

rekáža zľahčovanie toho, že predseda parlame
Povinne precitat: Toto je súbor otazok, ktoré dáva Zendulka na svojom predmete na FIT VUT. Obsah a osnova predmetu su velmi podobne, preto môžeme očakávať, že aj nám dá podobné otázky. Treba na ne postupne vypracovať odpovede, prípadne ak nilineaekto nájde otázku, ktorú sme nepreberali tak ju zmažte alebo to tam vyznacte. Vela otazok sa opakuje. Zatiaľ sme to nestihli vsetko porovnat iba sme sem dali otazky. Idealny stav : na konci by bol dokument kde je iba zoznam otazok s odpovedami. A skuste to nerozjebat.

DO BOJA SLOVENSKÍ ŠTUDENTI - Musíme dokázať, že prípravu zvládneme ako naši kolegovia v Brne



// ten pocit ked si neboli ani na 1 prednaske z OZ a nevies na ktorom z vyssie uvedenych obrazkov je prednasajuci :/ // prvý (hore napravo) je Zendulka (+1), vedla neho je Bartik nebudete lepsie vymazat tie otazky co sa opakuju nech su tu iba raz ?
:/ :/ :/ :/ :/ :/ ako to vidite s IAU? :)

-
https://docs.google.com/document/d/14Z3Z4sSP1qH_UZ3XL9lbVnUZHAmAVPvntkzhuYEvf

oo/edit?fbclid=IwAR3nzV94bnji62PWBpNQf9PV_AVCAC5kjJQ4VMvu8IAzP8PyYfTLUvDg8
6s#heading=h.l668w9fmfy9z

RT 2018 Inteligentna analyza udajov

https://docs.google.com/document/d/1iaFNDnbBl9beMPb5441lGVtio05sUpwN4_Nu35wIdKw/edit

skúška RT OZNAL 2017- Co tak to tu vypracovat?

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 // je to v doku arit. priemer, modus, median
 - b. 3 Metriky rozptylu, výhody a nevýhody // rozptyl, rozsah, kvartily, odlahle hodnoty, krabicove diagramy
 - c. Opíš krabicový graf, aké informácie v ňom sú, na čo sa dá použiť pri predspracovaní dát // je to dole
2. Zhlukovanie
 - a. podla rozdelovania (partition) výstižne opísať, uviesť jeden príklad na takúto metódu: opísať algoritmus, výhody, nevýhody
 - b. Podľa hustoty, to isté + porovnat s a. (teda podla hustoty vs. podla rozdelenia)
 - c. Metriky porovnania zhlukovania, 1 opísať na čom je založená
3. Klasifikácia
 - a. Opísať princíp, všeobecné kroky klasifikácie
 - b. opisat + jeho výhody a nevýhody
 - c. k-najblízšich susedov
 - d. opisat Naive Bayes a ako suvisí s podmienenou pravdepodobnosťou
4. Otázky okolo dolovania z webu
 - a. Opísať PageRanking a HITS
 - b. Opísať aké data ziskavame z analýzy struktury a použitia webu, na čo sa podobajú, opisat aké sú najcastejšie úlohy (asociacne, ...) // aké úlohy tu chcu pocut?
 - c. Kroky získavania dát použitia // predpsracovanie, ziskavanie znalostí, analiza vzorov
5. Asociačné pravidlá
 - a. Definuj frekv. Množinu a podporu vo frekv. množine
 - b. Nevýhody apriori a ako ich rieši FP strom
 - c. Definuj výpočet spoľahlivosti
 - d. Ziskanie asociačnych pravidiel z frekventovanej množiny

Nahradny RT OZNAL

1. Dolovanie frekventovanych grafov(nie mnozin), 2 metody a 1 priklad pouzitia // toto niekto ?
2. Co je to sekvencia
3. Ake typy atributov pri dolovani pozname
4. Metoda Apriori, popis výhody nevýhody postup
5. Dolovanie na webe
 - a. porovnat www vs textove databazy
 - b. co ma význam sledovať na webe z pohľadu dolovania dat
 - c. rozdelenie web mining (content usage structural + popisat)
6. Rozhodovacie stromy
 - a. Ako funguju
 - b. Blízsie vysvetlenie ako sa vytvara z dat rozhodovaci strom (Gain(), Info(), ...)
 - c. Porovnanie ID3, C4.5 a Gini index z pohľadu vyberu atributov pri vytvaraní stromu
7. Metriky určene na výhodnocovanie asociáčnych pravidiel // toto ?možno toto

Metriky:

$$s(A \Rightarrow B) = P(A \cup B) - \text{podpora}$$

$$c(A \Rightarrow B) = P(B|A) - \text{spolahlivosť}$$

// to som len tak narychlo spisal, struktura skusky bola rovnaka ako na RT (5 uloh, v kazdej abcd, ale casto sa miesala tematika otazok takze je to jedno)

// ked sa na to tak vseobecne pozeram tak si myslim, ze na OT vymysli nejake uplne nove otazky... co myslite vy ? //Urcite

Opravny termin 2017

1. Cistenie dat (neuplnosť, zasumenie, nekonzistencia, duplicity) plus vybrať dve a napisat spôsob riešenia
2. Co je to nekonzistencia a ako sa riesi
3. Znízovanie dimenzionality dat
4. Sekvencia a sekvenčný vzor...
5. ...
6. Viacurovnové asociáčné pravidla - vysvetliť, príklad tabuľky
7. Jednotkový neuron, nakresliť, popisať
8. Neuronová sieť - Backpropagation
9. výhody nevýhody neuronových sieti
10. stemovanie, stop-word list
11. tf, idf, tfidf
- 12.

Otzky sria 1

1) Cistn dat

(prednaska 02_predspracovani_FIIT)

- opis

Cistn je - Doplnen chybjcich hodnot, vyhlazen zaumennych dat, identifikace nebo odstran odlehlch hodnot, rejeni nekonzistenc

- problmy

realne dta s: neuplne, zasumene, rozsiahle, nekonzistentne
(nizka kvalita dat = nizka kvalita vysledkov)

- metody

1. osetrenie chybajucich hodnôt

- ignorovanie zaznamu
- doplnenie chybajucich hodnôt ručne
- automatické doplnenie
 - § globálnou konštantou (napr. neznáma)
 - § priemernou hodnotou atribútu
 - § najpravdepodobnejšia hodnota – odvodenie založené na bayesovskom klasifikátore alebo na rozhodovacom strome

2. Metody odstranení šumu //podla toho na čo sa pytajú

Pln (binning) - nejprve se data uspořádají a rozdělí do tzv. košů (stejné „šířky“ nebo „hloubky“) následně se data v koši vyhledí průměrem koše, mediánem koše, bližší hraniční hodnotou a pod.

Regres - vyhlazení vyrovnaním dat podle regresních funkcí

Shlukování - detekce a odstranení odlehlých hodnot

Kombinovaná kontrola počítacem a človkem - automatizovaná detekce podezřelých hodnot a kontrola človkem (např. potenciální odlehlé hodnoty)

2) Denclue

(prednáška 08_09_shlukovani_fiit slajd 58+)

- co to je

- DENsity-based CLUstEring
- zhlukovacia metoda založená na hustote
- **Centra shlukù**: místa v prostoru, kde se nacházejí lokální maxima celkové funkce hustoty
- **Funkce vlivu**: libovolná funkce odvozená od vzdálenostní funkce (např. Euklidovská vzdálenost)

- problmy

nastavení parametru σ

- parametry

// prednaskach som nenesiel, konkretnie sa opisuje iba metoda DBSCAN

// žeby σ ? máš to vyššie napísané // to je parameter lambda alebo ako sa vola // sigma

3) Rozhodovací strom

(03_05_klasifikace_fiit.pdf slide 9+)

- co to je

- Je to graf stromovej štruktúry.
- Vnútorný uzol - test hodnoty istého atribútu (student? = áno).
- Listy stromu - reprezentujú triedy, do ktorých bude daný objekt klasifikovaný.
- Atribúty musia mať diskrétny charakter.

- algoritmus pre vytvorenie

```
function CreateTree( $S, L$ ): tTree; //  $S$ -množina vzorů,  $L$ -seznam atributů
begin
    • Vytvör nový uzel  $N$ 
    • if (vzorky  $S$  jsou ve stejné triedě  $C$ ) then return  $N$  jako list dané triedy  $C$ 
    • if (seznam atributů  $L$  je prázdný) then return  $N$  ako list nejběžnější
        triedy ve množině vzorků  $S$ 
    • Vyber „vhodný“ atribut  $A$  ze seznamu  $L$  a odstraň jej z tohoto seznamu
    • Pojmenuj uzel  $N$  jménem vybraného atributu  $A$ 
    • for (každou možnou hodnotu  $a_i$  atributu  $A$ ) do
        Vytvoř větev z uzlu  $N$  pro podmínce „ $A = a_i$ “
        Nechť  $s_i$  je podmnožina vzorků z  $S$ , u nichž „ $A = a_i$ “
        if ( $s_i$  je prázdná) then připoj k věti list s nejběžnější triedou v
            množině  $S$ 
        else připoj k věti podstrom vzniklý rekurz. voláním
            CreateTree( $s_i, L$ )
    end
```

-ID3

- Algoritmus na generovanie rozhodovacieho stromu z datasetu
Vhodný atribút - atribút s **najväčšou rozlišovacou schopnosťou**

-jaké atributy upřednostňuje ID3

ID3 upřednostňuje atribúty A s najväčšou hodnotou Gain(A) alebo - ekvivalentne - najmenšou hodnotou $Info_A(S)$

-jak ID3 řeší spojité nediskretizované atributy

- zoradit hodnoty, vypočítat' pre všetky možné splitpoint $Info_a(S)$, vybrať to, ktorá má najmenší $Info_A(S)$

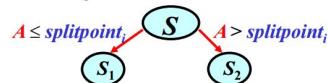
Diskretizace spojitéch hodnot 1/2

- Rozhodovací strom vyžaduje pouze atributy obsahující diskrétní hodnoty

- Nechť atribut A nabývá spojitéch hodnot
- Určení „vhodné“ dělící hodnoty (**split-point**) pro atribut A a následné rozdělení do dvou skupin:
 - Seřadíme hodnoty atributu A tak, aby tvořily rostoucí posloupnost: $a_1 < a_2 < \dots < a_i < a_{i+1} < \dots < a_v$
 - Postupně „vhodnou“ dělící hodnotu (**split-point**) pokládáme rovnou výrazu:
$$\text{splitpoint}_i = \frac{a_i + a_{i+1}}{2}$$

Diskretizace spojitéch hodnot 2/2

- 3) Postupně pro každou hodnotu splitpoint_i simulujeme vytvoření následujícího uzlu:



a vypočítáme hodnotu $\text{Info}_{A_i}(S)$

- 4) Vybereme tu hodnotu splitpoint_i , která má nejmenší hodnotu $\text{Info}_{A_i}(S)$.

Podle této hodnoty provedeme diskretizaci na dvě skupiny:

$$\begin{aligned} A &\leq \text{splitpoint}_i \\ A &> \text{splitpoint}_i \end{aligned}$$

2 základní metody pro ořezání stromu:

- Prepruning = již v průběhu vytváření stromu nejsou generovány větve, které mají malý význam pro rozhodování
- Postpruning = nejdříve vytvořen strom jako celek, teprve pak jsou větve s malým významem odstraněny
- Obě metody mají své výhody a nevýhody → mnohdy se ořezání řeší kombinací těchto metod

4) Dolování asociačních pravidel

Asociačné pravidlá sa generujú z frekventovaných množín.

Frekventovaná množina je množina položiek, ktorá má podporu vyššiu než je užívateľom zadaná hodnota.

Silné Asociacie Pravidla (asociacie pravidla) sa generujú z frekv. množín na základe minimálnej spoľahlivosti.

Na nájdenie frekv. množín sa najčastejšie používa algoritmus **Apriori**, ktorý stavia na podmienke, že každá frekventovaná množina môže byť frekventovaná, len ak sú všetky jej podmnožiny frekventované. Lepším riešením, je použiť algoritmus spolu s **FP stromom** (**Frequent-Pattern tree**), ktorý je kompaktný, kompletný a netrpí potrebou generovania obrovského množstva kandidátnych množín

5) Dolování v proudu dat

- co to je

Je to dolování z datového toku, který je spojitý, uspořádaný, proměnlivý, rychlý. Většinou nepotřebuji „přesnost“, obsahuje spíše skryté a obecnější vzory.

V proudech dat se dolují pouze přibližné frekventované vzory, protože získání přesných je nerealistické → přibližné postačí.

Podstata dolování: nesleduj položky, dokud se nestanou potenciálně frekventovanými. To má výhodu, že mám zaručenou ohraničenou chybu, ale nevýhodu, že musím uchovávat hodně informací. Např. alg. Lossy Counting

Typické úlohy: multidimenzionaální analýza, **dolování častých vzorů**, shluková analýza, klasifikace

Metody zpracování proudu dat: náhodné vzorkování, klouzavé okno, histogramy, skeče, randomizované algoritmy

- problémy

- Obrovský objem dat
- Data se rychle mění → potřebujeme odezvu v reálném čase
- Náhodný/přímý přístup k datům je drahý → jednopruhodové algoritmy
- Proud obsahuje nízkourovňová a vysoce dimenzionální data → musím je potom dále zpracovávat.

Nezmestia sa do ramky, prechádzanie dát môžeme väčšinou len raz (treba rýchle algoritmy), treba brať do úvahy, že staršie dáta môžu byť pre nás menej smerodajné

Oázky séria 2

1. Popište 2 bioinformatické problémy a jejich řešení + uveďte dolovací použité metody. (8b) toto by som cele vyhodil

(TOTO SME V PREDNASKACH NENASLI ale daco ma v knihe)

Dolovanie v sekvenciach DNA

Porovnanie nějaké sekvenči s celou databází sekvenčí. Pro řešení tohoto úkolu se používají nástroje BLAST nebo FASTA 1(případně jejich varianty). Vzhledem k počtu sekvenčí, které obvykle chceme porovnat, není možné využít běžného přístupu pro porovnávání sekvenčí. Místo toho se využívají nejrůznější indexy a heuristiky.

Problémy

Počet možných sekvenčních vzorů je obrovský.

Požadavky na dolovací algoritmus:

- je-li to možné, nalézt úplnou množinu vzorů splňujících podmínku minimální podpory,
- efektivnost, škálovatelnost, co nejmenší počet potřebných průchodů databází,
- schopnost zahrnout různé druhy uživatelem definovaných omezení.

Kategorie algoritmů

Generující kandidáty + využití Apriori vlastnosti (např.GSP—Generalized Sequential Pattern)

- Není-li sekvence S frekventovaná, pak není frekventovaná ani žádná její supersekvence.

Bez generování kandidátů (PrefixSPAN)

- Vytváří tzv. projektované databáze obsahující frekventované prefixy sekvencí postupně rostoucí délky. Ty představují sekvenční vzory.

2. Algoritmus k-means. Co to je, k čemu to je, vstupy, výstupy, jak pracuje, výhody, nevýhody. (9b)

- Zhlukovací algoritmus.
- **Vstupy** - k (požadovaný počet zhlukov), Dátový súbor D, ktorý obsahuje n objektov
- **Výstupy** - objekty s priradenými číslami zhlukov
- Algoritmus:
 - a. Vyberie K objektov ako centroidy
 - b. Každý neoznačený objekt priradí k najbližšiemu centroidu
 - c. Vypočítaj priemery jednotlivých zhlukov a nastav ich ako nové hodnoty centroidov
 - d. Opäť priradí jednotlivé objekty k najbližším centroidom
 - e. Opakuj kroky B - D pokým nie sú všetky objekty priradené k tým istým centroidom ako v minulej iterácii (alebo sa preradilo len malé množstvo (zadaná hodnota, napr. 5%) objektov k iným centroidom - nie je totiž zaručené, že algoritmus vždy skonverguje, častokrát nie)

```

Metoda k-means

Vstup: datový soubor D o n objektech,
        počet shluků k
Výstup: k shluků
        Zvol libovolných k objektů jako počáteční
        těžiště shluků;
repeat
    Přiřaď každý objekt do shluku s nejbližším
    těžištěm;
    Přepočítej těžiště (střed (mean)) shluků
    aktuálního rozdělení.
until nezměnilo se přiřazení žádného objektu;

```

- **Nevýhody:** každý objekt patří do jedného zhluku, tāžko určiť predpokladaný počet zhlukov, (neviem či sa nemôže zaseknúť na lokálnom minime)
- Nevýhody: citlivost na šum, odlehlé hodnoty, shluky rôznych velikostí, nekonvexní tvary
- **Výhody:** ľahká implementácia?
- Výhody: rýchlosť
- Výhody: funguje dobře tam, kde data tvoří kompaktní shluky

3. Rozdíl klasifikace a predikce. Uveďte fáze klasifikace/predikce. (3b)

klasifikace = zařazení daného objektu do jisté třídy na základě jeho vlastností (diskrétní hodnoty například rostliny, řády živočichů, druhy atd.)

predikce = předpověď jisté hodnoty (obecně spojitého charakteru) pro daný objekt na základě jeho vlastností

fáze (v opoře je to nadepsáno jako fáze klasifikace, ale pro predikci jsem to zvlášť nenašla):

1. fáze učení - vybíráme z databáze vzorky dat, u kterých musíme znát, do které třídy jsou zařazeny (trénovací množina). úkol klasifikátoru je zjistit klasifikační pravidla, pomocí kterých může objekty klasifikovat do daných tříd s jistou přesností.
2. fáze testování - opět vybírány vzorky dat (testovací množina), které jsou pomocí naučených klasifikačních pravidel zařazovány do tříd. opět musíme znát, do které třídy vzorky patří, abychom mohli spočítat procento úspěšnosti klasifikace. testovací množina by se neměla prolínat s trénovací.
3. fáze použití - klasifikace neznámých dat (neznáme třídu)

Z našich prednašok (FIIT STU):

Klasifikace = zařazení daného objektu do jisté třídy na základě jeho vlastností

- **Fáze klasifikace:**
 - I. **Trénování** - na základě trénovacích vzorů (u nichž víme, do jaké třídy patří) určení „pravidel“ klasifikátoru.
 - II. **Použití** pro klasifikaci neznámých dat, typicky po **Testování** - pravidla z kroku I. jsou testována na jiných vzorech dat (určení procenta úspěšnosti).
- **Predikce** = předpověď jisté hodnoty (ze spojité funkce) pro daný objekt

4. Asociační pravidla v **relačních DB**: (10b)

a. Jaké typy pravidel lze dovolat v databáz

- **Jednodimenzionální pravidla:** $\text{koupí}(X, \text{"mléko"}) \Rightarrow \text{koupí}(X, \text{"chléb"})$
- **Vícedimenzionální pravidla:** 2 dimenze nebo predikáty
 - Mezi-dimenzionální asociační pravidla (predikáty se neopakují) $\text{věk}(X, \text{"19-25"}) \wedge \text{zaměstnání}(X, \text{"student"}) \Rightarrow \text{koupí}(X, \text{"cola"})$
 - Hybridně-dimenzionální asociační pravidla (predikáty se opakují) $\text{věk}(X, \text{"19-25"}) \wedge \text{koupí}(X, \text{"popcorn"}) \Rightarrow \text{koupí}(X, \text{"cola"})$
- **Kategorické atributy (neviem ci patri pod pravidla)**
 - konečný počet možných hodnot, nemá uspořádání
- **Kvantitativní atributy (neviem ci patri pod pravidla)**
 - numerické, definováno uspořádání nad hodnotami mi atributu

Získávání asociačních pravidel – různé typy

- Booleovské vs. kvantitativní asociace (založeno na typu zpracovávaných hodnot)
 - $koupí(x, "SQLServer") \wedge koupí(x, "DMBook") \rightarrow koupí(x, "DBMiner") [0.2\%, 60\%]$
 - $věk(x, "30..39") \wedge příjem(x, "42..48K") \rightarrow koupí(x, "PC") [1\%, 75\%]$
- Jednodimenzionální vs. vícedimenzionální asociace (viz. předchozí příklad)
- Jednoúrovňová vs. víceúrovňová asociační pravidla
 - Které druhy čokolády jsou v asociaci s kterými druhy dětských plen?
- Různá rozšíření
 - Korelace, analýza kauzality
 - Asociace v sobě nutně nezahrnuje korelacii nebo kauzalitu
 - Maximální vzory a uzavřené množiny
 - Uplatnění omezení
 - Např., malé prodeje ($\text{sum} < 100$) způsobí velké nákupy ($\text{sum} > 1,000$)?

b. Uveďte příklad relační tabulky a Asociacích Pravidel z ní vydolované. Jaký je hlavní problém dolování v relačních databázích proti transakčním.

Age	Salary	Car	Country
19	15000	VW Golf	Czech Rep.
20	20000	Opel Astra	Germany
44	14000	Ferrari	Germany
20	21000	VW Golf	Czech Rep.
:	:	:	:

$\text{Car}(X, \text{VW Golf}) \Rightarrow \text{Country}(X, \text{Czech Rep.})$

Problémy: diskretizácia kvantitatívnych atribútov (tie v transakčnej asi moc nie sú)

c. Metody dolování Asociacích Pravidel v relačních databázích (uveďte 3 a demonstrujte na příkladu)

//že by toto ? +1

Techniky pro dolování vícedimenzionálních asociací ()

Hledání frekventovaných k-vzorů (predikátů):

- Např.: {věk, zaměstnání, koupí} je 3-vzor.
 - Techniky se dělí podle toho, jak je zpracován atribut věk.
1. **Statická diskretizace kvantitatívních atributů** - Kvantitatívni atributy jsou diskretizovány staticky s využitím předdefinované konceptuální hierarchie.
 2. **Kvantitatívní asociační pravidla** - Kvantitatívni atributy jsou dynamicky diskretizovány do "košů", založeno na rozložení dat.
 3. **Asociační pravidla založená na vzdálenosti** - Dynamický diskretizační proces, který bere v úvahu vzdálenosti mezi datovými body (hodnotami).

5. Dolování v textu: (10b)?

a. Popište reprezentaci dokumentu v boolovském, vektorovém a pravděpodobnostním modelu.

bool m. - jednotlivé indexy termů jsou v daném dokumentu přítomny nebo nepřítomny, Váhy indexů termů jsou v binární formě, Dotaz je tvořen pomocí indexů termů a logických spojek (např. car and repair, plane or airplane)

- Booleovský model určuje, zda je dokument relevantní či irrelevantní na základě shody dokumentu s dotazem

vektor m. - dokumenty a dotazy jsou m-rozmerne vektory (m-celkový počet indexů termov pre vsetky dokumenty), Jednotlivé souřadnice reprezentující dokument d jsou tvořeny TF-IDF vahami termů t, Stupeň podobnosti dokumentu d s ohledom na daný dotaz q je vypočten jako korelace mezi vektory (euklidova vzdálosť vektorov)

pravdepodobnostny m. - proces specifikace vlastností ideální odpovědi = prvy odhad

- První odhad zahrnuje vygenerování předběžného pravděpodobnostního popisu ideální odpovědi, která je použita k vyhledání první množiny dokumentů
- Pomocí interakce s uživatelem je pak dosaženo vylepšení pravděpodobnostního popisu výsledku dotazu

b. Co to je a kčemu: stemming, seznam stop-slov.

- Stemming - slová transformujeme na koreň (učenie, učiť => uč)
- Stop-slová - časté a nič nehovoriace slová (ktorý, a, ako...)
-

Oba způsoby jak Stemming tak StopList jsou využívány k predzpracování textových dat.

c. Vysvětlete váhování TF, IDF, TF-IDF.

- TF (Term Freq) - ako často je term v dokumente
- IDF (Inverse Document Freq) - dôležitosť termu je inverzná funkcia k tomu, v koľkých dokumentoch sa nachádza
- TF-IDF - násobok predchádzajúcich (najvyššie je keď je velakrát nejaké slovo v len jednom dokumente)

d. Jaké dvě metriky se používají k vyčíslení podobnosti dokumentů reprezentovaných vektorem. Jednu rozeberte.

□ Metriky podobnosti

- Relativní výskyt termů
- Kosinova vzdáenosť

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1| |v_2|}$$

- precision = $\frac{\text{relevant príenik retrieved}}{\text{retrieved}}$
 - recall = $\frac{\text{relevant príenik retrieved}}{\text{relevant}}$

- skalární součin vektorů TF-IDF a Normalizovaný skalární součin

- kosinová podobnost a euklidovská vzdálenost

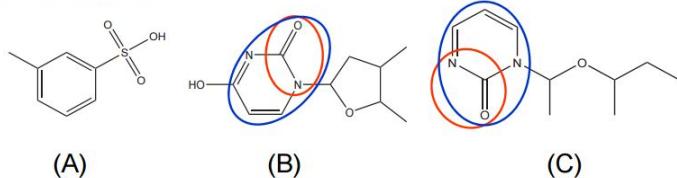
6. Dolování z grafů: (11b)

a. Vytvořte databázi 4 grafů a ukažte příklad frekventovaného a nefrekventovaného podgrafa, stanovte i podmínu, kdy je a kdy ne. Příklady volte dostatečně obecně, aby byla patrná i vlastnost graf - podgraf. (5b)

TOTO je priklad z prednasky kde bola databaza 3 grafov - prednaska 12 slajd 35

Příklad trekventovaných (pod)gratů

Databáze grafů



Frekventované vzory (grafy) (min_sup = 2)



frekventovaný graf - Frekventovaným grafem nazveme graf g , pro který platí $\text{support}(g) \geq \min_{\text{sup}}$ (práh minimální podpory).

b. Uveďte a popište dva prístupy generování frekventovaných grafů a porovnejte je. (6b)
Dva základné prístupy hľadania frekv. grafov:

Založené na Apriori. Rast vzoru

Apriori rieši duplicitu merge-ovaním. Ak sa uzol nenachádza v grafe tak ho pridá a previaže, ak sa nachádza uzol v grafe, tak ho len previaže s týmto uzlom.

Rast vzoru na prvej úrovni rieši duplicitu na druhej duplikuje grafy.

Otázky séria 3

1. ETL fáze při tvorbě OLAP, nějaké podrobnosti kolem toho ohledně transformací, co je to ROLAP a MOLAP (10 b.)

OLAP - Online Analytical Processing, ETL - Extract, transform, load

ROLAP MOLAP ani podrobnejšie sme nemali na FIIT

Typy OLAP serverů: ROLAP, MOLAP, HOLAP

OLAP servry prezentují uživateli multidimenzionální data, aniž bychom věděli, jak jsou data uložena. Implementace OLAP serverů mohou být následující:

Relační OLAP (ROLAP): Využívá relační nebo rozšířené relační databázové systémy pro uložení dat skladu a nástroje OLAP podporují zbyvající funkce, data jsou tedy uložena v relačních tabulkách. Výhodou je lepší škálovatelnost.

Multidimenzionální OLAP (MOLAP): Multidimenzionální pohled na data je podporován díky multidimenzionálním úložištěm založeným na polich. Struktury tedy mají strukturu datové kostky. Výhodou je, že podporuje rychlé indexování na přepracovanou sumarizovanou data. Je neefektivní, pokud data jsou řídká.

Hybridní OLAP (HOLAP): Jde kombinaci MOLAP a ROLAP technologie. Kombinuje škálovatelnost ROLAPu a rychlosť MOLAPu.

Specializované OLAP servry: Jde o relační databázové servry, které poskytují pokročilé dotazovací jazyky nad schématem hvězdy nebo vločky.

http://www.fit.vutbr.cz/~zendulka/VYUKA/Objavovanie_znalosti/oporazzN.pdf

2. Dolování z grafů: co je to ohodnocený graf a další věci, 2 hlavní metody dolování, Duplicity a jak se řeší. (10 b.)

Založené na **Apriori** a **Rast vzoru**

- Graf (ohodnocený/značený – labeled)**

$$G = \{V, E, L, LS\}$$

$V = \{v_1, v_2, \dots, v_m\}$ - množina vrcholů,

$E = \{(v_x, v_y) : v_x, v_y \in V\}$ – množina hran,

$LS = \{l_1, l_2, \dots, l_k\}$ – množina hodnot/značení,

$L: V \cup E \rightarrow LS$ – funkce ohodnocení/značení.

- Graf g_x je podgrafem grafu G , existuje-li izomorfismus z g_x do G .**

Izomorfizmus grafov - keď majú všetky vlastnosti rovnaké (vrcholy, hrany...).

Apriori rieši duplicitu merge-ovaním. Ak sa uzol nenachádza v grafe tak ho pridá a previaže, ak sa nachádza uzol v grafe, tak ho len previaže s týmto uzlom.

Rast vzoru na prvej úrovni rieši duplicitu na druhnej duplikuje grafy.

Eliminace duplicitných podgrafov = pasivní vs. aktivní

3. Adaboost

je to klasifikator

Princip - postupne upravuje vahy jednotlivych trenovacich vzoriek

Vstup - klasifikaciona metoda a mnozina trenovacich vzoriek

Vystup - klasifikacia pomocou vahovaneho hlasovania modelov vytvorených v jednotlivych iteraciach

Algoritmus:

1. Na zaciatku maju vsetky vzorky rovnaku vahu - urobi sa zakladny model
2. Testovanie trenovacich dat a uprava vah nespravne klasifikovanych vzoriek
3. Vytvorenie noveho modelu s novymi vahami. Cely postup sa opakuje k-krat

4 nevýhody K-means - uz je vyriesene vyssie

5. Diskretizace hodnot při dolování, proč je třeba a 3 metody + ke každé příklad

- Podstata: Rozdelení rozsahu kvantitatívного atributu na intervale, pripomienanie návštev intervalu
- Dôvody/prínosy:
 - Niektoré algoritmy (klasifikační) vyžadujú pouze kategórické atributy.
 - Dochází k redukcii dat.
- Metody:
 - plnení
 - shlukování
 - založené na entropii, ...

3 príklady

Plnení: tato technika vyhlazuje setříděná numerická data tak, že zohledňuje hodnoty v blízkém okolí. Provádí tedy lokální vyhlazení. Probíhá ve dvou krocích. V prvém se setříděné hodnoty rozdělí do tzv. košů (bins), zpravidla stejné frekvence (equal-frequency), tj. tak, že každý koš obsahuje přibližně stejný počet hodnot. Druhý krok představuje samotné vyhlazování. Hodnoty v koši se nahradí průměrem koše, mediánem koše nebo každá z hodnot koše tou hraniční hodnotou (minimální, resp. maximální hodnota v koši), ke které má daná hodnota blíže

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Shlukování: může být účinnou metodou pro nalezení odlehlych hodnot. Potenciálně to budou ty, které nebudou zařazeny do žádného shluku.

Entropia: - Priemerné množstvo informácie atribútu: $\text{Entropy}(A) = - \sum (p_i * \log_2(p_i))$; p_i je pravdepodobnosť vyskytu hodnoty atributu - v súvislosti s diskretizáciou netusim...

čotak diskretizácia na intervale rovnakej veľkosti, na intervali s rovnakým počtom prvkov, a intervaly v závislosti na triede (napr. chceme mať aspoň 3 ANO a 3 NIE)
// nie su toto nahodou bins ? ano su

6. Normalizace hodnot - proč se dělá + metody

Normalizace: změna měřítka(mierky) tak, aby hodnoty padly do specifikovaného (zpravidla malého) intervalu/rozsahu

- delame napríklad proto, že nekteré algoritmy umieji (dokážu) pracovat pouze se specificky predzpracovanými daty. Takýmto algoritmom je napríklad s použitím neurónových sietí. (alebo zhľukovacie algoritmy)
- Hodnoty rôznych atribútov môžu mať rôzny rozsah (napr. atribut vek a atribut plat)

Metody:

$$v' = \frac{v - \min_v}{\max_v - \min_v} \cdot (n_{\max_v} - n_{\min_v}) + n_{\min_v}$$

- min-max normalizace

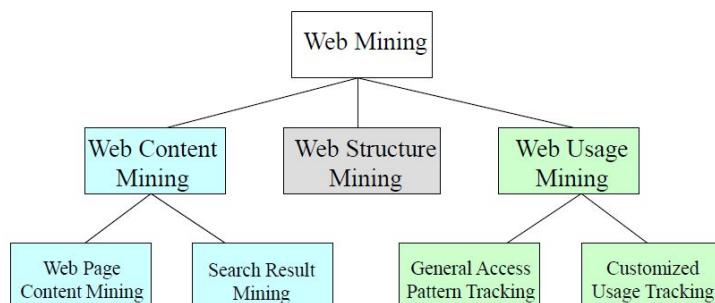
$$v' = \frac{v - \bar{x}_v}{\bar{s}_v}$$

- z-skóre normalizace

$$v' = \frac{v}{10^j}$$

- normalizace změnou dekadického měřítka

7. Dolování z užití webu - jaké věci se běžně snažíme zjistit, z čeho se doluje, fáze dolování.



Co dolujeme (Web Content Mining): //toto nie je odpoved na otazku(pretože "Dolovani z uziti webu" = web usage mining)

- Dolování strukturovaných dat
 - Seznamy produktů nebo služeb

- Srovnávací nakupování
- Meta-vyhledávání
- 1. Napsání extrahovacího programu založeného na formátovacích vzorech stránky
- 2. Indukce (učení) – uživatel manuálně označí množinu trénovacích stránek, ze které jsou generována pravidla, ta jsou aplikována
- 3. Plně automatický přístup – pevná šablona
- Dolování nestrukturovaných dat
 - Na webové stránky je pohlíženo jako na textové dokumenty
 - Současné přístupy založeny na strojovém učení a zpracování přirozeného jazyka k určení extrahovacích pravidel z manuálně označených příkladů
 - Tato oblast spadá do oboru Text Mining

Co dolujeme (Web Usage Mining):

- spracovava sa *pohyb používateľa na webe* (v troch fazach (nizsie))
- extrakcia vzorov *medzi prechodom webom* jednotlivych pouzivatelor
- pouzitie: analiza provozu (prevadzky) webu, optimalizace logicke struktury webovych stranek, personalizace

Tři fáze dolování z užití webu:

1. **Předzpracování**: vyčištění dat, identifikace uživatele/session, zkompletování cesty (tlačítko "zpět" a cache), identifikace transakce.
2. **Získávání znalostí**: počty shlédnutí, strávená doba, nejnavštěvovanější stránky, počet kliknutí, modelování chování uživatele (a rady, kam jít dál).
3. **Analýza vzorů**: používám OLAP analýzu, hledám zajímavé navigační vzory (objektivní a subjektivní metriky),

Z coho sa doluje

- logovacie subory (na ne su aplikovane dolovacie techniky)
- zapojenie heuristik, ktore pracuju so strukturou a obsahom webu, profilom pouzivatela, registracnymi udajmi,...

8. 3 způsoby jak popsat vzdálenost dvou shluků

Vzdálenost mezi shluky

- Minimální vzdálenost (Single linkage):

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

- Maximální vzdálenost (Complete linkage):

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

- Průměrná vzdálenost:

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

- Vzdálenost těžišť:

$$d_{centroid}(C_i, C_j) = |c_i - c_j|$$

- Vzdálenost středových objektů:

$$d_{medoid}(C_i, C_j) = |m_i - m_j|$$

Otázky séria 4

1. co je klasifikace a predikce?

V serii 2

2. dolování frekventovaných sekvenčních vzorů, jak to funguje, k cemu to je, srovnání s klasickými frekvencemi, princip apriori pro sekvenční vzory (co je sekvenční, podsekvenční, databaze sekvenční a sekvenční vzor)

Dolování sekvenčních vzorů

Sekvence: uspořádaná množina prvků nebo událostí. Nezajímá nás tedy frekventovaný vzor, ale sekvenční! Problémem je obrovský počet možných sekvenčních vzorů.

Podstata: mám množinu sekvenčních vzorů a chci najít množinu frekventovaných podsekvenčních vzorů.

Využitie: vzory spravania používateľov na webe, DNA sekvenční, sekvenční proteinov, postupnosti nakupov zakazníkov,..

Požadavky na algoritmus: najít úplnou množinu, efektivnost, škálovatelnost, zahrnout užívateľom definované omezenia. Prikľúčkom algoritmu je GSP

Apriori vlastnosti sekvenčních vzorů: není-li sekvenční frekventovaná, tak ani žádná její supersekvenční vzorec není frekventovaný.

GSP (generalized sequential pattern):

1. Na začátku je každá položka v DB kandidátem délky k=1
2. Loop until nelze nalézt žádnou kandidátní sekvenci:
 - Spočti podporu každé kandidátní sekvence;
 - Z frekventovaných sekvencí délky k generuj na základě vlastnosti Apriori kandidáty délky k + 1

Sekvence : < (ef) (ab) (df) c b >

Databáze sekvencí

SID	sekvence
10	<a(<u>abc</u>)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(<u>ab</u>)(df) <u>cb</u> >
40	<eg(af)cbc>

Prvkem sekvence může být množina.
Položky množiny jsou neuspořádané,
ale budeme je uvádět abecedně.

<a(bc)dc> je podsekvence
<a(abc)(ac)d(cf)>

Je-li práh podpory min_sup = 2, pak <(ab)c> je sekvenční vzor.

Délka sekvence – počet položek (u nás všech písmen) sekvence.

2. bayes, princip, bayes síte

Důležitý je vzorec pro podmíněnou pravděpodobnost: $P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$

Bayesovská klasifikace

Pokud máme vzorek dat $X = (x_1, \dots, x_n)$, které chceme zařadit do tříd C_1, \dots, C_m , tak ho zařadíme do třídy C_i takhle: $P(C_i|X)$ je **maximální**.

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}$$

Bayessův vzorec:

- **Výhody**: snadná implementace, většinou dává dobré výsledky
- **Nevýhody**: předpokládáme, že atributy jsou na sobě **nezávislé** (v praxi většinou nejsou) → toto řeší bayesovské sítě

Bayesovské sítě

Obsahují dvě části: **orientovaný acyklický graf** a **tabulky podmíněných**

pravděpodobností. Jednotlivé uzly grafu reprezentují atributy, závislosti mezi atributy jsou hrany. Tyto sítě je potřeba naučit (určit topologii sítě a doplnit hodnoty do tabulky PP). Pro vzorek dat x_1, \dots, x_n se pravděpodobnost výskytu vzorku vypočítá:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | Parents(Y_i))$$

3. dolovani na zaklade omezeni

- Omezení typu znalosti: určíme typ znalosti, např. asociační pravidla
- Datová omezení: specifikují data, která budou použita pro dolování
- Omezení dimenzí/úrovní dat: specifikují dimenze dat, které budou při dolování použity
- Omezení zajímavosti: specifikace prahů pro metriky zajímavosti, např. podpora a spolehlivost
- Omezení pravidel: specifikuje tvar výsledných pravidel. Omezení mohou být vyjádřena jako např. metapravidla.

Získávání znalostí založené na omezeních

- Interaktivní a efektivní dolování gigabajtů dat?
 - Je to reálné? — Ano, pokud se správně využijí omezení!
- Jaké druhy omezení mohou být použity?
 - Omezení typu znalosti: klasifikace, asociační pravidla atd.
 - Datová omezení: dotazy podobné SQL
 - Najdi dvojice produktů, které se prodávají současně v Olomouckém kraji v listopadu 2005.
 - Omezení úrovně/dimenze:
 - ve vztahu k regionu, ceně, druhu zboží, kategorii zákazníka.
 - Omezení pravidel
 - malé prodeje (cena < \$10) způsobí velké prodeje (sum > \$200).
 - Omezení zajímavosti:
 - silná pravidla ($\text{min_support} \geq 3\%$, $\text{min_confidence} \geq 60\%$).

4. dolovani struktury webu, alg. PageRank a HITS

Využití odkazů mezi stránkami

Přidělování vah www stránkám na základě vytvořeného grafu odkazů

Váha udává citovanost dané stránky

Reprezentace webu v databázi

Zachycení struktury webu pomocí popularity a seznamu odkazovaných stránek

Výsledkem model představující strukturu odkazů webu

Nalezení mikrokomunit na webu

PageRank (Google):

Používá se pro určení důležitosti stránky. V případě cyklických závislostí se výpočet komplikuje, takže to řešíme:

1. V prvním kroku přiřadit iniciální hodnoty
 2. Iterativně počítat hodnoty pro jednotlivé www stránky
 3. Výpočet skončí v okamžiku, kdy hodnoty začínají konvergovat (20 - 40 iterací).

T1, ..., Tn odkazují na stránku A

P(Ti) - PageRank stránky Ti

C(Ti) - počet odkazů ze stránky Ti

d - koeficient je PageRank stránky, na kterou nic neodkazuje

PageRank stránky A pak je:

$$P(A) = \left(1-d\right) + d\left(\frac{P(T_1)}{C(T_1)} + \dots + \frac{P(T_n)}{C(T_n)}\right)$$

Algoritmus HITS

Počítá seznam hubů a autorit pro zadané téma. Bere web jako orientovaný graf (uzly a hrany). Výsledkem algoritmu pro dané téma je seznam se stránkami s nejvyššími váhami hubů a stránky s nejvyššími váhami autorit. Ignoruje textový obsah, bere v potaz pouze odkazy.

Problémy: špatné výsledky pro přesně specifikované téma, stejná HTML šablona pro web → všechny stránky můžou mít nerelevantní odkazy.

5. vlastnosti vzdalenostních funkcí, hierarchické shlukování, jak se vypočítá vzdalenost prvku s intervalovými hodnotami

Vlastnosti

Vzdálenostní funkce (1)

- ## Vlastnosti metriky:

$$d(i, j) \geq 0$$

$$d(i,i)=0$$

$$d(i, j) = d(j, i)$$

$$d(i, j) \leq d(i, h) + d(h, j)$$

Hierarchické metody

- Je vhodné zadat ukončující podmínu (počet tříd/práh) \
- Shlukující hierarchické metody (častější) \\
 - Inicializace: každý objekt tvoří zvláštní třídu
 - Slučování nejpodobnějších tříd
 - Ukončení: všechny třídy spojené do jedné třídy nebo dosažení požadované úrovně shlukování
 - Např. Agnes (Agglomerative Nesting)

Nominální atributy

Zobecnění binárních proměnných

Metoda 1: Jednoduché párování kde m je počet shod a p je počet atributů, lze případně přiřadit váhy.

$$d(i, j) = \frac{p-m}{p}$$

Metoda 2: Zakódování pomocí binárních atributů: každému stavu nominálního atributu odpovídá jeden nový binární atribut + blízkost určena vztahy pro binární atributy.

Ordinální atributy

Lze zpracovat jako **intervalové**

1. Jednotlivým hodnotám přiřadíme pořadí

$$r_{if} \in \{1, \dots, M_f\}$$

2. Transformujeme rozsah všech proměnných na interval (0,1):

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

3. Pro hodnoty z_{if} použijeme libovolnou vzdálenostní funkci

Otzázky séria 6

1. Popsat stručně a vyjmenovat 5 druhů dat pro dolování.
Napsat co dolovat u každého.

relačná databáza - najčastejší zdroj dát, sprístupnenie pomocou SQL

dátové sklady - multidimenzionálna dátová kocka, sprístupnené pomocou OLAP operácií

dokumenty (textové, multimediálne) - po predspracovaní (vektor príznakov)

transakčné databázy - subor, ktorého kazdy zaznam reprezentuje jednu transakciu

prúdy dát - zo senzorov, kamier, potreba analyzovať a docasne ukladať data, specialne dolovacie algoritmy

- 2) Detekce odlehlé trajektorie

nemali sme

3. Popsat tri typy asociačnich pravidel+ priklady a u dvou si vybrat dve metody pomocí kterých je získáváme

Získávání asociačních pravidel – různé typy

- Booleovské vs. kvantitativní asociace (založeno na typu zpracovávaných hodnot)
 - koupí(x, "SQLServer") \wedge koupí(x, "DMBook") \rightarrow koupí(x, "DBMiner") [0.2%, 60%]
 - věk(x, "30..39") \wedge příjem(x, "42..48K") \rightarrow koupí(x, "PC") [1%, 75%]
- Jednodimenzionální vs. vícedimenzionální asociace (viz. předchozí příklad)
- Jednoúrovňová vs. víceúrovňová asociační pravidla
 - Které druhy čokolády jsou v asociaci s kterými druhy dětských plen?
- Různá rozšíření
 - Korelace, analýza kauzality
 - Asociace v sobě nutně nezahrnuje korelacii nebo kauzalitu
 - Maximální vzory a uzavřené množiny
 - Uplatnění omezení
 - Např., malé prodeje ($\text{sum} < 100$) způsobí velké nákupy ($\text{sum} > 1,000$)?

PCA? - Cíl: snížit dimenzionalitu lieárni kombinací rysů nesoucích nejvíce informace.
1-dimenzionálne booleovské: Apriori, FP strom
Viacúrovňové (v hierarchii): Zhora-dolu, s redukciami minimálnej podpory

4. DENCLUE - uz bolo vyssie

5. Zádány dva problémy. Krátké a dlouhé sekvence biologickych dat. Vysvetlit, zda se jedno resi jednoduchým markovovým modelem a druhý skrytým markovovým modelem

6. Dolování webu (co dolovat, rozdíl mezi dokumentem a webem, princip VIPS, priklady dolovani)

Rozdiely www vs dokument:

1. Stránka neobsahuje len obsah ale aj reklamy alebo odkazy na ine stránky

2. Na jednej stranke je viac tem
3. Zlozitejsia struktura - tu asi myslí len html vs obyčajny text
4. Vizualne informacie
5. Velmi rychly rast poctu stranok a caste zmeny v nich

Co mozeme dolovat zo stranky

Web content mining:

1. Dolovanie obsahu webovej stranky - napr identifikacia obsahu stranky na zaklade dotazovacieho jazyka, vyhľadanie cien produktov na roznych strankach a podobne
2. Dolovanie vo vysledkoch vyhľadavani
3. Dolovanie strukturovanych dat - napr zoznamy produktov
4. Dolovanie nestrukturovanych dat - na stranku sa pozera ako na textovy dokument

Web structure mining:

1. Vyuzitie odkazov medzi strankami
2. Reprezentacia webu v databaze
3. Najdenie mikrokomunit na webe
4. Vytvorenie modelu predstavujuceho strukturu odkazov webu

Web usage mining:

1. Spracovavanie pohybu pouzivatela na webe
2. Extrahovanie vzorov pouzivatelov
3. Aplikacia dolovacich technik na logovacie subory
4. Zapojenie heuristik

VIPS

Algoritmus VIPS (VIision-based Page Segmentation)

Motivace

V mnoha prípadech mohou byt téma odlišena pomocí vizuálních vlastností, jako pozice, vzdáenosť, font, barva

Cíl

Extrakce sémantické struktury stránky založené na vizuální prezentaci.

Procedura

Rozdelení stránky shora dolů založené na oddělovačích

Výsledek

- Stromová struktura, každý uzel stromu odpovídá jednomu bloku stránky
- Každému uzlu je přiřazena hodnota (stupeň koherence), která indikuje složitost obsahu bloku z hlediska vizuálního vnímání
- Každému bloku je přiřazena hodnota důležitosti
- Hierarchie nebo rovina

Oázky séria 7

1. Dolování dat na webu
2. Zdroje dat pro dolování
3. Biologická data - Markovy řetězce a Markovy skryté modely
4. Metoda DenClue

5. Metody klasifikace
 6. Pak něco s odlehlýma hodnotama tuším
- Myslim ze vsetky sú vyššie vypracovane**

Otázky séria 8

1. DENCLUE (co to je, jak jsou pocatecni parametry)
2. Nejake 2 priklady z biologickych dat - na ktery se pouziji Markovovy modely a na ktery\ Skryte Markovovy modely

Markov model - Method for performing a [random walk](#) will sample from the [joint distribution](#).

HMM (Hidden Markov Model) - One common use is for [speech recognition](#), where the observed data is the [speech audio waveform](#) and the hidden state is the spoken text.

3. Dolovani z webu - tri typicky druhy co se z webu doluje, tušim uvest nejaký algoritmus ktery se pro kazdy druh pouziva

Content - VIPS (VIision-based Page Segmentation)

Structure - PageRank, HITS

Usage - dolovanie na logovacie súbory, napr. zhľukovacie algoritmy (neurónky SOM...)

4. Tri ruzna deleni asociacnich pravidel, na jake typy to deli, vybrat 2 typy a napsat jakym zpusobem lze takovato asociaci pravidla ziskat
5. Uvest nejakych 6 typu dat ze kterych se da dolovat, charakterizovat je a uvest co se z nich typicky doluje (takze priklad: transakcni db, identifikator transakce a jednotlive polozky transakce, asociaci pravidla)

a

Otázky séria 9

1. redukce dimenziality dat (kdy, jak, proc a popsat principy)

Kedy ?

- Dat je hodně
- Analýza/dolování složitých dat by byla pro úplný soubor časově náročná

Ako ?

Základní metody redukce dimenziality:

- selekce rysů/atributů
- extrakce rysů/atributů

Výběr rysů/atributů (Feature Selection) – proces hledání nejlepší podmnožiny rysů/atributů zdrojových dat s ohledem na cíl zpracování nebo kritérium výběru.

Extrakce rysů/atributů (Feature Extraction) – proces hledání deskriptorů ve zdrojových datech, které nejlépe reprezentují zdrojová data s ohledem na cíl zpracování. Obvykle jde o transformaci do nového prostoru rysů/atributů.

Prečo ?

- Cíl: získání redukované reprezentace datového souboru, který je mnohem menší, ale při analýze/dolování dává stejně (nebo téměř stejně) výsledky.
- Cíl (PCA): snížit dimenzionalitu lineární kombinací rysů nesoucích nejvíce informace.
- identifying a smaller number of uncorrelated variables, called "principal components" zníženie výpočtovéj náročnosti ?

2. Laplaceova korekce u Bayesovské klasifikace (co, proc) + princip Bayesovských sítí

Laplaceova korekce (přidání jednoho prvku do všech množin), aby sme odstránili prípady, keď pravdepodobnosť nejakej udalosti je 0.

3. klasifikace vs. predikce
4. príklad na Bayesovskou klasifikaci: Zadány množiny atributu $\{a,b,c\}$ a $\{x,y,z\}$, trídy $\{0,1\}$. Zadán vzorek prvku (napr $\{a,x,0\}, \{b,z,1\} \dots$), vypočítať klasifikaci jednoho prvku (napr. $\{b,x,?\}$).

// Vyšla vám pravdepodobnosť 0 pre obe triedy?

5. uvést jednu metriku shlukovacích metod; uvést a popsat vlastnosti shlukovacích metod, které tyto metriky využívají

Dunn index - cím vyssi tym lepsie zhľuky.

- Průměr shluku: $\Delta(\Omega_i) = \max_{x,y \in \Omega_i} \|x - y\|$
- Vzdálenost mezi shluky: $\delta(\Omega_i, \Omega_j) = \min_{x \in \Omega_i, y \in \Omega_j} \|x - y\|$
- Dunnův index: $r = \min_i \min_{j, i \neq j} \frac{\delta(\Omega_i, \Omega_j)}{\max_k \Delta(\Omega_k)}$

Davies-Bouldinův index

- Vzdálenost uvnitř shluku: $s_i = \frac{1}{\text{card}(\Omega_i)} \sum_{x \in \Omega_i} \|x - v_i\|^2$
 - v_i je střed shluku i
 - $\|\cdot\|$ je libovolná vzdálenostní funkce
- Vzdálenost mezi shluky: $d_{i,j} = \|v_i - v_j\|^2$
- Poměr vzdáleností: $r_i = \max_{j, j \neq i} \frac{s_i + s_j}{d_{i,j}}$
- Výsledný index: $r = \frac{1}{c} \sum_{i=1}^c r_i$
- Optimální počet shluků je takový, pro který je výsledný index minimální.

Otázky séria 10

1. Popsat tri typy asociacnich pravidel+ priklady a u dvou si vybrat dve metody pomocí kterých je získáváme
2. DENCLUE
3. Zádány dva problémy. Krátké a dlouhé sekvence biologickych dat. Vysvetlit, zda se jedno resi jednoduchým markovovým modelem a druhý skrytým markovovým modelem
4. Dolování webu (co dolovat, rozdíl mezi dokumentem a webem, princip VISP, priklady dolovani)
5. Dolovani z webu - tri typicky druhy co se z webu doluje, tusim uvest nejaký algoritmus ktery se pro kazdy druh pouziva
6. Ake typy atributov pozname podla oboru hodnot.
7. Datove sklady, definicia, 4 vlastnosti podla neviem koho, rozdiel mezdi OLAP a OLTP
8. Dolovanie textu(nepamamatam si podotazky).
9. Popisat jeden z klasifikatorov zalozenych na hustote (si nespominam ktory).
10. Metody klasifikace
11. Tri ruzna deleni asociacnich pravidel, na jake typy to deli, vybrat 2 typy a napsat jakym zpusobem lze takovato asociaci pravidla ziskat
12. Uvest nejakych 6 typu dat ze kterych se da dolovat, charakterizovat je a uvest co se z nich typicky doluje (takze priklad: transakcni db, identifikator transakce a jednotlive polozky transakce, asociaci pravidla)
13. redukce dimenzionality dat (kdy, jak, proč a popsat principy) // vyššie

14. Laplaceova korekce u Bayesovské klasifikace (co, proč) + princip Bayesovských sítí
15. klasifikace vs. predikce
16. příklad na Bayesovskou klasifikaci: Zadány množiny atributů {a,b,c} a {x,y,z}, třídy {0,1}. Zadán vzorek prvků (např {a,x,0},{b,z,1}...), vypočít klasifikaci jednoho prvku (např. {b,x,?}).
17. uvést jednu metriku shlukovacích metod; uvést a popsat vlastnosti shlukovacích metod, které tyto metriky využívají
18. dolování víceúrovňových pravidel (co je víceúr.p.; příklad databáze, která toto umožňuje; metody určení podpory; algoritmus)

Položky často tvoří hierarchii

Položky na nižší úrovni zpravidla mívají nižší podporu

Pravidla na různých úrovních abstrakce mohou být užitečná

Transakční databáze může být zakódována podle úrovní

Víceúrovňová pravidla:

Progresivní zanořování (shrnutí)

- Shora dolů, přístup progresivního zanořování:
 - Nejprve získáme frekventované položky na nejvyšší úrovni:
mléko (15%), chléb (10%)
 - Pak získáme „slabší“ frekventované množiny na nižší úrovni
2% mléko (5%), tmavý chléb (4%)
- Různé hodnoty minimální podpory pro různé úrovně vedou k různým algoritmu:
 - Jestliže máme jedinou hodnotu *min_podporu* pro všechny úrovně
Pak je potřeba smazat *t* jestliže některý z předchůdců *t* není frekventovaný
 - Jestliže máme redukovanou *min_podporu* na nižších úrovních
Pak je potřeba zkoumat jen ty následníky, podpora jejichž předchůdce je nezanedbatelná x

Víceúrovňové asociace: konstantní vs. redukovaná podpora

- Konstantní podpora: stejná minimální podpora pro všechny úrovně
 - + Jedna hodnota min. podpory. Není potřeba zkoumat množiny obsahující jakoukoli položku, jejíž předchůdci nemají minimální podporu.
 - - Položky z nižších úrovní se nevyskytují tak často. Jestliže je podpora
 - příliš vysoká \Rightarrow ztráta asociací na nižší úrovni
 - příliš nízká \Rightarrow generuje příliš mnoho asociací na vyšší úrovni
- Redukovaná podpora: snížená minimální podpora na nižších úrovních
 - Existují čtyři strategie:
 - Nezávisle na úrovních
 - Filtrování úrovně k-množinou
 - Filtrování úrovně jednou položkou
 - Kontrolované filtrování úrovně jednou položkou

19. získávání silných asociačních pravidel z frekventovaných množin (postup; vzorec úrovně podpory; výpočet AP - příklad)

Asociační pravidlo, které má podporu a spolehlivost vyšší než je uživatelem zadaná hodnota, nazveme silné asociační pravidlo.

Základní postup:

Výpočet frekventovaných množin
na základě minimální podpory
časově náročnější krok

Generování silných asociačních pravidel z frekventovaných množin
na základě minimální spolehlivosti

Získávání asociačních pravidel - příklad

Trans. ID	Položky	Min. podpora 50%	Min. spolehlivost 50%
2000	A,B,C		
1000	A,C		
4000	A,D		
5000	B,E,F		

Frekv. množina	Podpora
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

- Pro pravidlo $A \Rightarrow C$:
 - podpora = $\text{podpora}(\{A \cup C\}) = 50\%$
 - spolehlivost = $\text{podpora}(\{A \cup C\}) / \text{podpora}(\{A\}) = 66.6\%$

Otázky séria 11

1. TL fáze pri tvorbe OLAP, nejaké podrobnosti kolem toho ohľedne transformací, co je to ROLAP a MOLAP (10 b.)
2. Dolování z grafu: do je to ohodnocený graf a další veci, 2 hlavní metody dolování, Duplicity a jak se reší. (10 b.)
3. Adaboost
4. 3 nevýhody K-means
5. Diskretizace hodnot pri dolování, proc je treba a 3 metody + ke každému príkladu
6. Normalizace hodnot - proc se delá + metody
7. Dolování z užití webu - jaké veci se bežne snažíme zjistit, z čeho se doluje, fáze dolování.
8. alespon 3 zpusoby jak popsat vzdáenosť dvou shluků

Otázky séria 12

1) Jaké popisné charakteristiky se používají pri určování stredu dat? 5b

Charakteristiky středu dat

- Aritmetický průměr (mean)
 - Vážený průměr:
 - Upravený průměr(trimmed mean): ořezání extrémních hodnot
- Medián
 - Prostřední hodnota seřazeného seznamu, je-li počet hodnot lichý,
jinak průměr dvou prostředních hodnot
 - Holistická míra (k výpočtu potřebujeme mít všechny hodnoty)
- Modus (mode)
 - Nejčastější hodnota v datech
 - Jedno-, dvoumodální, (multimodální)

2) Popsat tri druhy rozdelení asociacných pravidel a
ke dvou z nich pak napsat, jak je dolujeme. 5b

3) Popsat obecné shlukovací metody založené na hustotě, jejich výhody a nevýhody. Pak popsat DBSCAN: algoritmus, princip 10b

Metody založené na hustotě

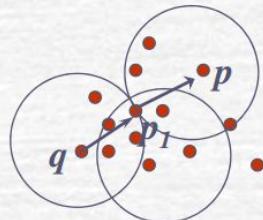
- ✓ **Shluky:** oblasti s velkou hustotou objektů oddělené oblastmi s malou hustotou objektů
- ✓ Shluk je zvětšován, dokud hustota objektů v sousedství neklesne pod danou hranici
- ✓ Umožňují:
 - Nalézt shluky různých tvarů
 - Odfiltrovat šum a odlehlé hodnoty
- ✓ **Nevýhoda:** nutnost definovat parametr hustoty

Metoda DBSCAN

- ✓ *Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density*
- ✓ **Shluk:** maximální množina bodů spojených na základě hustoty
- ✓ **ϵ -okolí:** okolí objektu o poloměru ϵ parametry
metody
- ✓ **jádro:** objekt jehož ϵ -okolí obsahuje alespoň určitý minimální počet (*MinPts*) objektů
- ✓ Objekt p je **přímo dosažitelný na základě hustoty** z objektu q v množině objektů D , jestliže p se nachází v ϵ -okolí objektu q a q je jádro.

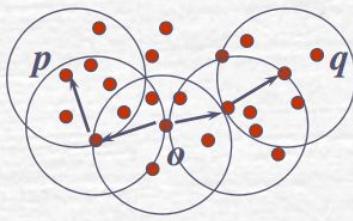
Metoda DBSCAN

- Objekt p je **dosažitelný na základě hustoty** z objektu q v množině objektů D , jestliže existuje řetězec objektů p_1, \dots, p_n , kde $p_1=q$ a $p_n=p$ takový, že objekt p_{i+1} je přímo dosažitelný na základě hustoty z objektu p_i , pro $1 \leq i \leq n$, $p_i \in D$.



Metoda DBSCAN

- Objekt p je **spojený na základě hustoty** s objektem q v množině objektů D , jestliže existuje objekt $o \in D$ takový, že objekty p a q jsou dosažitelné na základě hustoty z objektu o .



Metoda DBSCAN

Postup (zjednodušený):

Všechny objekty označ jako „nenavštívené“;

repeat

Náhodně vyber nenavštívený objekt;

if jde o jádro **then** Vytvoř nový shluk;

Označ objekty jako kandidáty pro expandování;

while lze expandovat **do**

Expanduj shluk;

end while end if

until je nějaký nenavštívený objekt;

Problém: určení parametrů ϵ a $MinPts$

4) Adaboost: co to je, princip, vstup, výstup, algoritmus 7b

5) Dolování z webu: rozdíl mezi dolováním z www stránky a dokumentem, popsat 3 oblasti, co mužeme dolovat z webu, popsat segmentaci, popsat VIPS 10b

6) Detekce odlehlé trajektorie: prístupy, algoritmy 10b

7) Bootstrap: co to je, kde se to používá 4b

V testovani modelov

Bootstrap

- Náhodně vybíráme vzorky pro učení: pro n-trénovacích dat provedeme n-krát výběr s navracením
- Vzorky v trénovací množině se mohou opakovat
- Pro rozumně se tak do výběru dostane cca 63% trénovacích dat
- Zbylých 37% dat využijeme pro testování

Otázky séria 13

1) definovať distributivné, algebraické a holistiké míry +
príklady v kontextu charakteristik dat

1. **distributivne** miery – majú rovnaké výsledky aj pri rozdelení dát na subčasti, napr. count(), min(), max(), aj keď spravíme nad jednotlivimi castami a potom spolu, výsledok je rovnaky
2. **algebraicke** miery – sú také, ktoré používajú viaceré distributívne miery, napríklad avg() je vlastne sum()/count()

3. **holisticke** miery – také, kde neexistuje algebraická miera pre tieto funkcie, napríklad median(), mode(), rank()
- 2) definovať ohodnocený graf, databázi grafu a frekventovaný graf, ilustrovať na príklade, charakterizovať dva základní postupy hľadania frekvencie grafu, problém duplicit a ako ich rešiť
- 3) popsat obecné shlukovanie na základe hustoty, princip metody DENCLUE, popsat jej parametre, výhody a nevýhody
- 4) skryté Markovské modely pro CpG ostruvky: na čo sa používajú, popsať je, postup hodnocenia, názvy používaných metod

5) co to je cross validate a proc se používá, napsat postup

Vstupná množina dat je rozdelená na podmnožiny. Jedna podmnožina slúži ako testovací množina, zbylé podmnožiny slúžia ako trénovacie množiny. Klasifikátor natreňuje model na trénovací množinu a pomocou testovací množiny testuje presnosť a výkonnosť tohto modelu. Tento proces sa několikrát opakuje, pokaždé s jinou podmnožinou tvorícou trénovací a testovací množinu.

□ Křížová validace (*n*-fold cross-validation)

- Rozdelení dat do *n* častí, vždy jedna časť ponechána pro testovanie
- Učení a testovanie sa opakujie *n*-krát
- Výsledek testovania sa průměruje

6) vlastnosti textových dokumentu (významné z hľadiska dolovania), jaké rysy se využívají, metody predzpracovania, co to je summarizace + jaké metody známe

Základní charakteristika

- Vysoká dimenzionalita rysů
 - reprezentace dokumentu je výrazne rozsáhlejší než klasických strukturovaných dat
- Řídkost rysů
 - většina rysů se vyskytuje v malém množství dokumentů
 - malá podpora vzoru
- Data jsou často semistrukturovaná
- Nejčastější rysy dokumentů
 - Znaky – většinou se nepoužívají, ale jde o úplnou reprezentaci
 - Slova – je jich mnoho, optimalizuje se
 - Termy – slova nebo spojení slov – nutnosť slovníku
 - Koncepty – slova, která se přímo nemusí vyskytovat v dokumentu – např. téma dokumentu

Typy metod dolování v textu

- Asociační analýza založená na klíčových slovech
- Automatická klasifikace dokumentů
- Detekce podobnosti
 - Shlukování dokumentů od stejných autorů
 - Shlukování dokumentů obsahující informace pocházející z jednoho zdroje
- Analýza souvislostí: neobvyklé korelace mezi entitami
- Analýza sekvencí: predikce opakující se události
- Detekce anomalií: nalezení informace, která se vymyká běžným pravidlům
- Analýza hypertextu
 - Zajímavé vzory v odkazech mezi dokumenty

Sumarizace textů

- Vyjmutí nejdůležitější informace ze zdrojového textu, která jej zestručňuje pro uživatele nebo další zpracování
- Rozdělení podle účelu
 - Indikativní – umožňují rozhodnutí, zda text stojí za to číst, součást vyhledávače, délka do 10% originálu
 - Informativní – 20-30% originálu, pro zběžné seznámení s obsahem
- Rozdělení podle počtu dokumentů
 - Jednodokumentové
 - Vícedokumentové

Luhnova metoda summarizace textu

Častá slova indikují hlavní téma textu

Předzpracování (stemming, stop slova)

Výpočet vah (TF-IDF)

Kroky metody

 Vyber termy s vahou vyšší než zadaný práh

 Zjisti shluky klíčových slov ve větách, které obsahují méně než 4 neklíčová slova

 Vypočti váhy shluků pomocí TF-IDF

 Vypočti váhy vět:

$$\text{weight}(S) = \text{SUMA}(\text{weight}(\text{terms_in_cluster}))$$

 Požadovaný počet vět s nejvyššími vahami představuje výsledek

7) napsat výhody využití FP stromu, nakreslit jeho příklad a na nem ukázat postup dolování FM

Výhody použití FP-stromu

□ Kompletnost:

- nikdy neporuší dlouhou množinu z jakékoli transakce
- udržuje kompletní informaci pro potřeby získávání frekventovaných množin

□ Kompaktnost

- redukce irrelevantní informace—nefrekventované položky jsou smazány
- seřazení podle frekvence výskytu: více frekventované položky budou spíše sdíleny
- nikdy nebude větší než původní databáze
- Příklad: Např. pro Connect-4 DB, komprese může být až 100 násobná

Dоловání frekventovaných množin s využitím FP-stromu

□ Hlavní myšlenka („rozděl a panuj“)

- Rekurzivní zvětšování frekventovaných množin s využitím struktury FP-stromu

□ Metoda

- Pro každou položku, vytvoříme její **podmíněný základ množiny**, a poté její **podmíněný FP-strom**
- Opakování tohoto procesu pro každý vytvořený podmíněný FP-strom
- Dokud není výsledný FP-strom **prázdný**, nebo obsahuje **jen jednu cestu**
 - Jedna cesta generuje všechny kombinace svých podcest, každá z nich je frekventovanou množinou

Otzázký séria 14

1. opravný fuj termín

- 1) 5 kroku predzpracování dat - úloha, důvod, (rešení?)...
 - 2) popsat ETL, ulohy, důvody řešení (ano, kapku redundantní s první otázkou, ale přej to je ok). Rozdíly ROLAP a MOLAP...
 - 3) rolosovací strom, jaké atributy upřednostňuje ID3, jak ID3 řeší spojité nediskretizované atributy, proc a kdy prorezává strom.
 - 4) Kvalita shluku. Vlastnosti, které to popisují. Jak je lze kombinovat. Název alespon jedné metody. (jak je lze počítat?)
- 5) algoritmus apriory - nevýhody. TIDlist / aprioriset, cím vylepšují algoritmus. Jak se dolují max. množiny pomocí max-mineru. (tak nejak...)

Jádro algoritmu Apriori:

Využití frekventovaných $(k - 1)$ -množin ke generování kandidátů na frekventované k -množiny

Využití průchodů databáze a porovnávání množin k zjištění počtu výskytů daného kandidáta

Hlavní problém Apriori: generování kandidátů:

Velké množiny kandidátů

Vícenásobné čtení z databáze - Potřebuje $(n + 1)$ čtení, n je velikost největší množiny

- Apriori:

C_k : Kandidáti na frekventovanou množinu velikosti k
 L_k : frekventované množiny velikosti k

```
 $L_1 = \{\text{frequent items}\};$ 
for ( $k = 1; L_k \neq \emptyset; k++$ ) do begin
     $C_{k+1} = \text{candidates generated from } L_k;$ 
    for each transaction  $t$  in database do
        increment the count of all candidates in  $C_{k+1}$  that are
        contained in  $t$ 
     $L_{k+1} = \text{candidates in } C_{k+1} \text{ with min\_support}$ 
    end
return  $\cup_k L_k;$ 
```

- Řešení: Návrh vhodnějších datových struktur pro uložení frekventovaných množin a informací o tom, které transakce je obsahují.

Algoritmy AprioriItemset, AprioriTIDList (datové struktury)

□ AprioriTID

TID	C_2
1	$\{\{A, B\}, \{A, C\}\}$
2	$\{\{A, D\}, \{B, C\}\}$
3	$\{\{A, B\}, \{A, D\}\}$
4	$\{\}$

□ AprioriItemset

C_2	BV
$\{A, B\}$	1010
$\{A, C\}$	1000
$\{A, D\}$	0110
$\{B, C\}$	0100

□ AprioriTIDList

C_2	RV
$\{A, B\}$	{1, 3}
$\{A, C\}$	{1}
$\{A, D\}$	{2, 3}
$\{B, C\}$	{2}

Algoritmy AprioriItemset, AprioriTIDList (operace)

□ AprioriItemset

- Přítomnost kandidátů v transakcích: Operace AND po bitech:
 - ▣ Bitový vektor pro $\{A, B, C\}$ se určí jako: BV $\{A, B\}$ AND BV $\{A, C\}$: {1010} AND {1000} = {1000}
- Podpora: počet jedniček vektoru

□ AprioriTIDList

- Přítomnost kandidátů v transakcích
 - ▣ Průnik množin RV u kandidátů (podmnožin daného kandidáta)
- Podpora: Počet prvků v množině RV

MaxMiner: Dolování maximálních množin

- 1. průchod: nalezení frekv. položek
 - A, B, C, D, E
- 2. průchod: nalezení podpory pro
 - AB, AC, AD, AE, **ABCDE**
 - BC, BD, BE, **BCDE**
 - CD, CE, **CDE**, DE,
- Je-li BCDE maximální množina, není potřeba kontrolovat BCD, BDE, CDE v dalším průchodu

Tid	Items
10	A,B,C,D,E
20	B,C,D,E,
30	A,C,D,F

Potenciální
max. množiny

6) definovat datový tok, 2 konkrétní zdroje, (ouliers? nevím...)

Datový tok a jeho vlastnosti

- Datový tok
 - **Datový tok** — spojitý, uspořádaný, proměnlivý, rychlý, obsahující obrovské množství dat (teoreticky nekonečný tok).
 - **Tradiční data** v databázích — uložena v konečných perzistentních množinách (souborech záznamů).
- Vlastnosti
 - Obrovský **objem** spojitých dat (ne nutně spojité hodnot), potenciálně nekonečný.
 - Rychle se **mění**, vyžaduje rychlou odezvu v reálném čase.
 - Vystihuje naše dnešní potřeby ve **zpracování dat** (\rightarrow aktuální problém).
 - Náhodný/přímý přístup k datům je drahý \rightarrow **jednopruhodové algoritmy**.
 - Ukládání pouze **sumárních charakteristik** dosavadních dat.
 - Většinou proud obsahuje **nízkoúrovňová** nebo **vysoce dimenzionální data**, proto se vyžaduje víceúrovňové a vícedimenziorní zpracování.

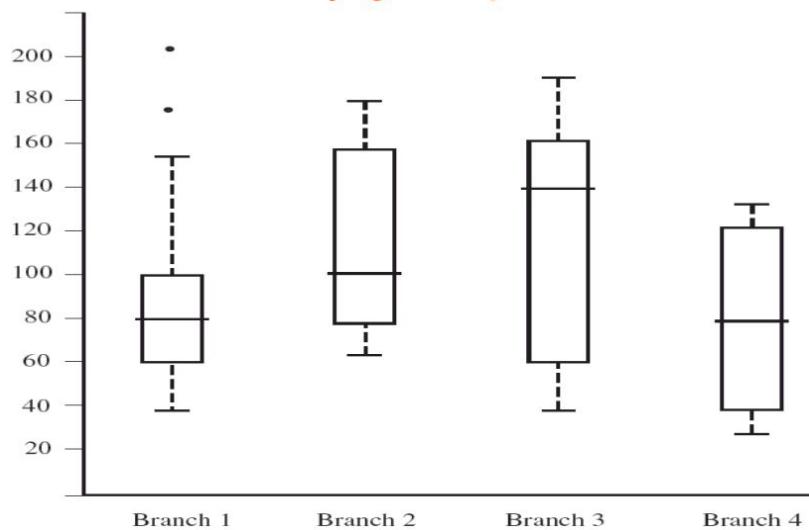
Otzázkы séria 15

2. opravný

1) Nakreslit a popsat grafy - Histogram, Krabicový graf, Kvantilový graf, Q-Q graf

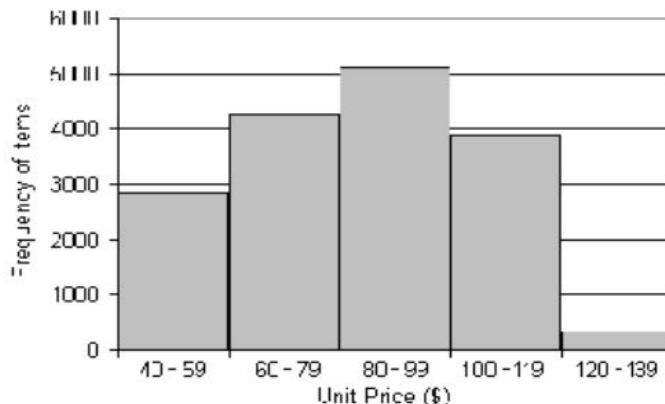
- **Krabicový graf**
 - Data reprezentována obdélníkem
 - Konce obdélníku jsou Q1 a Q3, tj. výška je IQR

- Medián je vyznačen úsečkou uvnitř
- Protažení: dvě úsečky k minimu a maximu, případně individuální odlehlé hodnoty



- **Histogram**

- Zobrazuje početnosti hodnot
 - Grafická metoda pro jeden atribut (proměnnou)

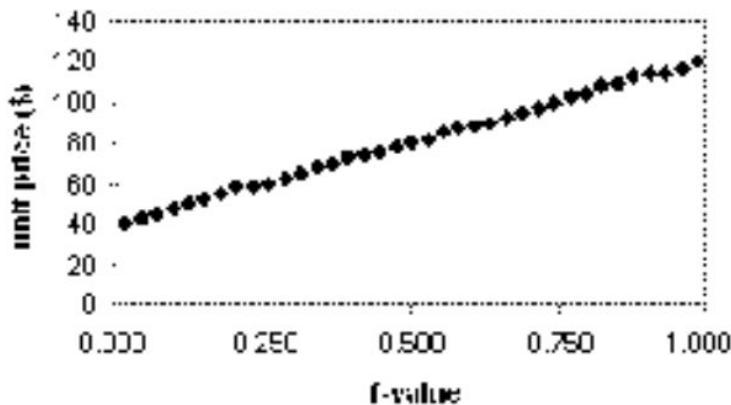


- **Kvantilovy graf**

- Zobrazuje všechna data (možnost posoudit celkové chování i neobvyklé výskytu)
- Vykresluje informaci o kvantilech
 - Pro prvek s hodnotou x_i souboru dat uspořádaného vzestupně, hodnota f i udává, že přibližně $100 f \%$ dat má hodnotu menší nebo rovnou x_i

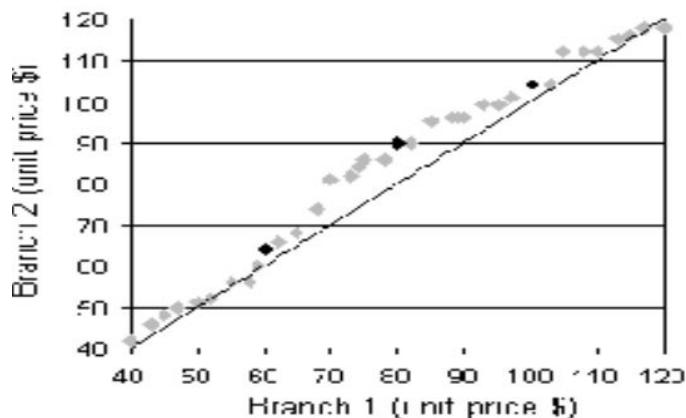
Vlavo => unit price (\$)

Vpravo => f-value



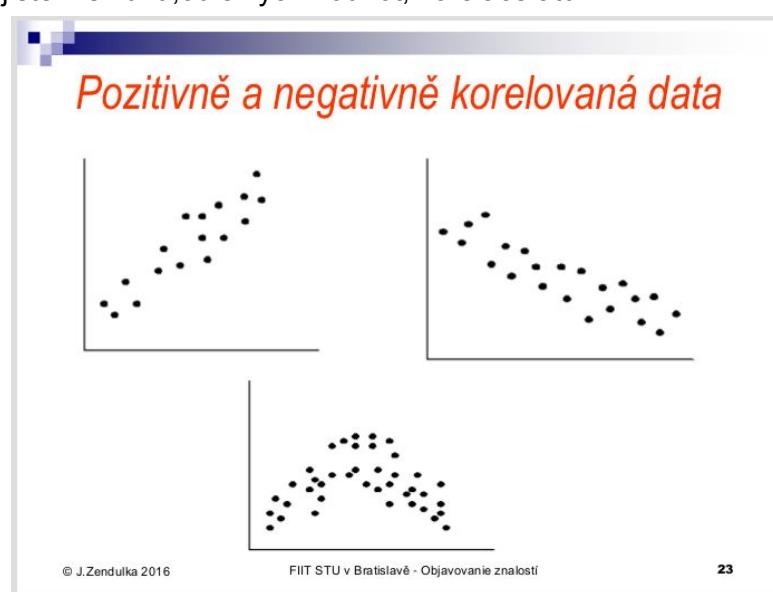
- **Q-Q graf**

- Zobrazuje kvantily dvou atributů (proměnných) navzájem, tj. bod grafu odpovídá stejné pravděpodobnosti na křivce distribuční funkce každého z atributů
- Umožňuje sledovat vzájemného posunu distribuce hodnot obou atributů



- **bodový graf**

- Umožňuje získat první názor na data dvou atributů (v 2D prostoru) pro účely zjištění shluků, odlehčit hodnot, korelace atd.



2) Sekvenční vzory - popsat a vysvetlit, uvést příklady, kde se to využívá

- Sekvence – uspořádaná množina prvků nebo událostí.
- Databáze sekvencí vs. databáze časových řad.
- Frekventované vzory vs. (frekventované) sekvenční vzory.
- Aplikace dolování sekvenčních vzorů
 - Posloupnosti nákupů zákazníků:
 - Během tří měsíců si nejprve koupí počítač, pak digitální kamery a pak USB diskové pole.
 - Lékařská ošetření, přírodní neštěstí (např. zemětřesení), procesy ve vědě a technice atd.
 - Vzory chování návštěvníků webu nebo nějaké interaktivní aplikace.
 - DNA sekvence, sekvence proteinů atd.

Dоловání sekvenčních vzorů

- *Podstata:* Dána množina sekvencí, chceme najít úplnou množinu **frekventovaných** podsekvencí.

<u>Sekvence</u> : < (ef) (ab) (df) c b >	
<u>Databáze sekvencí</u>	
SID	sekvence
10	<a(<u>abc</u>)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(<u>ab</u>)(df) <u>cb</u> >
40	<eg(af)cbc>

Prvkem sekvenčního vzoru může být množina. Položky množiny jsou neuspořádané, ale budeme je uvádět abecedně.

<a(bc)dc> je podsekvence
<a(abc)(ac)d(cf)>

Je-li práh podpory $min_sup = 2$, pak <(ab)c> je sekvenční vzor.

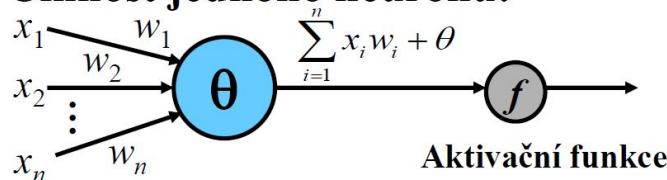
Délka sekvenčního vzoru – počet položek (u nás všech písmen) sekvenčního vzoru.

3. Neuronová síť:

- a) Vysvětlete možnosti využití 1 neuronu pro klasifikaci a způsob učení tohoto neuronu.

Cieľom je mapovať vstupy (x) na výstupy (y), kde hľadajú vzájomné vzťahy (angl. correlation) medzi premennými. Snaží sa nájsť funkciu $F(x)$, napríklad " $4x - 3$ ", ktorá by sa rovnala " y ". Číslo 4 je v tomto prípade váha premennej " x ". Algoritmus neurónovej siete mení váhy vstupných premenných a upravuje hľadanú funkciu.

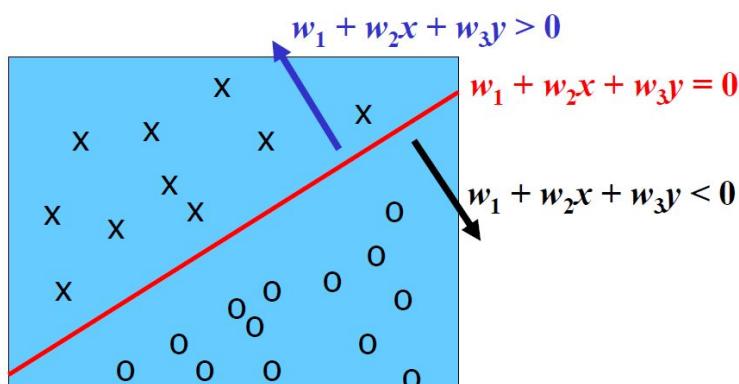
Činnost jednoho neuronu:



Aktivační funkce může mít skokový charakter nebo spojity

$$\text{Příklad aktivační funkce: } y = \frac{1}{1+e^{-x}}$$

- Umí klasifikovat jen do dvou tříd $\{1, -1\}$ a to jen data, která jsou lineárně separovatelná



- Učení:
- Nechť $Z = X$, pokud X je klasifikován do třídy 1 a $Z = -X$, pokud X je klasifikován do třídy -1
- opakovaně procházej vstupní data X a postupně měň W , dokud všechna data nejsou správně klasifikována:
- if $ZW > 0$ then W nemodifikuj else $W := W + Z$

b) Uveďte alespoň 3 výhody a 3 nevýhody klasifikace pomocí neuronové sítě.

Medzi výhody tejto metódy patria paralelné spracovanie informácií a univerzálnosť.

Hlavnou nevýhodou je dlhé učenie.

Výhody – neuronové siete sú jednoduché na implementáciu, proces klasifikacie je časovo nenarocny, narocne je iba učenie, skalovateľne

- nevýhody – neuronové siete sú black box, nevidime dovnutra. Trenovanie je narocne a je potrebna veľka množina trenovacich dat

c) K čemu slouží koeficient učenia a proč je doporučeno jeho hodnotu postupně snižovať?

Parameter "learning rate" predstavuje rýchlosť učenia a nadobúda hodnotu z intervalu $(0, 1)$. Používa sa na zaistenie konvergencie (zbiehania hodnôt). Pri vysokých hodnotách môže prestreliť optimálnu hodnotu. Neuronky majú ešte atribút momentum (hybná sila), čo má približne rovnaké použitie. Odporúča sa nízke momentum a vysoké learning rate a znižovať learning rate a zvyšovať momentum na dosiahnutie konvergencie (aby sme cieľovú najlepšiu hodnotu dosiahli čo najskôr - chceme úspešnosť nad 50+%).

d) Popište základní strukturu vícevrstvé dopředné neuronové sítě.

Ak každý neurón jednej vrstvy vysiela signály na každý neurón nasledujúcej vrstvy hovoríme o dopredných neurónových sietiach. U dopredných sietí neexistujú spojenia medzi neurónmi tej istej vrstvy, ani medzi neurónmi vzdialených vrstiev.

Sieť sa skladá z aspoň 3 vrstiev, kde na každej vrstve sa nachádza určitý počet neurónov.

Na vstupnej vrstve sa získajú dátu a na výstupnej vrstve sa odoberú výsledné hodnoty.

Medzi týmito vrstvami sa nachádza určitý počet skrytých vrstiev, kde prebiehajú hlavné výpočty a váhovanie. Ak môže existovať viac ako jedna skrytá vrstva, neurónová sieť sa nazýva hlboká (angl. deep).

(na toto sa nepýtal, ale pre istotu) Backpropagation

- Tak ako sa mení chyba, musí sa prispôsobiť aj zmena váhy

4) Rozdíl mezi symetrickými a asymetrickými binárními promennými plus príklad u každého

- symetrické - obidva stavy rovnako dôležité, napr. pohlavie
- asymetrické - nie, napr. lekársky test pozitívny, negatívny

b) Uveďte, jakým způsobem je vhodné určit vzdáenosť dvou objektů, které jsou popsány několika symetrickými binárními promennými, a vzdáenosť dvou objektů, které jsou popsány několika asymetrickými binárními promennými.

		objekt <i>j</i>		sum
		1	0	
objekt <i>i</i>	1	q	r	q+r
	0	s	t	s+t
sum		q+s	r+t	p

✓ Symetrické:

- míra odlišnosti: $d(i, j) = \frac{r + s}{q + r + s + t}$

✓ Asymetrické:

- míra odlišnosti: $d(i, j) = \frac{r + s}{q + r + s}$

- míra podobnosti (Jaccardův koeficient):

$$\text{sim}_{\text{Jaccard}}(i, j) = 1 - d(i, j) = \frac{q}{q + r + s}$$

5) Vyhledávání informací v textových datech (information retrieval)

a) Definujte co nejpřesněji úlohu vyhledávání informací v textových datech.

Textové databáze (dokumentové databáze)

Velké kolekce dokumentů z různých zdrojů: novinové nebo výzkumné články, knihy, digitální knihovny, e-maily, www stránky a další

Data jsou velmi často semistrukturovaná

Tradiční techniky se ukazují jako nevhodné, díky velkým objemům dat

Information retrieval (vyhledávání informací)

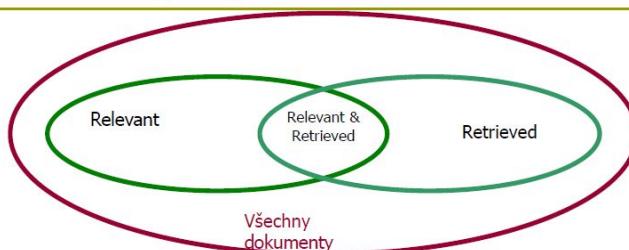
Obor, který byl založen paralelně s databázovými systémy

Informace je organizována do (velkého počtu) dokumentů

Problém IR: nalezení relevantních dokumentů, založené na vstupu od uživatele, např. klíčová slova, příklad dokumentu

b) Vysvětlete rozdíl mezi dvěma metrikami úspěšnosti vyhledávání informací: přesnost a úplnost.

Hlavní metriky vyhledávání informací



- Přesnost: procento z nalezených dokumentů, které odpovídají zadanému dotazu

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- Úplnost (Recall): Procento z relevantních dokumentů, které skutečně byly nalezeny

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

c) Popište alespoň dvě možné reprezentace textového dokumentu pro účely vyhledávání informací

Booleovský model - Předpokládáme, že jednotlivé index termy jsou v daném dokumentu přítomny nebo nepřítomny

Váhy index termů se předpokládají v binární formě

Dotaz je tvořen pomocí index termů a logických spojek: not, and, a or

např.: car and repair, plane or airplane

Booleovský model určuje, zda je dokument relevantní či irelevantní na základě shody dokumentu s dotazem

Vector Space Model - Reprezentace dokumentu pomocí vektoru termů

Term: základní pojem, např., slovo nebo fráze

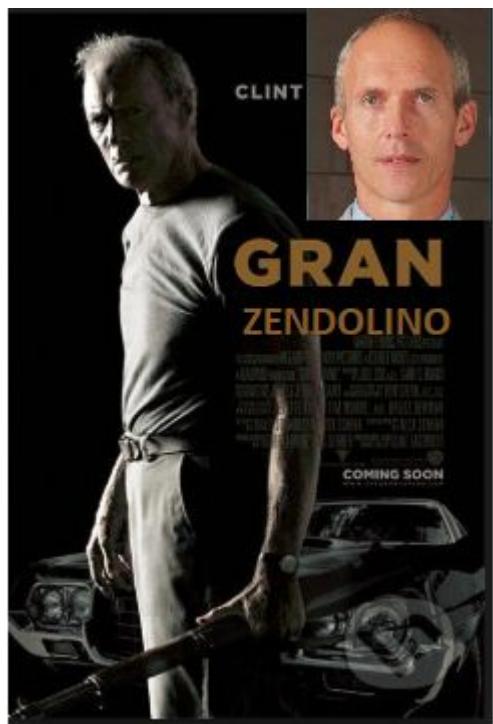
Každý jeden term definuje jednu dimenzi

N termů reprezentuje N-dimenzionální prostor

Jeden element vektoru koresponduje váze daného termu

Např., $d = (x_1, \dots, x_N)$, x_i je "důležitost" termu i

Nový dokument je přiřazen k nejvíce pravděpodobné kategorii na základě podobnosti vektorů



dafuq