České vysoké učení technické v Praze

Fakulta elektrotechnická

Katedra kybernetiky
Katedra počítačů

Vytěžování dat – cvičení III

Bayesovská predikce, ztráta, riziko, jednoduchá klasifikace

Filip Železný: zelezny@fel.cvut.cz
Pavel Kordík: kordikp@fel.cvut.cz
Monika Žáková: zakovm1@fel.cvut.cz

Program cvičení

- Bayesovská predikce
  - Bayesovská predikce na základě jednoho příznaku
  - Riziko
  - Podmíněné riziko
  - Bayesovská predikce na základě více příznaků
- Jednoduchá klasifikace
  - Klasifikace podle jednoho vstupu
  - Hodnocení klasifikátoru

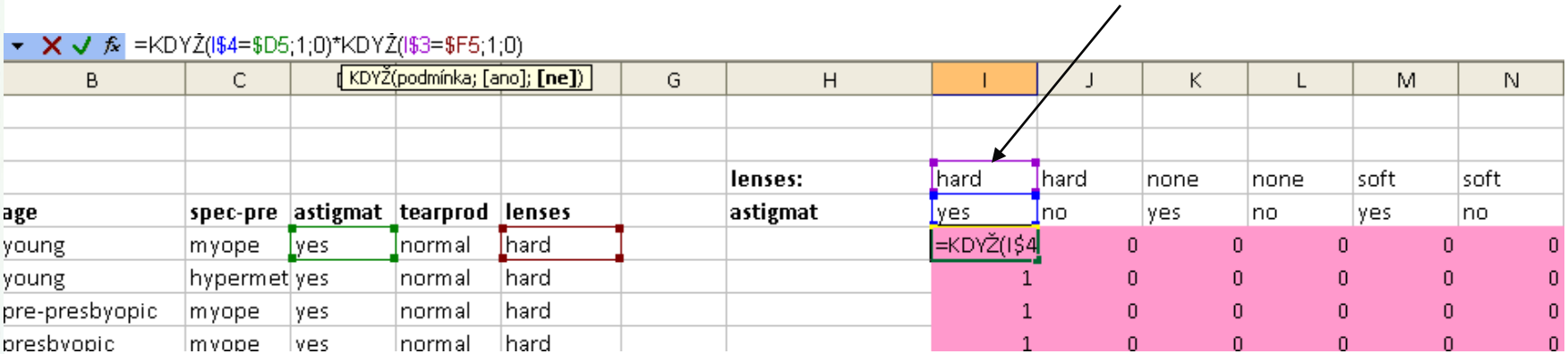
Predikce na základě jednoho příznaku

Predikce typu contact-lenses na základě hodnoty astigmatismu pomocí maximalizace podmíněné pravděpodobnosti

- Vyplňte nejprve pomocnou tabulku pro výpočet relativních četností a z ní odhadněte pravděpodobnosti.
- Z toho sestavte tabulku sdružených pravděpodobností a vypočtěte i marginální pravděpodobnosti.
- Diskutujte hodnoty apriorní pravděpodobnosti třídy contact-lenses a jejich význam.

Návod

Pro každý sloupec (jev) spočítáme četnost



B	C	D	E	F	G	H	I	J	K	L	M	N
						lenses:	hard	hard	none	none	soft	soft
age	**spec-pre**	**astigmat**	**tearprod**	**lenses**		**astigmat**	yes	no	yes	no	yes	no
young	myope	yes	normal	hard			=KDYŽ(I$4	0	0	0	0	0
young	hypermet	yes	normal	hard			1	0	0	0	0	0
pre-presbyopic	myope	yes	normal	hard			1	0	0	0	0	0
presbyopic	myope	yes	normal	hard			1	0	0	0	0	0

- Abychom v tabulkovém procesoru urychlili práci, používáme kopírování vzorců s fixací řádku (I$4), sloupce ($D5) nebo buňky(F5)
- Kopírování vzorce aktivní buňky – označit, levé tlačítko na levém spodním rohu buňky (kříž)

Predikce na základě jednoho příznaku

- Sestavte dále tabulku podmíněných pravděpodobností.
- Diskutujte její hodnoty a jejich význam ve srovnání s výše uvedenými apriorními pravděpodobnosti.
- Zopakujte předešlé výpočty na dalším pracovním listu, tentokrát pro atribut tearprod (tear production)
- Diskutujte rozdíl mezi podmíněnou a apriorní pravděpodobností třídy.

Riziko

- máte k dispozici sdružené a podmíněné pravděp. rozložení pro contact-lenses vs tearprod.
- Znáte vzorec pro výpočet středního rizika:

$$R(f) = \sum_{\mathbf{x}} \sum_{y} P(x, y)[L(f(\mathbf{x}), y)]$$

- definujte si ztrátovou funkci tak, že ztráta je 0, rovná-li se predikovaná třída skutečné, jinak 1
- vyzkoušejte, jak se střední riziko mění pro různé predikce
- je možné nalézt lepší predikci, než podle podmíněné pravděpodobnosti?

Podmíněné riziko

- Střední riziko lze počítat jako střední hodnotu podmíněných rizik pro jednotlivé hodnoty pozorovaného příznaku:

$$R(f) = \sum_{\mathbf{x}} P(\mathbf{x}) E[L(f(\mathbf{x}), y) | \mathbf{x}] \equiv E_{\mathbf{x}}[r(f, \mathbf{x})]$$

- je predikce podle jedné hodnoty podmínky závislá na ostatních?
- upravte ztrátovou funkci – např. pro predikci soft při skutečnosti hard, dejte ztrátu = 10
- jak se liší predikce podle středního rizika a predikce podle podmíněné pravděpodobnosti?

Predikce na základě více příznaků

- přizpůsobte výpočet pravděpodobnostního rozložení tak, abychom třídu contact-lenses predikovali na základě obou příznaků astigmat. a tearprod.
- kolik příkladů v datech bychom minimálně potřebovali, aby se každá kombinace hodnot atributů objevila alespoň jednou
- předpokládejte *n* atributů, každý s *m* možnými hodnotami.

Jednoduchá klasifikace

- Stejná data jako minulé cvičení
- Klasifikace do dvou tříd:
  - auta původem z Ameriky (origin 1)
  - auta z Evropy a Japonska (origin 2,3)

- Vyrobte graf, který rozmístí jednotlivé vozy podle atributů x=mpg a y=weight, barva vozu bude znázorňovat třídu (viz výše).

Klasifikace podle jednoho vstupu

- Vyrobte klasifikátor, který na základě hodnoty atributu mpg rozhodne, do které ze dvou tříd vůz patří.
- Spočítejte chybně klasifikované vozy a vypočtěte procentuální úspěšnost klasifikace.
- To samé udělejte pro klasifikátor, který o každém vozu prohlásí, že pochází z Ameriky (patří do třídy 1).

Hodnocení klasifikátoru I

- Vypočítejte četnosti v jednotlivých třídách.
- Spočítejte TP,TN,FP a FN

true = dobrá klasifikace

false = špatná klasifikace

positive = klasifikován jako americký

negative = klasifikován jako neamerický

		origin:	
klasifikace:		amerika	mimo
	amerika	TP	FP
	mimo	FN	TN

Hodnocení klasifikátoru II

- TP=americký označen jako americký
- TN=neamerický označen neamerickým
- FP=označí ho jako americký, ale ve skutečnosti je evropský nebo japonský
- FN=americký vůz není označen jako americký
- Vypočtěte FN rate, FP rate, specificitu a senzitivitu.

Hodnocení klasifikátoru II

- FP rate=procento neamerických aut klasifikovaných jako americká
- FN rate=procento špatně klasifikovných amerických vozů
- specificita=pravděpodobnost správné klasifikace neamerických vozů
- senzitivita=pravděpodobnost, že americké auto bude klasifikováno jako americké
- Jak to bude např. pro klasifikaci pacientů?

Hodnocení klasifikátoru IV

Diskutujte graf - co se dozvídáme o klasifikátoru?