

w Toto vyzerá na výcuc z ktorého je vhodné sa učiť, ten druhý čo ste dali, tak vtedy bol prednášajúci ten Čech. Či sa mýlim?

-

// 2018 nie je tu toho moc či? **//nestacia ti otazky z minulych rokov??**

Prednášky majú dokopy cez 400 strán :D Kašlat na výcic // :D

MR: Vytvoril som kopiu dokumentu, ktorý je na fiitkarovi k roku 2007, ale otázky sa podobajú na 2011. Necham tu aj povodne odpovede ľudí, čo to vypracovávali, ak by bolo niečo zle tak to oznacte a napiste dovod preco je to zle.

RT2013:

<https://www.dropbox.com/s/a8lx9mt7bjm2t6o/2013-05-30%2009.53.57.jpg> //hadze 404

<https://www.dropbox.com/s/v5ugh9s3qif1k74/2013-05-30%2009.54.10.jpg> //hadze 404

Pripadne (Dropbox je citlivy na premenovanie a presunutie suborov):

<http://cl.ly/3E0g2Q2X3r0h>

**RT2014: len kombinacia predoslych prikladov**

**RT2015: len kombinacia predoslych prikladov // toto nie je celkom pravda, niektore ulohy tam boli nove, resp. v nich boli podstatne zmeny. Vytvoril som novu kapitolu RT 2015 kde sa pokusim spisat novinky**

## **Obsah:**

[Skúška RT 2018](#)

[Skúška RT 2017\(nahradny\)](#)

[Skúška RT 2017](#)

[Skúška RT 2007](#)

[Doplnenie zo skúšky 2011](#)

[Ďalšie dôležité pojmy](#)

[Doplnenie zo skúšky 2012](#)

[Skúška OT 2013](#)

[Skúška RT 2013](#)

[Skuska RT 2015](#)

a

## Skúška OT 2018

Multichoice

1. Základné parametre atribútu (7b)
  - a. Medián
  - b. Centroid
  - c. Aritmetický priemer
  - d. Single link
  - e. Average link
  - f. X2 štatistický
  - g. ???
2. R

## Skúška RT 2018

Multichoice

1. Pocet binarnych atributov (3b)
  - a. K
  - b. K-1**
  - c. ? nieco jak pocet potrebný na zakodovanie premennej do binárnej sústavy
2. Ako sa odstraňujú atribúty (7b)
  - a. Histogram
  - b. Korelačný koeficient
  - c. PCA**
  - d. Rozhodovací strom
  - e. chi-kvadrat
  - f. ??
  - g. ??
3. Bayes teorka
4. Identifikacia stupňa polynomu teorka
5. Asociacné pravidla teorka
6. k-means

Pisomne

- + Sefredaktor a clanky (13b)
- + Vzdialenosti text, bin, num (3b)

- + Akú techniku vyhodnocovania zhukovania by ste použili na opis charakteristík jednotlivých zhukov?

## Skúška RT 2017(nahradny)

- 1 - A. co je to dolovanie z dat a na com je zalozene  
B. popiste jednotlivé kroky predspracovania a príklad ku kazdemu  
C. vymenujte druhy atributov dolovania a ku kazdemu príklad
- 2 - A. aky princíp ma dolovanie v sekvenciach a vymenujte 2 príklady  
B. aky princíp ma dolovanie v grafoch a povedzte príklad aplikacnej oblasti  
C. charakterizujte povahu toku dat a niektore jeho problémy pri jeho spracovaní
- 3 - A. co je to asociacne pravidlo a povedzte jeho metriky  
B. ako funguje algoritmus apriory a uvedte jeho apriory pravidlo  
C. ako sa robia asociacne pravidla z frekventovaných množín
- 4 - A. ako funguje algoritmus vytvorenia klasifikačného stromu a v ktorej fáze sa vytvára  
B. aky je rozdiel medzi metodami ID3, C2.nieco a gini index
- 5 - A. ake znalosti sa dajú vytážiť z webu  
B. čím je web odlišný od textu  
C. uvedte 3 metódy dolovania vo webe

## Skúška RT 2017

Na niektoré som určite zabudol, ale 3-5 bodov na podotázku, teda 12-20 otázok

1. Metriky
  - a. Metriky stredných hodnôt, 3 opísať, výhody nevýhody, typy atribútov, na ktoré sa používajú
  - b. Metriky rozptylu
  - c. Opíš krabicový graf, aké informácie v ňom sú
2. Zhukovanie
  - a. podľa rozdelenia (partition) opísať, uviesť príklad, opísať algoritmus, výhody, nevýhody
  - b. Podľa hustoty, to isté
  - c. Metriky porovnania zhukovania, 1 opísať na čom je založená
3. Klasifikácia
  - a. Opísať princíp, kroky
4. Otázky okolo dolovania z webu

- a. Opísať PageRanking a HITS
  - b. Opísať ako data získavame z analýzy štruktúry a použitia webu, na čo sa podobajú
  - c. Kroky získavania dát použitia
5. Asociačné pravidlá
- a. Definuj frekv. Množinu a podporu
  - b. Nevýhody apriori a ako ich rieši FP strom
  - c. Definuj výpočet spoľahlivosti

## Skúška RT 2007

**1. Atribút teploty môže nadobúdať hodnoty z intervalu  $<0, 100>$ . Klasifikátor, ktorý chcete použiť však potrebuje, aby všetky hodnoty boli binárne. Vysvetlite ako pretransformujete tento numerický atribút . Jednotlivé kroky transformácie vysvetlite. (4 body)**

Treba ratat s tým že teplota je spojita velicina a preto nie je mozne priame mapovanie medzi numerickymi atributmi a binarnymi hodnotami. Preto treba vykonat nasledovne kroky :

1. Zoradime vsetky hodnoty napr. od najmensieho po najvacsie
2. Spravime binning tj. vytvorenie intervalov z takto usporiadanej mnoziny cisiel. V tejto situacii sa zda najlepsie riesenie konkretne equal-interval binning (znamena to ze rozsah jednotlivych intervalov je rovnaky -> to neznamenava ze pocetnost je rovnaka)

3. Po binningu nam teda ostane 101 binarnych atributov (interval  $<0,100>$  je uzavrety z oboch stran).

Pozor : 101 binarnych atributov je preto lebo sme si zvolili velkost intervalu 1. Kludne si mozeme zvolit aj vacsie rozmedzie intervalu a potom bude pocet binarnych atributov iny. Tuto je mozno vhodne spomenut, ze pomocou si mozeme histogramami, ktore nam ukazu rozlozenie teplot a na zaklade toho vieme ze co je pre nas najlepsi sposob pri binningu (je kludne mozne pouzit aj binning na zaklade diskretizacie podla tried napr. 3 z 1, alebo 6 z 1, zalezi na datach)

// nie je ale binarnych atributov 100? 1.:0-1...2.:1-2....3.2-3 atd az po 100. //+3

// nenazyval by som uz to v tomto kroku *binarnymi* atributmi //to budu este len zaokruhlene cisla ci? //no nie zaokruhlene, ale jednotlivé intervaly s rovnakou width, do ktorých sa ide roztriedovať // hm tak som to teda asi zle pochopil, myslel som že napr číslo 0,4 by to dalo na číslo 1. // ja to chapem tak že take číslo vložis do intervalu  $<0, 1)$  a tým pádom z neho spravíš diskretný atribut (zaradil si ho do príslušného *bin-u*). //to dava vacsi zmysel :D dik

//ak by boli intervaly takto  $<- , 1)$ ;  $<1, 2)$  ...  $<99, +)$ ;

195Sauce: [http://www.saedsayad.com/unsupervised\\_binning.htm](http://www.saedsayad.com/unsupervised_binning.htm)

Tak ich je 100 pricom width je 1

4. To ako sa to presne robí nie je nikde presne popisane. Kazdopadne toto je jedno z mozných

rieseni. Bud tieto cele cisla budeme priamo prevadzat do binarnej sustavy, alebo proste len ake velke cislo, tolko jednotiek bude mat vid tabulka :

	a1	a2	a3	a4	...	a100
0	0	0	0	0	...	0
1	1	0	0	0	...	0
2	1	1	0	0	...	0
3	1	1	1	0	...	0
...						
100	1	1	1	1	...	1

Takto sa zabezpeci, ze susedne hodnoty maju vzdy rozdiel len v jednom atribute a povodna hodnota sa da zrekonstruovat aj keby sa atributy poprehadzovali.

### Iné riešenie //+1

Podľa prednášky 02\_preprocessing, slide 65:

## Konverzia diskretných atribútov na numerické

- diskretné: ordinálne, kde môže byť nejaké usporiadanie
- transformovanie k-hodnotového atribútu do k-1 binárnych atribútov  $a_1, \dots, a_{k-1}$ 
  - i premenná vyjadruje, či je hodnota pôvodného atribútu  $\geq i+1$
  - zoradenie: susedné atribúty sa líšia v 1 atribúte
  - pre  $\{1, \dots, 5\}$ :

	$a_1$	$a_2$	$a_3$	$a_4$		$a_1$	$a_2$	$a_3$	$a_4$		$a_1$	$a_2$	$a_3$	$a_4$		
1:	0	0	0	0		3:	1	1	0	0		5:	1	1	1	1
2:	1	0	0	0		4:	1	1	1	0						

f

Teda stačí 99 binárnych atribútov. Samozrejme, ak sú pôvodné hodnoty spojité, treba ich transformovať na diskretné (celé čísla). //nestaci 99 atributov, treba 100 kedze je 101 hodnot

// ako sa transformuje realne cislo na cele? len zaokruhlenim?

// nie len zaokruhlenim, ale zaradenim do prislusneho intervalu +1

**2. Uvedte čo je cieľom normalizácie a vysvetlite, kedy je potrebné ju použiť. Uvedte jednu konkrétnu metódu normalizácie (vzorec). (3 body)**

Cielom je previesť hodnoty atributu z povodneho rozsahu do standardizovaneho rozpatia, napr  $<0,1>$  alebo  $<-1,1>$ , cim sa zabezpeci rovnocennost hodnot normalizovanych atributov, ktore su merane v roznych jednotkach. Toto je dolezite ak chceme pracovat s datami ako s rozdielovymi.

### Min-Max normalizácia

- pre každý atribút

- minA - najnižšia hodnota
- maxA - najvyššia hodnota

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

do intervalu  $<0,1>$

keď príde v budúcnosti inštancia, ktorej hodnota atribútu je mimo  $<\min_A, \max_A>$ , konci chybou

#### Zero-mean normalization

dáta sú normalizované na základe

- aritmetického priemeru atribútu A
- štandardnej odchýlky z hodnôt atribútu A:  $S_A$

$$v' = \frac{v - \bar{A}}{S_A}$$

• keď príde v budúcnosti inštancia, ktorej hodnota atribútu je mimo  $<\min_A, \max_A>$ , nie je to problém

- **nie je to** do intervalu  $<0,1>$ , ale mean je = 0

?? možno aj toto: <http://stat.ethz.ch/R-manual/R-patched/library/base/html/scale.html>

Keď to správne chapem tak interval môže byť teoreticky  $<-\infty, +\infty>$

<http://math.stackexchange.com/questions/362918/value-range-of-normalization-methods-min-max-z-score-decimal-scaling>

Cielom normalizácie je aj

- prevod numerických hodnôt na pomerne
- Skalovanie hodnôt do intervalu (vyššie)
- Standardizovanie jednotiek (penazne meny, rozdiel v kalendároch)

**3. Vyjadrite podporu (ang. support) a spoľahlivosť (ang. confidence) pre pravidlo IF maslo AND moka THEN vajcia pre množinu transakcií z tabuľky 1. Nie je potrebné vypočítať presné číslo, stačí ho vyjadriť v tvare číselného výrazu. (6 bodov)**

číslo	Položky
1	Mlieko, vajcia, cereálie
2	Maslo, vajcia, múka
3	Mlieko, maslo, vajcia, múka
4	Maslo, múka
5	Múka, vajcia

Tabuľka 1. Dáta nákupného košíka

Pravidlo IF maslo AND muka THEN vajcia môžeme vyjadriť ako  $X \Rightarrow Y$   
kde X a Y sú dve množiny definované ako:

$X = \{\text{maslo, muka}\}$

$Y = \{\text{vajcia}\}$

**Podpora(support)** vyjadruje pravdepodobnosť, že sa uvedené prvky nachádzajú v košíku (v transakciách).

Výpočet:

$s(X,Y) = \frac{|X \cup Y|}{n}$  (X zjednotenie Y) / počet všetkých transakcií

Takže pre prvky [maslo,muka,vajcia] je podpora = 0,4. Pretože táto kombinácia sa v nákupných košíkoch nachádza 2 krát a teda  $2/5 = 0,4$ .

**B:** podľa wiki je support podiel tých transakcií, v ktorých sa nachádza X oproti všetkým transakciám, čiže by to bolo % // +4

[https://en.wikipedia.org/wiki/Association\\_rule\\_learning#Useful\\_Concepts](https://en.wikipedia.org/wiki/Association_rule_learning#Useful_Concepts)

podpora (angl. support): podiel tých instancií, ktoré vyhovujú pravidlu

$(X \cup Y)$  oproti všetkým instanciám (n) //prednaska asociacne pravidla cize podľa tohto je správna odpoveď 0,4

V prednáške 2018 je: podiel instancií vyhovujúcich pravidlu vs všetky instance

<https://www.youtube.com/watch?v=iuIEVnhWtlw> // M: z tohto videa možno vidieť, že **su dva spôsoby, akými sa chape support**, teda buď hovoríme o pravdepodobnosti, že naše pravidlo platí  $A \Rightarrow B$ , kde je výsledok ako p (v našom prípade p = 2) alebo support ako podiel výskytov A (antecedent t) - v našom prípade t = 3. Potom výsledná spoľahlivosť (confidence) sa vypočíta ako  $p/t = \frac{2}{3}$ .

**Spolahlivost(confidence)** - podiel tych instancií, ktoré vyhovuju pravidlu  $(X \cup Y)$  oproti tým, na ktoré sa dá pravidlo aplikovať ( $X$ )

[maslo, muka] -> sa nachadza v 3 prípadoch -> to je naše  $X$

[maslo, muka, vajcia] ->  $X \cup Y$  -> sa nachadza 2 krát

$\alpha = \frac{2}{3} = 0.66666$

- veľké množstvo asociačných pravidiel  $X \Rightarrow Y$ , preto nás zaujímajú len tie, ktoré

- pokrývajú dostatočné množstvo inštancií – *podpora* (angl. support): podiel tých inštancií, ktoré vyhovujú pravidlu  $(X \cup Y)$  oproti všetkým inštanciám ( $n$ )

- sú dostatočne presné – *spoľahlivosť* (angl. confidence): podiel tých inštancií, ktoré vyhovujú pravidlu  $(X \cup Y)$  oproti tým, na ktoré sa dá pravidlo aplikovať ( $X$ )

// Podľa prednášky má byť teda support % a confidence % :) /+1 Spýtame sa na skúške keď tam bude taká otázka a je :D

// je tam jasne napísané  $X$  zjednotenie  $Y$ , takže support = %. A už tu nespekulujte dookola...

4. Aké techniky vyhodnocovania zhlukovania by ste použili na opis charakteristík jednotlivých zhlukov? Vysvetlite prečo. (5 bodov) Akú techniku zhlukovania by ste použili na opis charakteristík jednotlivých zhlukov? Vysvetlite prečo. Aké charakteristiky zhlukov takto získate? //takto presne znela otázka na skúske VIE NA TOTO NIEKTO ODPOVEDAť?

//Nejde tu náhodou o techniku pravdepodobnostného zhlukovania ? Tam sa jednotlivé zhľuky opisujú cez parametre, ktoré musí mať instancia aby patrila do zhluku. Ak som to správne pochopil. - s tým že každý zhľuk má 3 parametre - 2 pre normálne rozdelenie a 1 pre pravdepodobnosť že sa niečo vyskytuje v danom zhluku ale to neviem presne 05\_zhlukovanie.pdf - 42 až 48ú

// Konceptuální shlukování je proces, který se snaží pro objekty nalézt klasifikační schéma. Na rozdíl od běžného shlukování nehledá pouze skupiny objektů, ale zároveň se snaží nalézt charakteristický popis pro každou skupinu objektů (třidu). // Nájdené v opore z vlašajška (od Zendulky). Podľa mňa toto je správna odpoveď a možno vtedy sa to ešte vyučovalo ale keď teraz nie je toto v našich prednáškach tak podľa mňa nehrozí, že nám dajú túto otázku

// nemože to byť toto? Jedine čo mi dáva "zmysel", ak to má byť naozaj *technika zhlukovania*, je tam príklad so sieťnicou a mozgovou korou za tým



## Samoorganizujúce sa mapy

- **neurónové siete** – učenie učiteľ'a
- **adaptácia váh** odzrkadľuje **štatistické vlastnosti trénovacej množiny**
- **topografické mapy** – zobrazenie zachovávajúce topológiu (**charakteristických črt**)
  - **výstupná vrstva** - pravidelná štruktúra (mriežka, reťaz)
    - podobné vstupy evokujú odozvy na fyzicky blízkych neurónoch
  - inšpirované biologickými neurónovými sieťami vyšších cicavcov (mozgová kôra)
  - efektívny spôsob reprezentácie parametrov vstupných dát
  - **projekčné oblasti** – mapa povrchu tela, vypočítané – vizuálne a sluchové mapy

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ≡ ≡ ↺ 🔍 ↻

## //What?

**MR:** dal by som tam len vymenovane techniky zhlukovania + ako sa robia opis, viac neviem povedat.

MF: kedze tu chceme charakterizovat zhluky, nemoze to byt: purity, NMI, rand index a f-measure? ←SOM ZA TOTO (J)

**MZ:** podľa mňa nie, pretože tie ukazovatele (purity, nmi...) sa používajú na porovnanie konkrétnych zhlukov so zlatým štandardom (klastrovanie odborníkom). neopisujú charakteristiky jednotlivých zhlukov, ale akoby správnosť už celého výsledku oproti tomu, ako človek povedal, že by to malo byť.

- porovnanie s tzv. *zlatým štandardom*
  - zlatý štandard – priradenie inštancií do skupín - ideálne produkované ľuďmi
- externé kritériá pre vyhodnocovanie výsledkov zhukovania
  - vyhodnocuje ako dobre sa zhukovanie zhoduje so zlatým štandardom
  - miery:
    - purity
    - normalized mutual information
    - rand index
    - F-measure

Edo: zeby <http://en.wikipedia.org/wiki/Tf%E2%80%93idf>

MR: No neviem, to mi celkom nesedi. Tato otazka je pre mna velkou zahadou :D

PG: tf-idf je iba pre texty myslim ze tu je potrebne nieco vseobecnejsie, ale co to netusim, ale priklanam sa k MF, aj ked to su charakteristiky, neviem ci to spada aj pod techniky???

**MS: Podla mna zle napisana otazka [+6] :(**

**MI:**

**tie metody tam spominane su aj v prednaskach ak sa nemylim**

MZ: vzhľadom na to, že tá otázka nedáva zmysel - techniky zhukovania sa používajú na zhukovanie, nie jeho vyhodnocovanie, tak odhadujem, že to malo byť "Aké techniky **vyhodnocovania** zhukovania by ste použili na opis charakteristík jednotlivých zhukov" - vypadlo jedno slovo :)

**5. Uvažujte problém porovnania dvoch metód na zhukovanie aplikovaných na jednorozmerné dáta a ich rozdelenie na 2 zhuky. Výsledok oboch metód sú pravdepodobnosti  $P_j(x_k|x_1)$  kde  $x_1 = (x_1..x_k)$  je množina dát  $x_k$   $k=\{1,2\}$  predstavuje zhuk a horný index  $j = \{1,2\}$  predstavuje označenie metódy na zhukovanie. Ako by ste merali kompaktnosť zhukov, ktorý by sa dala použiť na porovnávanie výsledkov použitých metód. Uvedte vysvetlenie a formálny vzťah. (6 bodov)**

Kompaktnosť zhukov sa meria Dunnovým alebo Davies-Bouldinovým indexom. Zhuk je kompaktný, keď vo vnútri zhuku sú navzájom podobné dáta a podobnosť medzi zhukami je čo najmenšia.

nevieme vizualizovať veľ'arozmerný priestor

- cieľ: identifikovať kompaktné a dobre separované zhľuky
- pre množinu zhľukov  $\mathcal{C} = C_1, \dots, C_k$  – indexy využívajúce:
  - vzdialenosti zhľukov  $\delta(C_j, C_m)$  (intercluster)
  - vzdialenosti v rámci zhľuku  $\Delta(C_j)$  (intracluster)

J: Čo tak siluetová validačná metóda - priemerné s(i) zhľuku? (meria ako pevne sú dáta zoskupené v zhľuku) // výborný nápad! +1

MZ: Ale tu je priradenie inštancií do zhľukov vyjadrené pravdepodobnosťou. Samozrejme, ak sa inštancie priradia k tomu zhľuku, pre ktorý je najväčšia pravdepodobnosť, že doň patria, tak tieto miery sú podľa mňa OK. Ale myslím si, že je dobré využiť práve uvedenú pravdepodobnosť, potom vyhodnotenie - vierohodnosť (v prednáškach označaná ako L) sa môže vypočítať pre obe metódy a porovnať. Alebo využiť AIC a BIC. AIC - meria, ako dobre štatistický model reprezentuje dáta, BIC - kritérium pre výber modelu z triedy parametrických modelov s rôznym počtom parametrov. Oba počítajú s vierohodnosťou L,

#### ■ *Dunnov index*

$$V(\mathcal{C}) = \min_{j,m: 1 \leq j \leq k, 1 \leq m \leq k, m \neq j} \left\{ \frac{\delta(C_j, C_m)}{\max_{1 \leq i \leq c} \Delta C_i} \right\}$$

- maximalizovať

#### ■ *Davies-Bouldinov index*

$$DB(\mathcal{C}) = \frac{1}{c} \sum_{j=1}^k \max_{m \neq j} \left\{ \frac{\Delta C_j + \Delta C_m}{\delta(C_j, C_m)} \right\}$$

- minimalizovať

ale treba poznať počet parametrov k.

Podľa mňa je cieľom tejto otázky práve napísať viac možností a povedať, ktorá sa kedy hodí a v čom je dobrá.

## 6. Vysvetlite pojem stratifikovaná krížová validácia (stratified cross validation). Čo získame jej použitím? (3 body)

Množina instancií sa nahodne rozdelí na k podmnožín, pričom sa vždy jedna podmnožina použije ako testovacia a ostatných k-1 podmnožín na trenovanie. Tento postup sa opakuje tak, aby sa ako testovacia podmnožina použila každá z k podmnožín. Výsledky z testovania sa nakoniec skombinujú do výsledného odhadu. V "stratified" verzii je rozdeľovanie do podmnožín riadené tak, aby priemerná výstupná hodnota bola rovnaká pre všetky trenovacie a testovacie podmnožiny, toto je vhodné ak sú výstupné hodnoty dichotomné a tieto dve rôzne hodnoty sú nevyvážene zastupované v dátach.

stratifikácia: nahodný výber tak, aby každá trieda bola približne rovnako zastupovaná v trénovacej

aj testovacej množine+

// dichotomne znamena nieco ako binarne? //(ano boolovske, binarne)

**7. Vysvetlite pojem PageRank a uveďte rekurzívny vzťah na jeho výpočet. (3 body)**

- autorom je Larry Page (používa google)
- meria prestíž - nezávisle na dopyte (narozdiel od HITS)
- prestíž stránky je proporcionálna suma prestíží stránok, ktoré na ňu odkazujú
- Predpokladajme že: graf je silno prepojený (z každého vrcholu v existuje cesta do vrcholu +u)  
používateľ začne na náhodnom uzle u s pravdepodobnosťou  $p_0[u]$  a kliká náhodne do nekonečna

Vztah na vypocet PageRank je nasledovny :

$$p[v] = (1 - d) \left( \frac{p[u_1]}{N_{u_1}} + \dots + \frac{p[u_n]}{N_{u_n}} \right) + d$$

označenie:

- $u_1, \dots, u_n$  – stránky odkazujúce sa na stránku  $v$
- $d$  – tlmiaci faktor  $d = 0.15$
- $N_{u_i}$  – počet liniek vychádzajúcich zo stránky  $u_i$
- $p(\cdot)$  – PageRank (prestíž) stránky

Rekurzívny je prave preto ze ked dostanem  $p[v]$  tj. pagerank  $v$ , tak tym padom sa musia prepocitat pagerank ostatnych stranok na ktore stranka  $v$  ukazuje (stranka  $v$  obsahuje linky, na ine stranky)

**8. Uvažujte problém diagnostikovania pomocou klasifikátora Naive Bayes. Sledované symptomy su kýchanie, horúčka, kašeľ, pričom každý z týchto symptomov môže nadobúdať hodnoty ÁNO, NIE. Diagnóza nadobúda jednu z hodnôt Zdravý, Alergia, Prechladnutie. Tabuľka 2 obsahuje vierohodnosti (podmienené pravdepodobnosti jednotlivých symptómov pre dané diagnózy). Apriórna pravdepodobnosť diagnóz je  $P(\text{zdravý}) = 0.8$ ,  $P(\text{alergia}) = 0.1$ ,  $P(\text{prechladnutie}) = 0.1$ . Určte výslednú diagnózu pacienta s týmito symptomami: kýchanie = ÁNO, kašeľ = ÁNO, horúčka = NIE. Svoje rozhodnutie podložte výpočtom. (5 bodov)**

Symptóm/diagnóza	Zdravý	Alergia	Prechladnutie
Kýchanie = ÁNO	0.1	0.8	0.8
Kašeľ = ÁNO	0.1	0.7	0.8
Horúčka = ÁNO	0.01	0.4	0.7

Tabuľka 2. Vierohodnosti – podmienené pravdepodobnosti symptómov pre jednotlivé diagnózy

Instancia:

$d = \{\text{kýchanie, kasel, } \neg\text{horucka}\}$

Hodnoty z tabuľky sa dajú prepísať takto:

$P(\text{kýchanie} \mid \text{zdravý}) = 0.1$      $P(\text{kýchanie} \mid \text{alergia}) = 0.8$      $P(\text{kýchanie} \mid \text{prechladnutie}) = 0.8$

$P(\text{kasel} \mid \text{zdravý}) = 0.1$      $P(\text{kasel} \mid \text{alergia}) = 0.7$      $P(\text{kasel} \mid \text{prechladnutie}) = 0.8$

$P(\text{horucka} \mid \text{zdravý}) = 0.01$      $P(\text{horucka} \mid \text{alergia}) = 0.4$      $P(\text{horucka} \mid \text{prechladnutie}) = 0.7$

Keďže máme v instancii negáciu horucky, tak pravdepodobnosti sú len komplement k tretiemu riadku:

$P(\neg\text{horucka} \mid \text{zdravý}) = 0.99$ ,  $P(\neg\text{horucka} \mid \text{alergia}) = 0.6$ ,  $P(\neg\text{horucka} \mid \text{prechladnutie}) = 0.3$

Vyjadríme pravdepodobnosť pre každú z možných tried v závislosti od našej instance: zdravý, alergia, prechladnutie (Bayesov vzorec):

$P(\text{zdravý} \mid d) = (P(d \mid \text{zdravý}) * P(\text{zdravý})) / P(d)$

$P(\text{alergia} \mid d) = (P(d \mid \text{alergia}) * P(\text{alergia})) / P(d)$

$P(\text{prechladnutie} \mid d) = (P(d \mid \text{prechladnutie}) * P(\text{prechladnutie})) / P(d)$

$P(d \mid \text{zdravý}) = P(\text{kýchanie} \mid \text{zdravý}) * P(\text{kasel} \mid \text{zdravý}) * P(\neg\text{horucka} \mid \text{zdravý}) = 0.1 * 0.1 * 0.99$   
 $= 0.0099$  // z čoho vyplýva, že to môžeme rozdeliť na takýto súčin? // prednáška 3 slide 41

$P(d \mid \text{alergia}) = P(\text{kýchanie} \mid \text{alergia}) * P(\text{kasel} \mid \text{alergia}) * P(\neg\text{horucka} \mid \text{alergia}) = 0.8 * 0.7 * 0.6$   
 $= 0.336$

$P(d \mid \text{prechladnutie}) = P(\text{kýchanie} \mid \text{prechladnutie}) * P(\text{kasel} \mid \text{prechladnutie}) * P(\neg\text{horucka} \mid \text{prechladnutie}) = 0.8 * 0.8 * 0.3 = 0.192$

$P(d \mid \text{zdravý}) * P(\text{zdravý}) = 0.0099 * 0.8 = 0.00792$  /// (odkiaľ sa bere to 0.8)? // zo zadania+1

$P(d \mid \text{alergia}) * P(\text{alergia}) = 0.336 * 0.1 = 0.0336$

$P(d \mid \text{prechladnutie}) * P(\text{prechladnutie}) = 0.192 * 0.1 = 0.0192$

$P(d) = 0.00792 + 0.0336 + 0.0192 = 0.06072$

\*

$P(\text{zdravý} \mid d) = 0.00792 / 0.06072 = 0.1304$

$$P(\text{alergia} | d) = 0.0336 / 0.06072 = 0.5534$$

$$P(\text{prechladnutie} | d) = 0.0192 / 0.06072 = 0.3162$$

Skuska spravnosti je taka, ze sucet vsetkych pravdepodobnosti, pre vsetky triedy by mal byt 1. Najvacsia pravdepodobnost je pre alergiu, takže instancia by bola klasifikovana ako ALERGIA.

**9 Uvažujte klasifikačný problém rozdelenia filmov podľa toho, či sa nám páčia do dvoch tried ÁNO, NIE na základe žánru filmu a režiséra. Tabuľka 3 obsahuje 12 trénovacích príkladov. Uveďte, ktorú veličinu je potrebné vypočítať na určenie, ktorý z atribútov bude koreňom rozhodovacieho stromu pri použití klasifikátora ID3. V tvare číselného výrazu vyjadrite túto veličinu pre oba atribúty (nie je potrebné hodnotu vypočítať). Uveďte podľa akého kritéria atribút vyberiete. (7 bodov)**

Film	Žáner	Režisér	Páči sa?	Film	Žáner	Režisér	Páči sa?
F1	Triler	Bergman	NIE	F7	Dráma	Bergman	NIE
F2	Komédia	Spielberg	ÁNO	F8	Dráma	Spielberg	NIE
F3	Komédia	Spielberg	NIE	F9	Dráma	Hitchcock	ÁNO
F4	Triler	Bergman	NIE	F10	Komédia	Spielberg	NIE
F5	Komédia	Hitchcock	ÁNO	F11	Triler	Spielberg	NIE
F6	Dráma	Bergman	ÁNO	F12	Triler	Hitchcock	NIE

Tabuľka 3. Dáta pre príklad 9.

Riešenie, ak nie je správne, opravte.

**Uveďte, ktorú veličinu je potrebné vypočítať na určenie - Informačný zisk**

**Pre atribúty:** Žaner, Režiser

**Žaner:**

komedia = [2,2]

triler = [0,4]

drama = [2,2]

-----  
spolu [4,8]

Vzorec pre informacny zisk:

$$\text{gain}(\text{zaner}) = \text{info}[4,8] - \text{info}([2,2],[0,4],[2,2])$$

postupne vyratame jednotlivé casti:

$$\begin{aligned} \text{info}[4,8] &= H(4/12, 8/12) = 4/12 * \log_2(12/4) + 8/12 * \log_2(12/8) = 0,3333 * 1,5849 + 0,6666 \\ &* 0,5849 = 0,5282 + 0,3898 = 0,918 \end{aligned}$$

//pls vysvetlite niekto preco je tam log2 ??? v prednaske som nasiel ze tam je log ... a ked si to prepocitavam tak mi to nevychadza ... co je log2 vlastne? logaritmus dvojky alebo logaritmus pri zaklade 2?

z prednášky 03, slide 65: logaritmus je obyčajne so základom 2

//  $\log_2(8) = 3$ , cize zaklad je 2

$$\begin{aligned} \text{info}([2,2],[0,4],[2,2]) &= (4/12) * \text{info}[2,2] + (4/12) * \text{info}[0,4] + (4/12) * \text{info}(2,2) = 2 * [(4/12) * \\ \text{info}[2,2]] &\quad // \text{info}[0,4] = 0 \end{aligned}$$

$$\begin{aligned} \text{info}[2,2] &= H(2/4, 2/4) // \text{komedie aj dramy obsahuju po 4 instance} \\ &= 2/4 * \log_2(4/2) + 2/4 * \log_2(4/2) = 1 \end{aligned}$$

dosadime do vzorca:  $\text{info}([2,2],[0,4],[2,2])$

$$\text{a dostaneme: } 2 * [0,3333 * 1] = 0,6666$$

dosadime do hlavneho vzorca:  $\text{gain}(\text{zaner}) = \text{info}[4,8] - \text{info}([2,2],[0,4],[2,2])$

$$\text{a dostaneme: } \text{gain}(\text{zaner}) = 0,918 - 0,6666 = 0,252$$

postup opakujeme aj pre atribut reziser a atribut s vacsim informacnym ziskom sa stava korenou stromu

$\text{gain}(\text{reziser}) = 0,1012$  //pokiaľ nenastala chyba niekde vo vypočte. Mohol by to niekto prepocitat pre kontrolu.

**EDIT: mne to vyslo inak:**

**B [1,3]**

**S [1,4]**

**H [2,1]**

$$\text{info}[4,8] - \text{info}([1,3],[1,4],[2,1])$$

$$0,9183 - (4/12 * 0,811 + 5/12 * 0,7223 + 3/12 * 0,918) = 0,1185$$

**tvoj vysledok vychadza vypoctom  $0,9183 - (4/12 * 0,808 + 4/12 * 0,7223 + 4/12 * 0,918)$ , nemali by byt vsade 4/12**

**aj podľa mňa to je 0.1185 avšak na výsledku to nič nemení +1**

**Zdroj:** 03\_klasifikacia, 57slide+

//ako to teda je, má sa to ratovať ako 4/12, 5/12, 3/12 //toto +1

alebo je správne všade dať 4/12?

//4/12, 5/12, 3/12 - tam sa berie pravdepodobnosť, že pôjdeme po tej danej vetve, čiže keď od Spielberga je 5 filmov z 12-tich, treba pre množstvo informácií ratovať s pravdepodobnosťou 5/12 z prednášky: "Priemerné množstvo informácií pre uzol (atribút): množstvo informácií pre jednotlivé vetvy váhované počtom inštancií " // +1

**MR:** Pre bližšie pochopenie problematiky stacia slajdy 62,64,66 z 03\_klasifikacia.pdf

**Záver:** vyberie sa atribút žánru ako koreň stromu

**10. Mobilný operátor ABC každý mesiac opúšťa 1% zákazníkov. Priemerná cena získania nového zákazníka 500SKK. Manažér spoločnosti sa rozhodol, že je lepšie poslať 300SKK poukážku každému zákazníkovi, ktorý sa chystá od spoločnosti odísť. Tak by si spoločnosť mohla týchto zákazníkov udržať. Ako konzultant firmy ABC pomôžte nájsť zákazníkov, ktorí sa chystajú prestať využívať služby ABC tak, aby sa to firme oplátiло. Odpovedzte na tieto otázky:**

- a) Aké atribúty by ste zbierali o zákazníkoch?
- b) Z akého časového obdobia by ste zbierali dáta?
- c) Aké DM techniky by ste použili (a v akom poradí)?
- d) Aký typ znalostí očakávate ako výsledok zvolených DM techník?
- e) Ako by ste overovali svoje výsledky?

a)

Tu je to úplne jedno ale povedzme že : vek, pohlavie, zamestnanie, mesačné faktúry, ako dlho je klientom,

*keďže mobilný operátor, tak má k dispozícii presné dáta o volaní a využití služieb (internet, sms a pod.) podľa týchto dát potom môže ponúknuť lepšiu ponuku*

// + ak by som v c) zvolil klasifikáciu, tak má zaujímavé dáta o zákazníkoch ktorí už odišli, nad ktorými by som natrénovával model

b)

Je to veľmi individuálne ale povedzme že posledný rok.

c)

redukcia dimenzionality (odstránim tie atribúty, ktoré nie sú potrebné), //ako môžem začať s



redukciou dimenzionality ked este neviem (pred natrenovanim a klasifikovanim) ktore atributy niesu potrebne resp. maju najmensiu vahu?ci tu niesu vypisane v poradí?

skontrolovanie atributu (z viacerých atribútov viem napríklad spraviť jeden atribut -> napr. vek zákazníka)

agregácia -> niektoré (hlavne časové) hodnoty sa dajú ľahko agregovať napr. na základe mesiacov

*equal-interval binning pre vek ?*

Použijem Naive Bayes lebo je asi najefektívnejší

DM techniky: klasifikácia, zhukovanie

a napríklad lift chart??

d) Očakávame transformovanú množinu, ktorú budeme môcť priamo použiť v niektorom z konkrétnych DM algoritmov (napr. ak chcem použiť k-nearest, tak sa budem snažiť optimalizácie aby boli atribúty rozdielové)

*Vystup: Natrenovaný klasifikátor, ktorý je schopný určiť na základe atribútov o existujúcich klientoch, že s akou pravdepodobnosťou opustia firmu - aby sme vedeli komu majú kontaktovať aby firmu neopustil*

e)

Najlepší spôsob je sledovať či naša klasifikácia správne určila ľudí, ktorí odišli. Ak sme sa netrafili, je nutné upraviť trenovaciu množinu (len tu treba dávať pozor aby sme ju nepreucili)

+ cross-validácia trenovacej množiny // *prípadne možno ešte stratifikácia (tych, čo odišli bude asi oveľa menej ako tých čo zostali) čo tak lift chart alebo ROC krivka ktoré sú na toto presne určene? // Určite by som dal aspoň ROC krivku, kde vidíš hneď TP a FP*

---

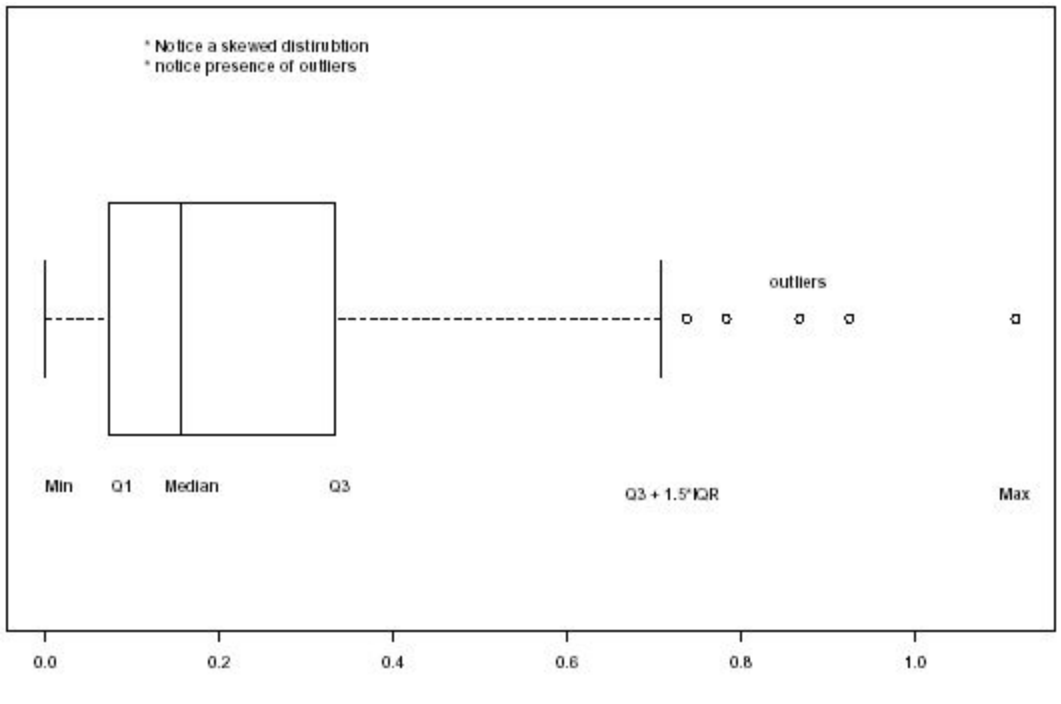
## 11. Boxplot, Scatterplot: čo sú, v ktorej fáze OZ sa používajú.

Obe sú nástroje vizualizovanej analýzy dát, ktoré sa používajú vo fáze výberu a predspracovania atribútov, tj. príprava dát.

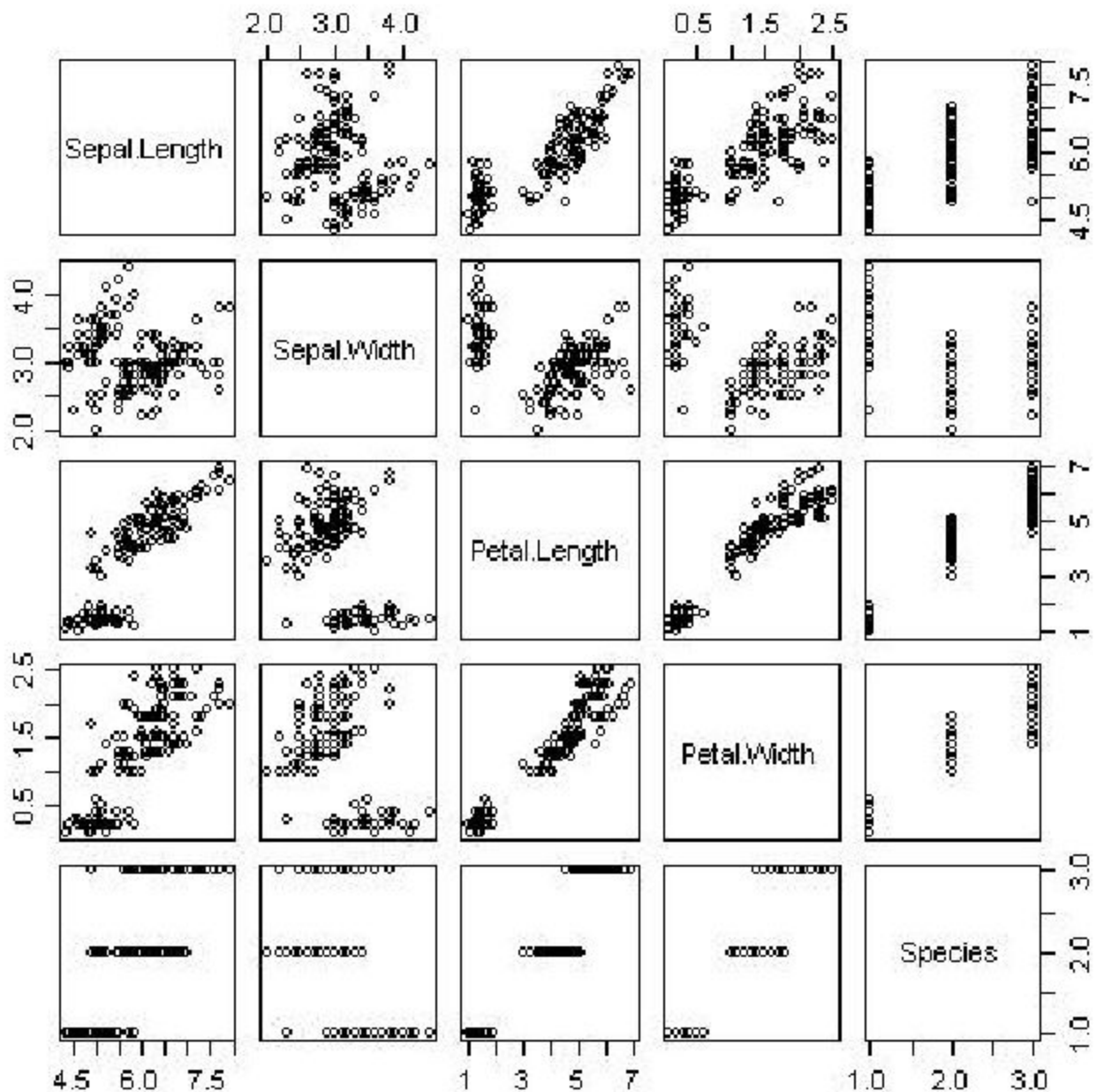
**Boxplot:** jeho cieľom je sumarizovať dáta a rýchlo zobrazíť, či sú dáta symetrické, či majú outlierov.

5-číselná sumarizácia

- rámček: dolný kvantil, medián, horný kvantil
- medián: v zoradených hodnotách - prostredná
- dolný kvantil: medián ľavej polovice čísel po mediáne (vrt.)
- horný kvantil: medián pravej polovice čísel od mediánu (vrt.)
- zarážky ("fúzy"): vyjadrujúce minimum a maximum (max 1,5x šírka rámčeka)
- kolieska: outlieri



**Scatterplot:** pomocou nej sa zisťujú vzťahy medzi dvojicou premenných (2D graf) - scatter matrix: sú grafy všetkých kombinácií dvojíc premenných



## 12. Kedy a prečo nemusí stačiť *podpora* a *spol'ahlivosť* pri dolovaní asociačných pravidiel?

dolovanie asociacnych pravidiel je identifikovanie vsetkych asociacnych pravidiel  $X \Rightarrow Y$  s minimalnou podporou  $s$  a spol'ahlivosťou  $a$  (alfa).

dolovanie asociacnych pravidiel prebieha v nasledujucich fazach:

1. hľadanie frekventovaných množín položiek (množiny položiek, ktoré majú väčšiu podporu ako prahová hodnota  $s$ )
2. generovanie pravidiel z frekventovaných množín položiek

pre kazde asociacne pravidlo  $X \Rightarrow Y$ , musi byt'  $X \cup Y$  frekventovana mnozina poloziek.  
Nachadzanie frekventovanych mnozin poloziek hoci je jednoduché ale je casovo narocne (pre  $m$  poloziek:  $2^{m-1}$  kandidatov tj. pre  $m=30$  to znamena 1 073 741 823 kandidatov).

Ztoho vyplyva aj odpoved na tuto otazku (nepotvrdena odpoved):

- ak sa spoliehame iba na podporu a spolahlivost, mozeme sa ocitnut v situacii, ze budeme mat vyssi pocet moznosti a teda aj prakticky nezvladnuteľny pocet kandidatov!!

//[J] správna odpoved' začína odtiaľto :)

**Metriky podpora a spolahlivost ignoruju  $P(B)$ ...** (07\_asoc\_prav slide 37) // co je  $P(B)$ ?

// podpora a spolahlivost porovnavaju len zhodne hodnoty v transakciach a nie ich koreláciu, teda nevyjadruju podobnosť transakcii // pri dolovaní takisto vytvaraju bez obmedzení veľke množstvo kandidatov na frekventované množiny // aplikovaním metrik zaujímavosti, alebo použitia pravidla získame lepšie vydolované pravidla

## Meranie kvality AP

- $s(A \Rightarrow B) = P(A, B)$
- $\alpha(A \Rightarrow B) = P(B|A)$
- ignorujú  $P(B)$ , preto iné miery:
  - $\text{interest}(A \Rightarrow B) = \frac{P(A, B)}{P(A)P(B)}$  (meria koreláciu medzi položkami v pravidle)
  - $\text{conviction}(A \Rightarrow B) = \frac{P(A)P(\neg B)}{P(A, \neg B)}$  (keďže  $A \Rightarrow B = \neg(A \wedge \neg B)$ , meria nezávislosť negácie implikácie)

//to sa neda povedať nejak rozumnejšie? Kto sa ma toto naucit? Nejake zhrnutie do jednej vety?

Jednou vetou asi takto: **nestaci to pri merani kvality asociacnych pravidiel, pretože podpora a spolahlivost ignoruju  $P(B)$  // doplnil by som: a neberie do úvahy koreláciu medzi položkami +1**

## Doplnenie zo skúšky 2011

**3. Definujte nasledovne pojmy :**

**krizova validacia (cross-validation)** - iteratívne sa vymienia trenovacia a testovacia množina.

Zvyčajne sa data rozdelia do  $n$  množín (nazýva sa to  $n$ -fold). Algoritmus pracuje tak že jedna množina je stanovená za testovaciu a zvyšné sú považované za trenovacie. Postupne sa iterujú tak, aby každá množina bola **testovacou**. Takýmto spôsobom vlastne dosiahneme

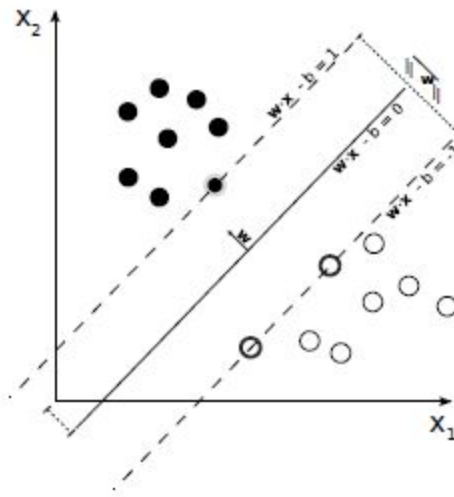
najvacsiu moznu testovaciou mnozinu a iterativne sa zlepšuje **vyhodnotenie**.

**stratifikacia** - nahodny vyber tak, aby kazda trieda bola priblizne rovnako zastupena v trenovacej aj testovacej mnozine

**bootstrap** - pri vybere instance z databazy do trenovacej mnoziny, instancia zostava v databaze. N-krat vyber s opakovanim z databazy -> niektore instance sa opakuju. Instance, ktore neboli vybrane do trenovacej mnoziny idu do testovacej

## 5. Opiste SVM algoritmus

SVM klasifikátor rozdeľuje trénovacie vstupy hyperplochou optimálne, čiže tak, že body dvoch tried ležia vždy na opačnej strane hyperplochy a vzdialenosť hyperplochy od bodov je čo najväčšia.



SVM pracuje na princípe prevedenia dimenzie vektora do vyššej dimenzie, kde je veľmi často možné problém lineárne separovať hyperplochou. Prevod vektora na inú dimenziu využíva SVM takzvanú kernel funkciu. Používané kernel funkcie sú:

- lineárny,
- polynomiálny,
- gaussian (RBF),
- sigmoid.

Nevýhoda tohto klasifikátora je schopnosť klasifikovať iba do dvoch tried. Je preto nutné klasifikátory reťaziť.

Algoritmus urci hyperplochu, ktorá zabezpeci optimalne oddelenie dvoch tried vstupnych vzorov, ktore su LINEARNE separovatelne. /+1

Vzory premieta do n-dimenzionalneho priestoru co ho robi neskutocne narocnym na vypocet. Prave preto sa vyuzivaju kernelove metody, ktore ako tak vedia pracovat v n-rozmernych priestoroch

(J) Pozn: premietanie do viacrozmerneho priestoru sa využíva, ak data nie sú lineárne separovateľné (toto premietanie ich takými spraví). Lenže tým sa zvýši dimenzionalita a vzrastá zložitosť. Funkcia pre výpočet však závisí iba od súčinu vektorov (tuším skalárneho //správne). Na to sú kernelové funkcie - umožňujú vykonávať tento skalárny súčin vo viacrozmerom priestore bez toho, aby bolo potrebné samotné premietanie, takže pracujeme v tom menej rozmerom.

## 6. Definujte siete maleho sveta, co je to a na ake ulohy sa používajú

trieda nahodnych grafov, kde vacsina uzlov nie su vzajomnymi susedmi, ale vacsina uzlov moze byt' dosiahnuta z ostatnych uzlov na maly pocet krokov  
prednaska 09

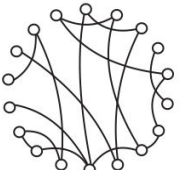
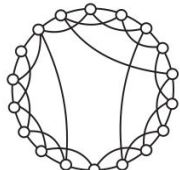
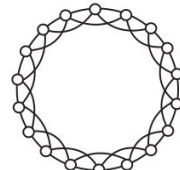
spôsob vytvorenia - preferencne pripojenie

novy vrchol (stranka) sa nespaja s existujucimi vrcholmi na webe podl'a rovnomerneho rozdelenia, ale s vacsou pravdepodobnost'ou sa bude odkazovat' na vrcholy s vysokymi stupnami (winners take all)

Používa sa na dolovanie štruktúry webu.

d - vzdialenosť medzi vrcholmi, je kratka lebo stranka sa pripoji cez preferovane uzly

C - klasterizacny koeficient, je vysoký lebo stranka bude odkazovat na jeho blizke okolie

	náhodné siete	siete malého sveta	mriežkové siete
$d$	krátka	krátka	dlhá
$\bar{C}$	malý	vysoký	vysoký
			

## Ďalšie dôležité pojmy

## 01. Inferencia 1R pravidiel

Pre všetky atributy

- Každé pravidlo zodpovedá prave jednej hodnote atributu
- Priradenie tej triedy, ktorá sa pre danú hodnotu vyskytuje najčastejšie

Alebo vypočítame error rate (1-accuracy) a vyberieme ten atribut, ktorý má najmenší error rate

Success rate = (1-error rate).

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

**2. Množstvo informácie (entropia)** -  $\text{Sum}(p_i \log(1/p_i))$

**3. Informačný zisk** -  $\text{gain} = \text{info}(D) - \text{SUM}(p(D_i) \text{info}(D_i))$

**4. Pruning** – orezavanie rozhodovacieho stromu

- v prípade, že nahradíme strom listom, sa minimálne zmení error rate – len vtedy to robíme

## 5. Zhľukovanie

- **Vzdialenosti (pre numerické data)** v zhľukoch určujeme cez : centroid, radius, diameter, medoid (vzdialenosť s ostatnými instanciami v zhľuku je minimálna), average //nema tu byť Euklidova a Manhatanska? // To by som dal sem isto ++
- **Single link** - najkratšia vzdialenosť medzi instanciami z dvoch zhľukov
- **Complete link** - najdlhšia vzdialenosť medzi instanciami z dvoch zhľukov
- **Vzdialenosti pre binárne data** – hamingova vzdialenosť, tanimotova miera rozmanitosti, jaccardova

•

**1. Hierarchické zhľukovanie** – aglomeratívny (zdola nahor), divízny – zhora nadol

- nie je potrebné určiť dopredu počet zhľukov
- **Dendrogram** – kompletný strom –
  - koreň – 1 zhľuk obsahujúci všetky instance
  - list – zhľuk obsahujúci 1 instanciu
  - vnútorné uzly – zhľuky vzniknuté spojením/rozdelением zhľukov

**2. Zhľukovanie rozdeľovaním** (partitional clustering)

- Je potrebné vedieť počet zhľukov (najmenšia kostra, k-means, PAM, CLARA)

**3. Pravdepodobnostné zhľukovanie - Zmiešaný model**

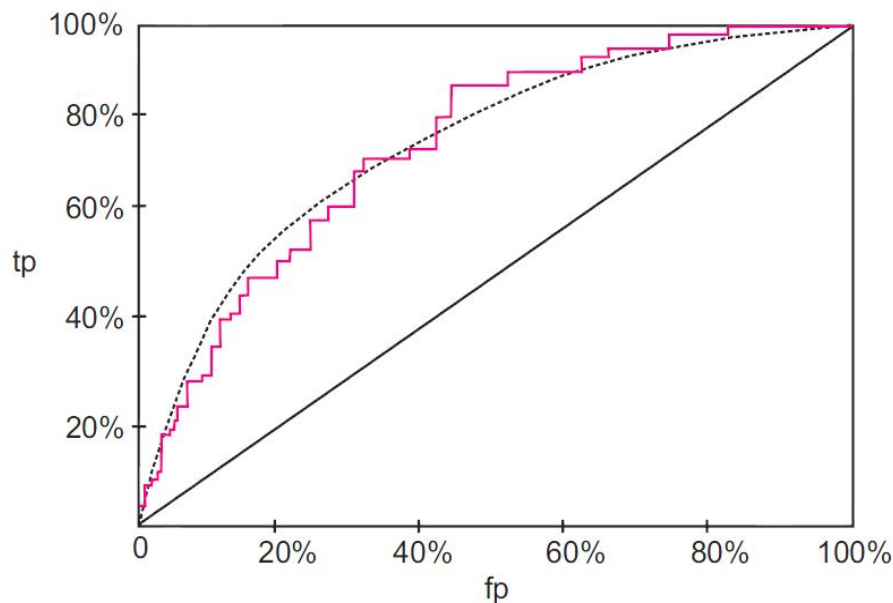
- cieľ je najst' najvierohodnejšiu množinu zhľukov pre dané data
- instance nie sú kategoricky priradené do zhľukov, ale prislusnosť do zhľukov vyjadruje pravdepodobnosť

## 6. Markovove modely (spracovanie sekvencnych dat)

- predikovať nasledujúce pozorovanie v čase na základe predoslych hodnôt
- predpoklad, že budúce predikcie sú závislé len na X posledných pozorovaniach – markov model x-teho radu

## 7. Co je to ROC krivka a aký je rozdiel oproti krivke navysenia (lift chart)

- ROC krivka charakterizuje výkon binárneho klasifikačného systému. Je to 2D graf, ktorého osi sú FPR a TPR (false/true positive rate). V prípade, ak by ROC graf mal tvar  $x = y$  - binárny klasifikátor by vôbec nefungoval (lepšie povedané by sa nelisil od náhodného klasifikátora). Čím je krivka viac natiahnutá smerom k ľavému hornému rohu, tým je výkon binárneho klasifikátora lepší. Oplatí sa spomenúť aj AUC (area under curve) - čím je táto plocha pod krivkou väčšia, tým je klasifikačný model kvalitnejší.



**Krivka navysenia - Lift chart** vyjadruje ako dobre predikčný systém pracuje na zvolenej množine dát - napríklad ak je našim cieľom najst tých, ktorých bude zaujímať reklamná kampan, tak aké percento pokrytia pozitívnych hodnôt bude vzhľadom na veľkosť vstupných dát.

# Doplnenie zo skúšky 2012

## 1. Stručne vysvetlite výhody (silné stránky):

- **k-means zhlukovania**
  - jednoduchý iteratívny algoritmus



- Časovo rýchly a jednoduchý na implementáciu
- dopredu si vieme určiť počet zhlukov // +2 //musíme si určiť, je to výhoda? // nevýhoda
- Citlivý na inicializáciu - spustiť viackrát - nevýhoda
- konverguje k lokálnemu minimu // toto nie je výhoda +1 // čo to znamená? - čím konverguje?
- hierarchického zhlukovania
  - nie je potrebné dopredu určiť počet zhlukov

2. Rovnako ako príklad s operatorom a zákazníkmi avšak s dokumentami v redakcii, ktoré treba utriediť.

## Skúška OT 2013

1. množiny {modrá, červená, zelená, žltá} a {malá, stredná, veľká} ako pretransformovať do binárneho tvaru + dodržať usporiadanie ak treba, prepísať zadane 3 instance do binárneho tvaru (napr. {červená, malá})

Tieto hodnoty nie je potrebné normalizovať ani spraviť binning, keďže nie sú spojité.

Hodnoty nie sú ordinálne. // toto je bullshit

-> z prednášky 02\_preprocessing, slide 5:

ordinálne - je možné ich usporiadať

napr. horúci > teplý > vlažný > chladný > studený

napr. starý > v stredných rokoch > mladý

Takže {malá, stredná, veľká}: veľká > stredná > malá

Ďalej podľa slidy 72 robíme transformáciu:

veľká ... 1

stredná ... 2

malá ... 3

vytvoríme atribúty s1, s2, s3, vďaka ktorým zachováme usporiadanie: // MATE TO ZLE n

kategorických hodnôt sa mapuje na n-1 binárnych atribútov čiže 00 10 a 11 zbytočný atribút je tam navyše

	s1	s2
veľká	0	0
stredná	1	0

mala            1        1

//+1

Potom by som pre prvý atribút rezervoval 2byty a namapoval by som ich: // takisto bullshit

	f1	f2	f3
modra	0	0	0
cervena	1	0	0
zelena	1	1	0
zlta	1	1	1

Pre instanciu {cervena, mala} by mapovanie vyzeralo takto:

100 11 // toto je blbost, nie? Pre kategorický atribút bez poradia (farba) robím N binárnych atribútov a pre diskretný/ordinalný (veľkosť) zásu N-1 -> t. j. malo by to byť niečo typu 0100 11

//Odkiaľ to máš že na N? Nikde to nevidím v prednáške // TU je to pekne zhrnuté

//+5 // Mohol by to sem teda niekto napísať správne?// pre každý kategorický atribút urobíš zvlášť stĺpec (napr. teda farbu zaznamu rozlíšiš pozíciou jednotky - ostatné farby budú vždy 0) a pre skupinu tých ordinalných urobíš N-1 stĺpcov a rozlíšiš ich počtom jednotiek (lebo to ti zachová pomer/poradie)

**Takto je to správne**

	s1	s2
velka	0	0
stredna	1	0
mala	1	1

Takto zakodovať to je správne podľa prednášky /ved je!

	f1	f2	f3	f4
modra	1	0	0	0
cervena	0	1	0	0
zelena	0	0	1	0
zlta	0	0	0	1

Nech to máme správne :) /ok

Takže (cervena,mala) by bolo 0100 11 ano? / ano

**2. data s veľa atribútmi, aký klasifikátor by ste použili aby ste nemuseli v predspracovaní odoberať atribúty a prečo?**

Mohol by to byť rozhodovací strom, lebo mu nezáleží na počte atribútov. Sam rozhodne ktoré atribúty sú dôležité (information gain?). V prípade že vznikne zložitý strom treba ho orezať (pre

prunning, post pruning ?). Prilis zlozity strom moze byt preuceny a to nechceme. Taktiez je narocne ho pochopit. /+1

//nie radšej SVM? SVM je známe tým že sa jednoducho vysporadúva s veľkou dimenzionalitou

### 3. vysvetliť (najmä rozdiely): a, klasifikácia b, zhľukovanie, c, predikcia d, regresia

**Klasifikácia:** Učenie s učiteľom, vieme do akých tried chceme instance umiestniť.

**Zhľukovanie:** Učenie bez učiteľa, nepoznáme triedy. Instance priradujeme do tried (zhľukov) na základe podobnosti hodnôt ich atribútov.

**Predikcia:** Na základe predošlých meraní (vykonávaných pravidelne napr. denne) hodnôt atribútov odhadujeme budúce hodnoty atribútov. (môže ísť o klasifikáciu alebo regresiu)

**Regresia:** Učenie sa funkcie zobrazovania. Data sa zobrazujú nejakou funkciou známeho typu.

//aký je rozdiel medzi predikciou a analýzou časových radov ?

// IMO: analýza časových radov patrí do predikcie, je to jedna z metód predikcie

// Analýza časových radov opisuje ich vlastnosti (napr. stabilitu, priemer) a vytvára z týchto vlastností nejaký model. Predikcia časových radov z tohto modelu odhaduje ďalší prvok v rade.

### 4. ROC krivka:

a, čo musí splňať klasifikátor aby sa dal použiť?

klasifikátor musí byť binárny

// toto vraj nestaci, treba este doplnit nejaku podmienku! // AKU ??

//nemôže to byť toto s vahami???? // nie, toto nie je podmienka, ktorú musí spĺňať, je to iba vysvetlenie tej krivky, to môžete použiť ako odpoveď na otázku “b) čo vyjadruje jeden bod na ROC krivke?”

// Nesmie mať tvar  $x=y$ , inak by nefungoval (bol by to vlastne náhodný klasifikátor) // +2

//The classifier or diagnosis result can be a real value (continuous output), in which case the classifier boundary between classes must be determined by a threshold value (for instance, to determine whether a person has hypertension based on a blood pressure measure). Or it can be a discrete class label, indicating one of the classes.

## ROC krivka

- z angl. receiver operating characteristic (detekcia signálu)
- podobne ako lift chart
- vertikálna os: počet pozitívnych inštancií obsiahnutých vo vzorke ako percento zo všetkých pozitívnych inštancií
- horizontálna os: počet negatívnych inštancií obsiahnutých vo vzorke ako percento zo všetkých negatívnych inštancií
- používa pravdepodobnosti (váhy) – meníme prah, pri ktorom bude výsledok klasifikácie interpretovaný ako "áno"
- citlivé na výber testovacej množiny  $\Rightarrow$  priemer cez behy cross-validation

### b, čo vyjadruje jeden bod na ROC krivke?

vyjadruje pomer medzi false positive rate a true positive rate

// čo to je to false/true positive?

// false positive - klasifikoval si ako positive pritom to je negative (cize chyba)

// false positive rate - (false positive) / (false positive + true negative) FPR

// true positive - klasifikoval si ako positive a naozaj to aj je positive

// true positive rate - (true positive) / (true positive + false negative) TPR

// tieto vypocty vyssie nie su zrejme spravne - podla paperu

<http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf> sa to predeluje FP/ all negatives a TP/all positives  
taktiez je to tak aj v tomto velmi dobrom videu <https://www.youtube.com/watch?v=OA16eAyP-yo>

### c, ak by sme mali ROC krivku pod diagonalou čo by to znamenalo?

binarny klasifikator by nefungoval, nelisil by sa od nahodneho klasifikatora

// nahodny klasifikator je ked ROC krivka lezi **na** diagonale, podla mna ked je **pod** tak sa nejako vymenili navzajom triedy /osi FPR  $\leftrightarrow$  TPR

<http://www.navan.name/roc/> tu sa s tym da hrat :D ked je to v strede tak triedy spolu splynuli,  
pod ciarou by to zrejme islo ak by sa triedy vymenili?

//klasifikator je horsi ako nahodny, riesenie->vymenit vysledky navzajom

Zdroj: <http://www0.cs.ucl.ac.uk/staff/ucacbb/roc/>

### 5. Który algorytmus zhlukowania by ste použili, keď dopredu neviete výsledný počet zhlukov a prečo? Ako by ste výsledky vyhodnocovali?

hierarchické algoritmy, zhora nadol a zdola nahor

Tieto algoritmy nevyžadujú na vstupe počet zhlukov. Rozdelia celú množinu instancií na jeden zhluk. Tento zhluk delia na ďalšie a ďalšie podľa atributov.

Výsledok by som zobrazil dendrogramom, čo je stromová štruktúra. Na vyhodnotenie by som použil interne alebo externe kritéria. Pravdepodobne indexy.

### 6. Vysvetliť Page Rank.

//vysvetlene vyššie

### 7. Zadane množiny A,B,C a ich prieniky - hodnoty (A=10, B=15, ..., A a B = 15, A a B a C = 15), zadane 3 pravidlá A a B => C, B a C => A, C => A, pre min. s = 0.25 a alfa = 0.75, určte či alg. APRIORI odhalí zadane pravidlá.

$x = [a, b]$

$y = [c]$

$s(x, y) = X \cup Y$  / počet všetkých transakcií

$x \cup y = (a \text{ prienik } b) \cup c = 15 + \dots$

n - počet všetkých transakcií

$\alpha = (x \cup y) / x$

// Nerozumiem... čo sa tým zistilo ani nič

Podľa toho čo som pochopil je to takto: Aby algoritmus odhalil pravidlá musí spĺňať aj min Confidence (alpha) aj min Support(s) ale z daných čísiel si to nevieme určiť lebo nepoznáme C a teda ani celkový počet. Domyslíme si teda C = 20 z čoho celkový počet vychádza na 45. Potom:

Support (A a B => C) =  $15/45 = 0.33$  -> čo je väčšie ako 0.25 -> spĺňa support

$\alpha(A \text{ a } B \Rightarrow C) = 15/15 = 1$  -> je väčšie ako 0.75 -> confidence spĺňa

Vychádzame z definície confidence: podiel tých instancií, ktoré vyhovujú pravidlu (A a B a C = 15) oproti tým, na ktoré sa dá pravidlo aplikovať (A a B = 15)

Support(B a C => A) - nevieme vypočítať lebo nemáme dosť dát na to

$\alpha(B \text{ a } C \Rightarrow A)$  - nevieme vypočítať lebo nemáme dosť dát na to

Support(C=>A) =  $20/45 = 0.44$  -> support vyhovuje

$\alpha(C \Rightarrow A)$  - nevieme vypočítať lebo chýbajú údaje

### 8. Bayes - príklad zdravy, alergia, prechladnutie (2007)

//vysvetlene vyssie

9. Nový sefredaktor v spoločnosti XYZ. Jeden priecinok so všetkými článkami. Vy ako expert v objavovaní znalostí roztriedte články bla bla ....  
+ a,...,f otázky ako pri ostatných ulohách tohoto typu

## Skúška RT 2013

Farbou **farba** označujem text, ktorý som domyslel ak máte lepší návrh tak ho napíšte.

1. (4 body) Atribút vzdialenosť môže nadobúdať reálne hodnoty z intervalu  $<0;1000>$ . Klasifikátor, ktorý chcete použiť však potrebuje, aby všetky atribúty boli binárne. Vysvetlite, ako pretransformovať tento atribút. Použite všeobecný postup s ohľadom na možnosť voliť výsledný počet atribútov a potrebou zachovať usporiadanie hodnôt.

- a. Discretization: reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.
- b. Binning methods – equal-width, equal-frequency → toto sú metódy bez učiteľa, t.j. že nemáme info o tom kam aká hodnota patrí vopred
- c. Riešenie:
  1. Využijeme rovnako veľké intervaly:
    - a. Najdeme najväčšiu a najmenšiu hodnotu atribútu
      - i. dosadíme do vzorca  $z = \min - \max/k$ , kde  $k$  je počet tried ktoré vzniknú, my máme len dve triedy (ma to byť binárne).
    - b.  $z$  je hranica medzi triedami napr. ak  $z = 25$  tak hodnoty menšie ako 25 sú 0 a väčšie ako 25 sú 1

// podobné ako [2007/1](#)

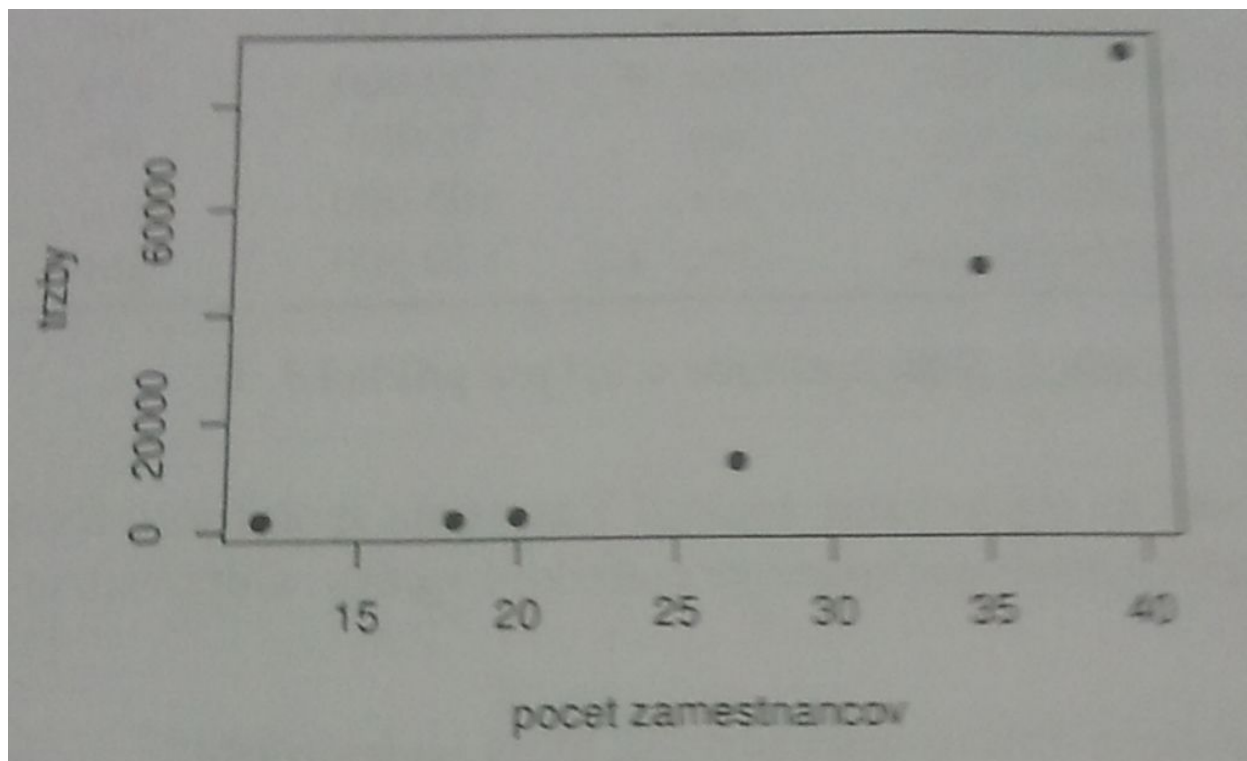
2. (6 bodov) Akú techniku **vyhodnocovania** zhľukovania by ste použili na opis charakteristík jednotlivých zhľukov? Vysvetlite prečo. Aké charakteristiky zhľukov takto získate? (Bez vysvetlenia a opisov charakteristík odpoveď nebude uznaná.)

// komentár 2015 - tu je asi odpoveď Silhouettova

([https://en.wikipedia.org/wiki/Silhouette\\_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))) validácia metóda, lebo meria aj ako pevne sú zoskupené instance v zhľukoch aj celkovo ako dobre je dátová množina rozklastrovaná

3. (6 bodov) Štatistický úrad zbiera informácie o rôznych veľkých podnikoch (veľkosť je meraná počtom zamestnancov). Náhodným výberom sa vyberie množina podnikov, ktoré poskytnú údaje o tržbách. Príklad získaných dát je na obrázku 1. Ako by ste dopočítali

(odhadli) tržby pre spoločnosti s 30 a 50 zamestnancami? Vysvetlite akú úlohu objavovania znalostí by ste riešili a akú metódu by ste zvolili.



Obrázok 1: Tržby vybraných spoločností v závislosti od počtu zamestnancov

Použil by som predikciu na základe regresie. Regresia by mi objavila funkciu na zobrazenie dát a na základe nej by som mohol dopocitať ďalšie hodnoty. Použil by som polynomiálnu regresiu.

Nedopocítané hodnoty by som mohol vypočítať aj interpoláciou.

// Tuto by sa asi dalo, že vidíme že to pripomína nejakú funkciu toto konkrétne  $e^x$ , a podľa toho vieme potom dopočítať chýbajúce hodnoty.// do not drink after 10pm, vycikat a spat // :D Nemáš mi to :((((((

4. (6 bodov) V akom ohľade sú miery pokrytie a spoľahlivosť nepostačujúce? Aké nevhodné typy pravidiel môžeme pomocou dolovania asosiačných pravidiel získať, keď budeme používať len tieto dve miery?

// podobné ako [2007/12](#)

// tie nevhodné typy pravidiel niekto nasiať?

// pokrytie = podpora(support)?

K tým nevhodným typom pravidiel, ako je spomenuté v 2007, support neberie do úvahy pravdepodobnosť pravej strany pravidla, teda  $P(Y)$  v pravidle  $X \Rightarrow Y$ . Ak má  $Y$  veľmi

vysokú pravdepodobnosť napr. 1, tak sa môže objaviť na pravej strane mnohých pravidiel, aj keď by tam ani nemusel byť.

5. (6 bodov) Uvažujte klasifikačný problém na množine dát, v ktorej sú rádovo viac zastúpené pozitívne ako negatívne inštancie a vzhľadom na celkové veľké množstvo inštancií je rozumné použiť vzorkovanie (angl. undersampling) z pozitívnych inštancií. Vysvetlite, ako budete porovnávať presnosť **daných/dvoch/niečo na d** klasifikátorov pre takéto dáta. Ako výsledky vyhodnotíte? Vysvetlite všetky dôležité kroky v **proces**e porovnania.

??Stratifikovaná?? Cross validácia ?

// f-measure kde FP je penalizované z inou váhou ako FN?

// čo je teda odpoveď k tejto otázke ? vie niekto ?

//nemohlo by sa použiť aj špecificita?

// A nemôže tu byť TP, FP -> confusion matrix, a vyhodnotenie pomocou Precision a Recall? A v prípade zlej klasifikácie by sa spravila cross validácia.

// prednáška 8 ? aby sa spravil t-test ? keďže ide o klasifikačný problém



- Klasifikácia
  - porovnávanie metód – t-test
  - počítanie ceny
  - lift charts, ROC, error, cost krivky
- Numerická predikcia
  - stredná kvadratická chyba, stredná absolútna chyba
  - relatívna kvadratická chyba, relatívna absolútna chyba
  - korelačný koeficient
- Zhľukovanie
  - MDL
  - indexy: Dunnov, Davies-Bouldinov
- Asociačné pravidlá
  - support, confidence
  - interest, conviction

6. (5 bodov) Mnohé algoritmy na zhľukovanie aj klasifikáciu počítajú vzdialenosti medzi jednotlivými inštanciami. Pre rôzne typy dát sa hodia rôzne spôsoby počítania vzdialeností.

(a) Aké vzdialenostné miery by ste použili pre tieto typy dát:

i. binárne dáta

hamingova vzdialenosť, tanimotova miera rozmanitosti, počet bitov v ktorých sa líšia //+1

ii. textové dáta

- vzdialenosti v texte by som meral na základe vzdialenosti slov, slova na základe vzdialenosti jednotlivých znakov

- alebo vzdialenosť na základe vzdialenosti hashov slov // dobrá bľbosť zmena jedného znaku predsa v hashi spôsobí obrovský rozdiel takže tie hashe by neboli ani len trochu podobné, takýmto princípom by si mohol matchnúť úplne ine slova



- **Levenshtein distance** = Edit distance - toto tu neuznavala, tusim chcela cosine similarity

**iii. numerické dáta**

euklidova, manhattan, cosinus distance //+1

**(b) Vysvetlite, čo by ste urobili, ak by ste mali k dispozícii dáta zmiešané napr. niektoré atribúty binárne, niektoré textové niektoré numerické.**

- všetky data by som previedol na spoločny formát, napríklad na binárne .... zároveň by som data normalizoval a použil binning tam kde je to potrebné

**7. (4 body) Vysvetlite ako pracuje algoritmus podporných vektorov SVM (support vector machine). Na aké úlohy je ho možné použiť?**

// [2011/5](#)

8. (6 bodov) Uvažujte dáta z tabuľky 1 a použitie naivného Bayesovho klasifikátora. Vyjadrite v tvare číselného výrazu (nie je potrebné vypočítať presné číslo):

(a) pravdepodobnosť, že keď máme záujem o byt, je to rekonštruovaný byt v Dúbravke

(b) pravdepodobnosť, že keď ide o rekonštruovaný byt v Dúbravke, tak o **neho máme** záujem

c. vzorky	lokalita	rekonštruovaný	cena	záujem
1	Dúbravka	áno	80 000	áno
2	Dúbravka	nie	85 000	nie
3	Dúbravka	nie	80 000	áno
4	Karlova Ves	nie	90 000	nie
5	Karlova Ves	nie	128 000	áno
6	Karlova Ves	áno	125 000	nie
7	Karlova Ves	áno	130 000	áno
8	Karlova Ves	nie	76 000	nie
9	Dúbravka	nie	105 000	nie
10	Dúbravka	áno	110 000	áno

Tabuľka 1: Dáta o záujme o byt pre príklad č. 8

a)

Seky: Vytvoríme si pomocnou tabuľku ako v prednaske 03\_klasifikacia slide 46.

	lokalita				rekonst rukcia			zaujem ?	
zaujem	ano	nie			ano	nie		ano	nie
V stĺpci vypisujem hodnoty lokality:				V stĺpci vypisem hodnoty rekonstru kcie:					
dubravka	3	2		ano	3	1		5	5
karlovka	2	3		nie	2	4			

dubravka	3/5	2/5		ano	3/5	1/5		5/10	5/10
karlovka	2/5	3/5		nie	2/5	4/5			

Pre lavu cast hladame lokalita x zaujem. V strednej casti rekonstrukcia x zaujem. V pravej casti len pocitame riadky zaujem. V dolnej casti tabulky cislo predelime suctom stlpca.

dubravka + zaujem = 3 \\\\\\\\\\\\\\\\\\\

dubravka + nezaujem = 2

Inak povedane spocitaj riadky ktore maju uvedenu ako lokalitu dubravka a nie je o nich zaujem.  
// pls vysvetlite niekto hodnoty rekonstrukcie, ako ich dostaneme (vysvetlit podobne ako lokalitu :)

Spocitaj riadky je kde rekonstrukcia ANO a je to ZAUJEM, potom ANO a nie je ZAUJEM.V dalsom riadku su nerekonstruovane a je zaujem. Nerekonstruovane a nie je zaujem.

$d_{n+1}$  je instancia, kt. ma atributy {Dubravka, zrekonstruovany}

$$P(d_{n+1} | \text{mam\_zaujem}) = P(\text{dubravka} | \text{mam\_zaujem}) \cdot P(\text{rekonstruovany} | \text{mam\_zaujem}) \\ = \frac{3}{5} \cdot \frac{3}{5} = \frac{9}{25}$$

Takto je to sto percent spravne a tie vypocty co tu boli predtym boli sprostosti.

// tak potom je zle aj to bcko, ale nema sa tu pouzit ta podmienena pravdepodobnost  $P(H|d)$  ?

//nie, to sa pouzije az v B. toto je jeden z vyrazov, ktory sa prave v B bude pouzivate pri vypocte. Ak mas vyraz  $P(\text{instancia} | \text{trieda})$ , dokazes to vyratat takymto sucinom. Je to v prednaske slide 42.

b)

// pls ako ste vyratali to b?

$$P(\text{mam\_zaujem} | d_{n+1}) = P(d_{n+1} | \text{mam\_zaujem}) \cdot P(\text{mam\_zaujem}) / P(d_{n+1})$$

Z tohto jednotlivé výrazy:

- $P(d_{n+1} | \text{mam\_zaujem}) = 9/25$  (uloha A)
- $P(\text{mam\_zaujem}) = 5/10$  (5 z 10 mam zaujem)
- $P(d_{n+1}) =$   
 $P(d_{n+1} | \text{mam\_zaujem}) \cdot P(\text{mam\_zaujem}) + P(d_{n+1} | \text{nemam\_zaujem}) \cdot P(\text{nemam\_zaujem}) =$   
niektore hodnoty pozname, tak ich rovno pisem  
 $9/25 \cdot 5/10 + P(\text{dubravka} | \text{nemam\_zaujem}) \cdot P(\text{rekonstruovany} | \text{nemam\_zaujem}) \cdot 5/10 =$   
 $9/25 \cdot 5/10 + \frac{2}{5} \cdot \frac{5}{10} = 9/50 + 2/50 = 11/50$

Dokazeme teda vyratať pociatočný vzorec

$$P(\text{mam\_zaujem} | d_{n+1}) = 18/50 \cdot 1/2 / 11/50 = 9/50 / 11/50 = 9/11 \text{ //+3 // } 9/20$$

**9. (7 bodov)** Ako odborníka na objavovanie znalostí Vás najala spoločnosť **Bcereálie** aby ste pomohli jej zamestnancom automaticky detekovať spamové e-mailové správy.

Odpovedzte na tieto otázky a svoje odpovede zdôvodnite.

**(a) Ako definujete typ úlohy objavovania znalostí ktorý budete riešiť?**

klasifikácia, cieľom je vytvoriť klasifikátor

**(b) Ako budete pripravovať dáta, čo potrebujete od spoločnosti Bcereálie a jej zamestnancov ?**

Najprv by som zachytával všetky potrebné dáta o emailoch aké môžem. Potom pri príprave by som si zvolil, ktoré atribúty budem používať pre klasifikáciu. Napríklad IP adresy, porty, odosielateľa, príjemateľa, kľúčové slová z obsahu mailu.

**(c) Ako dáta reprezentujete?**

Dáta by som reprezentoval objektom, OOP princíp. Objekt Email by mal všetky atribúty, ktoré potrebujeme. Čiže štruktúrované.

**(d) Aký typ predspracovania navrhujete použiť?**

Kontrola formátu údajov, doplnenie null hodnôt. Vytvorenie histogramov. Získanie kľúčových slov z obsahu emailu. Rozdelenie dát na testovacie a tréningové. Možno boxplot a scatter plot na numerické hodnoty.

// MD: neviem si predstaviť žiadnu num. hodnotu ktorá by bola relevantná na to (čas? chodí len poobede? IP adresa?)

**(e) Aké DM techniky by ste použili? Odpovedajte na úrovni konkrétnych algoritmov**

Na vytvorenie klasifikátora Naive Bayes. // pretože klasifikujeme na základe nezávislých atribútov? // **Prečo tu nemôže byť aj Decision Tree s Adaboostom, alebo SVM atď..? // podľa mňa môžeš ale potom treba aj predspracovanie podľa DM techniky popísať**

**(f) Aký typ znalostí očakávate ako výsledok zvolených DM metód?**

Výsledkom bude natrenovaný klasifikátor. Tento klasifikátor nám v budúcnosti pomôže klasifikovať nový email či ide o spam.

// To je znalosť? nie skor ze znalosť či ide o spam alebo nie? jkk je

**(g) Ako by ste overovali svoje výsledky?**

Cross validáciou, možno aj ROC krivka by sa dala použiť

## Skuska RT 2015

Na skuske bolo veľa minuloročných :

2013OT - otázka 1

a niekoľko nových:

**OTAZKA: Akú techniku zhľukovania by ste použili pre data, pri ktorých neocakavate, že výsledné zhľuky budú konvexné (pravidelné). Ako by ste vyhodnotili kvalitu takehoto zhľukovania.**

Dbscan

Cez indexy interne a externe

**OTAZKA: Vysvetlite pojmy: klasifikácia, zhľukovanie, regresia, stratifikácia**

**OTAZKA: Otázka koncipovaná podobne ako RT 2013 s Včeraiami (tie podulohy), ale zadanie bolo iné: Islo o cieľenu reklamu pre videopozicovnu, kedy chceme zákazníkovi na základe jeho vypožičok odporučiť ďalšie filmy**

**OTAZKA: Aké meranie vzdialenosti by ste použili pre binárne data a odovodnite PRECO**

// tu chcela naozaj hlavne ten dôvod, porovnaním s nejakými inými meraniami a napísanie prečo práve toto meranie je výhodné použiť.

Hamingova vzdialenosť - pretože tu sa počíta len počet bitov, o koľko sa daný reťazec líši od druhého a je to rýchle.

Tanimotova miera rozmanitosti - je použiteľná iba pre binárne dáta, a jeho hodnoty (pri použití na dátach) sú v blízkosti priemerných hodnôt oproti Manhattan alebo Euclidean distance, kde nie sú.