$$např.\ A(i) = a.\left(1 - \frac{i}{i_{\max}}\right)^{q_1} \quad a\ B(i) = -b.\left(1 - \frac{i}{i_{\max}}\right)^{q_2}$$



Modifikované Reedovo kritérium

- zobecněný věrohodnostní poměr se srovnává s prahem $G_r(i) = g_r\left(1 - \frac{i}{i_{\max}}\right)^{q_r}$
- přičemž pokud $\Lambda_i(\mathbf{x} \mid \omega_r) < G_r(i)$, třída ω_r vyloučena z dalšího rozhodování
- jinak je postup stejný jako u klasického Reedova kritéria

Poznámka

- nelze dopředu říci, která klasifikační metoda bude pro daná data fungovat nejlépe → potřebné vyzkoušet více klasifikačních metod a zvolit nejvhodnější pro daná data
- u velkých datových souborů je obtížné dopředu určit, zda je možné data oddělit lineárně nebo ne → potřebné vyzkoušet lineární i nelineární klasifikační metody

Hodnocení úspěšnosti klasifikace a srovnání klasifikátorů

Hodnocení úspěšnosti klasifikace - úvod

Vstupní data

Subjekt	voxel 1	voxel 2	voxel 3	...	Skutečnost (správná třída)
1					pacient
2					pacient
3					pacient
4					kontrola
5					kontrola
6					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

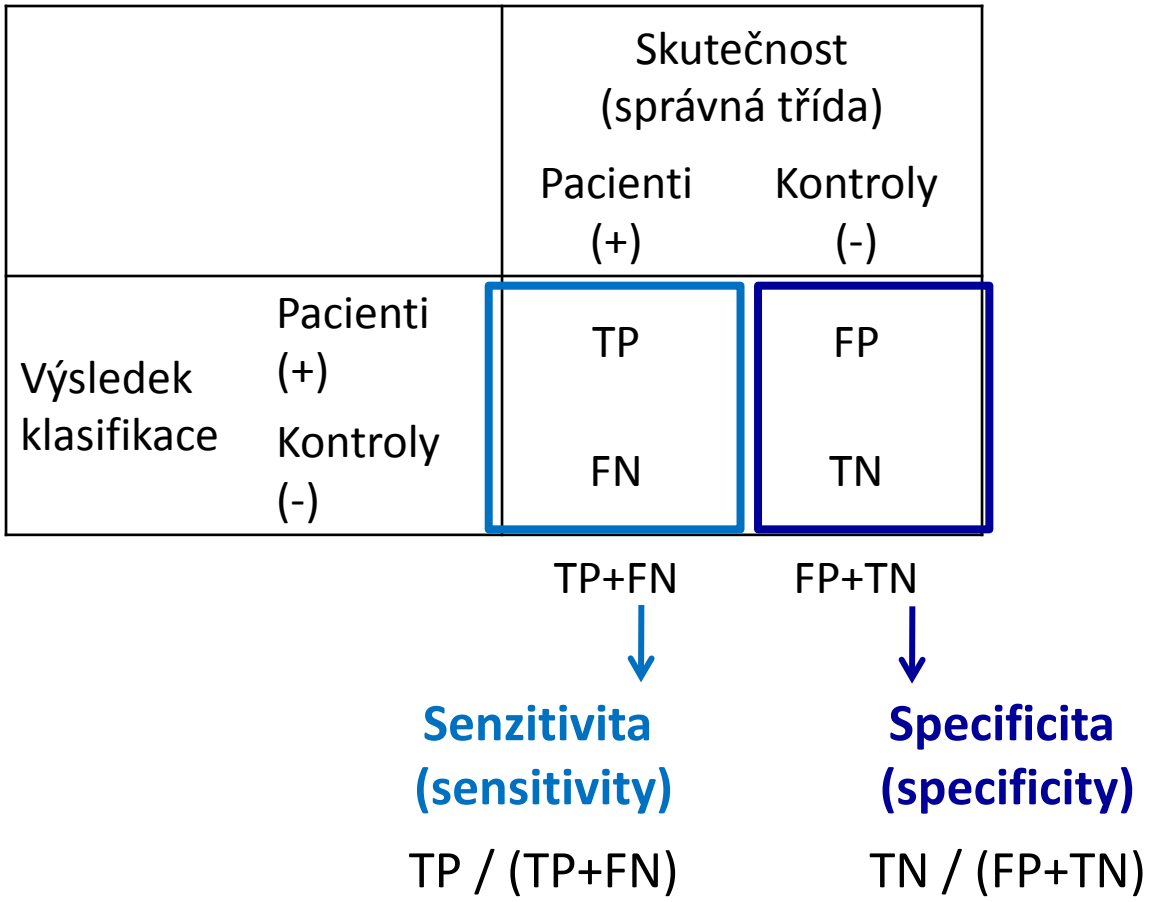
TP („true positive“) – kolik výsledků bylo skutečně pozitivních
(tzn. kolik pacientů bylo správně diagnostikováno jako pacienti).

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 FLDA

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

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 celkovou správnost

- celková správnost: $\frac{TP+TN}{TP+FP+FN+TN}$
- z toho plyne: $\hat{P}_A = \frac{N_{cor}}{N}$ (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:

$$\left[\hat{P}_A - 1{,}96 \cdot \sqrt{\frac{\hat{P}_A\left(1-\hat{P}_A\right)}{N}}; \hat{P}_A + 1{,}96 \cdot \sqrt{\frac{\hat{P}_A\left(1-\hat{P}_A\right)}{N}}\right]$$

Příklad – pokračování

Správnost: (TP+TN)/(TP+FP+FN+TN) = 0,67

IS pro správnost:

$$\left[\hat{P}_A - 1{,}96 \cdot \sqrt{\frac{\hat{P}_A\left(1-\hat{P}_A\right)}{N}}; \hat{P}_A + 1{,}96 \cdot \sqrt{\frac{\hat{P}_A\left(1-\hat{P}_A\right)}{N}}\right]$$

$$\left[0{,}66 - 1{,}96 \cdot \sqrt{\frac{0{,}66(1-0{,}66)}{6}}; 0{,}66 + 1{,}96 \cdot \sqrt{\frac{0{,}66(1-0{,}66)}{6}}\right]$$

$$[0{,}29; 1{,}00]$$

Trénovací a testovací data

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

1. resubstituce

- stejná trénovací a testovací množina
- **výhody:**
  + jednoduché
  + rychlé
- **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)

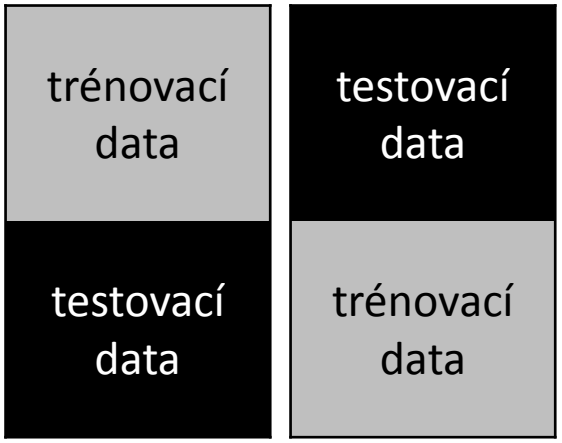
- použití části dat (většinou dvou třetin) na trénování a zbytku dat (třetiny) na testování

- **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

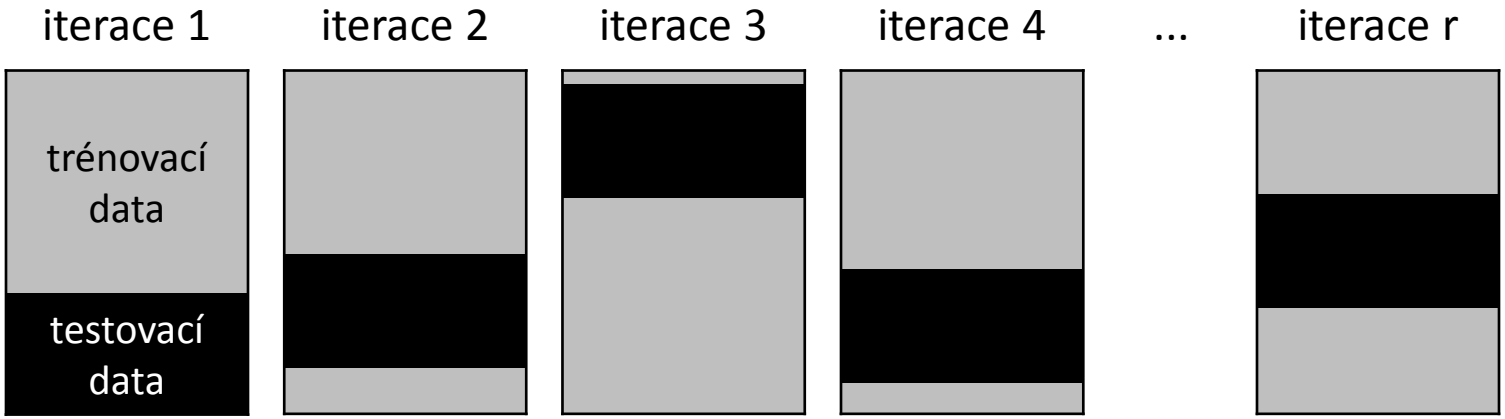


- **výhody:**
  + nezávislá trénovací a testovací sada

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

3. predikční testování externí validací (hold-out) – modifikace 2

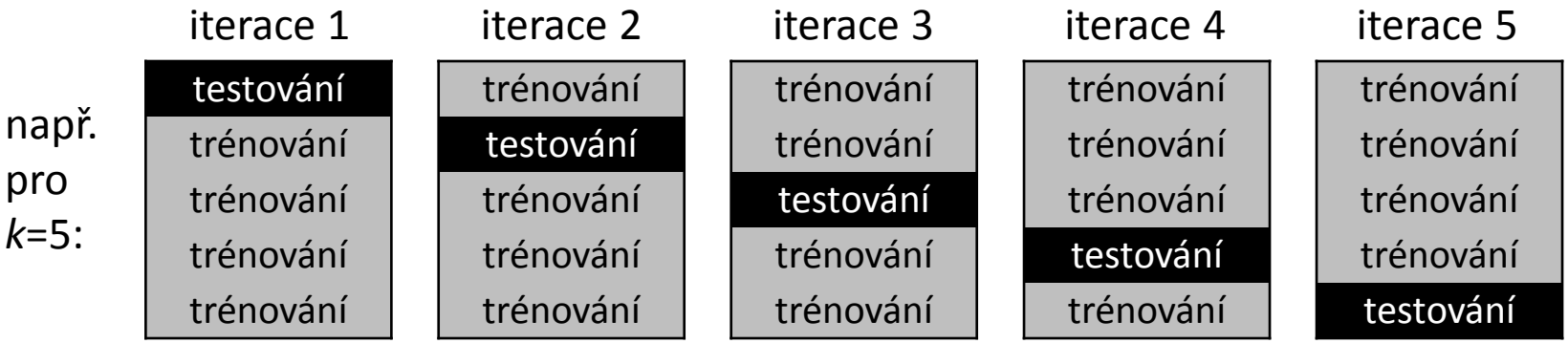
- *r*-krát náhodně rozdělíme soubor na trénovací a testovací data (většinou dvě třetiny pro trénování a třetinu pro testování) a *r* výsledků zprůměrujeme



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

4. *k*-násobná křížová validace (*k*-fold cross validation)

- používán též název příčná validace
- 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)
- speciálním případem je „odlož-jeden-mimo“ (leave-one-out) CV (pro *k*=N)



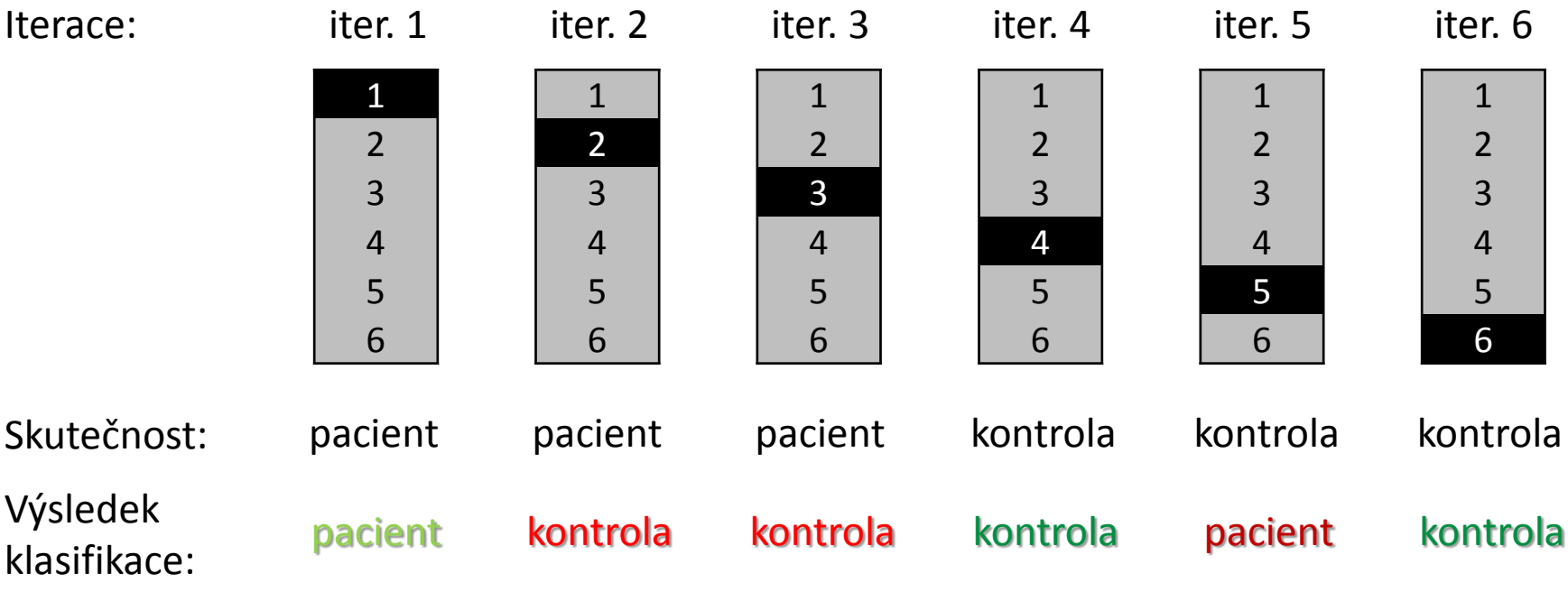
- **výhody:**
  + testovací sady se nepřekrývají
  + poměrně přesný odhad úspěšnosti klasifikace

- **nevýhody:**
  - časově náročné

„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 čtyřmi komentáři:
  - časově nejnáročnější ze všech možných k
  - velmi vhodná pro malé soubory dat
  - na rozdíl od jakékoliv k-fold CV dostaneme vždy pouze jeden výsledek úspěšnosti (tzn. výsledek úspěšnosti nezávisí na tom, jak se jednotlivé subjekty „namíchají“ do jednotlivých skupin)
  - v některých článcích se uvádí, že lehce nadhodnocuje úspěšnost → doporučuje se 10-násobná křížová validace

Příklad - „odlož-jeden-mimo“ křížová validace



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

Senzitivita: 1/(1+2)=0,33

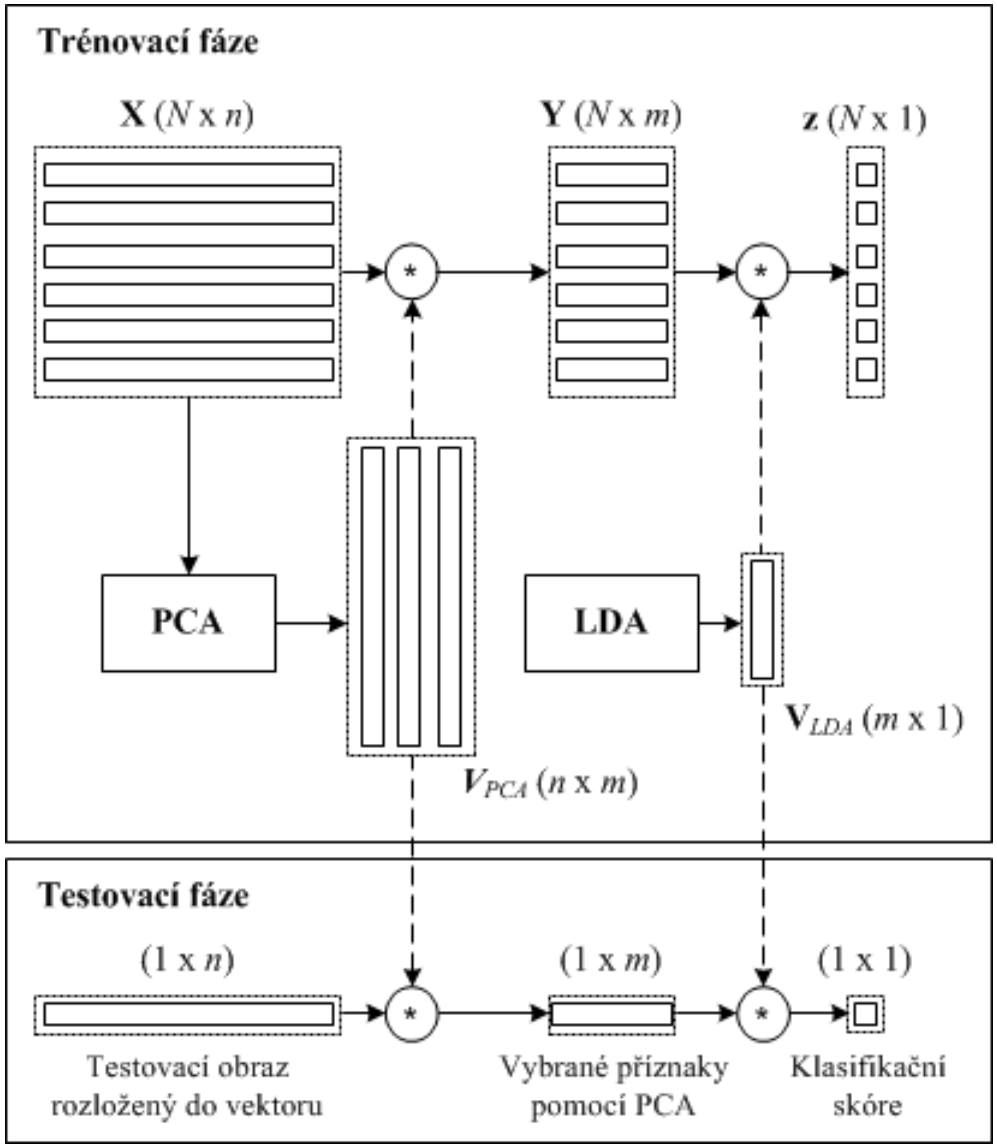
Specificita: 2/(1+2)=0,67

Správnost: (1+2)/(1+1+2+2)=0,50

Chyba: (1+2)/(1+1+2+2)=0,50

Upozornění !!!

Je potřebné rozdělit soubor na trénovací a testovací ještě před redukcí dat, jinak dostaneme nadhodnocené výsledky!!!

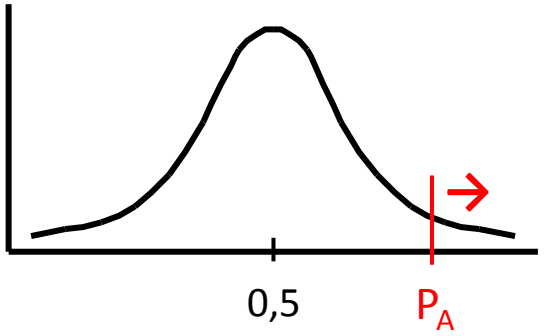


Je klasifikace lepší než náhodná klasifikace?

- 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 (v případě vyrovnaných skupin), 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
- $z = \frac{P_A - P_{A_0}}{\sqrt{(P_{A_0}(1-P_{A_0}))/N}}$
- 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): $P_A = 0{,}67$, $N = 6$, $P_{A_0} = {}^{N_i}/_{N} = 0{,}5$
- $z = \frac{P_A - P_{A_0}}{\sqrt{\left(P_{A_0}(1-P_{A_0})\right)/N}} = \frac{0{,}67-0{,}5}{\sqrt{\left(0{,}5(1-0{,}5)\right)/6}} = 0{,}83$
- 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)
- nezamítnutí nulové hypotézy vyplývá už i z vypočteného intervalu spolehlivosti (0,29 – 1,00), protože tento interval spolehlivosti obsahuje hodnotu 0,5

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:

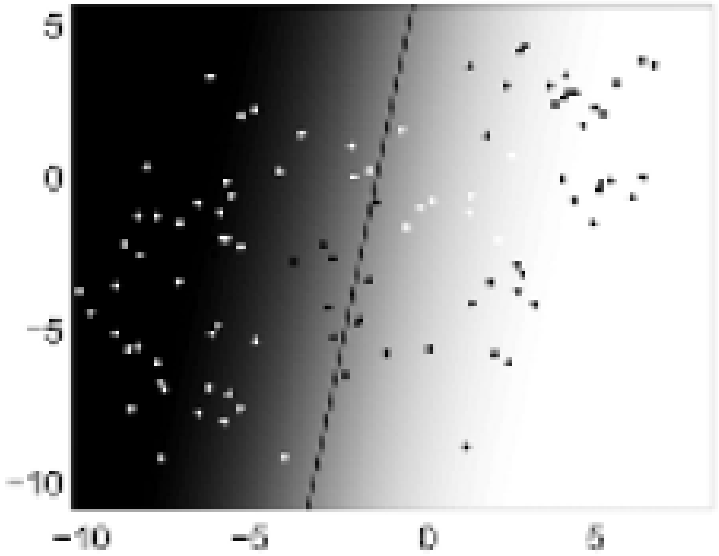
$$z = \frac{p_1 - p_2}{\sqrt{(2p(1-p))/(N_{ts})}} \qquad p_1 = \frac{N_{11} + N_{10}}{N_{ts}}; \qquad p_2 = \frac{N_{11} + N_{01}}{N_{ts}} \qquad p = \tfrac{1}{2}(p_1 + p_2)$$

Pokud $|z| > 1{,}96$, zamítáme nulovou hypotézu H_0 o shodnosti podílu správně klasifikovaných subjektů dvou klasifikátorů

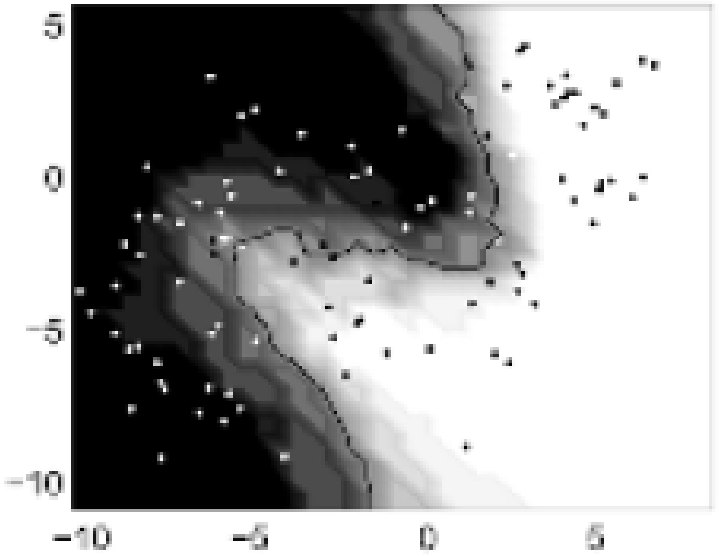
Dvouvýb. binomický test předpokládá nezávislost (tzn. že každý klasifikátor byl testován na jiném testovacím souboru) → raději používat McNemarův test

Příklad – srovnání 2 klasifikátorů

Lineární diskriminační analýza (LDA)



Metoda 9 nejbližších sousedů (9-nn)



Příklad – srovnání 2 klasifikátorů

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.

Příklad – srovnání 3 a více klasifikátorů

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 pro obor Matematická biologie

je podporována projektem OPVK

č. CZ.1.07/2.2.00/28.0043

„Interdisciplinární rozvoj studijního oboru Matematická biologie“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ