

Data mining

UAI/691 Přednáška 1

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda]

- Úvod do oblasti data miningu
- Knowledge Discovery in Databases (KDD)

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

[Motivace]

Data dostupná v
elektronické podobě



Problém

Jak je využít



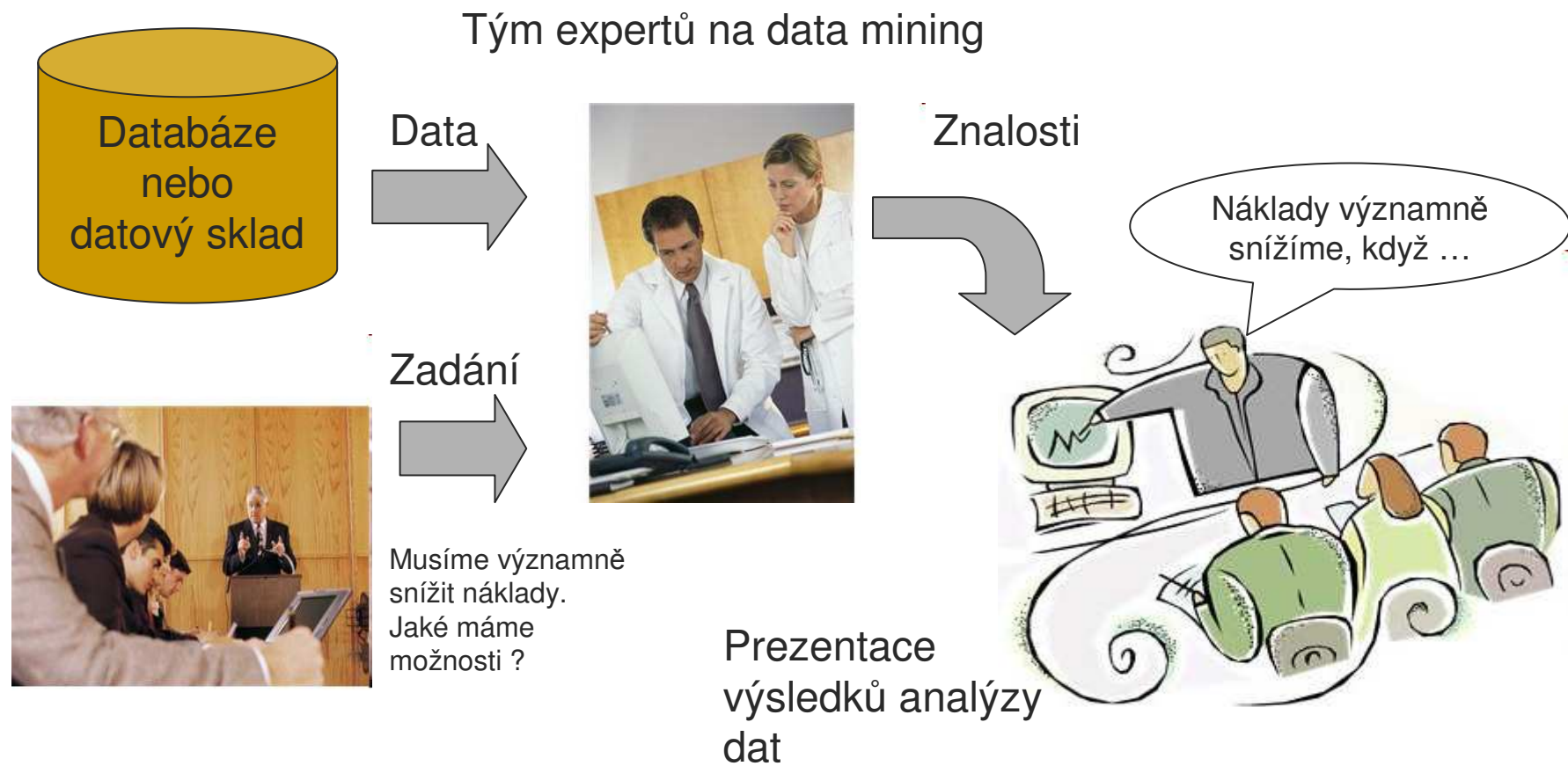
Řešení

Experti na data mining,
to je řešení !



Podniky
banky,
státní správa,
zdravotnictví,
obchodní řetězce,
mobilní operátoři,
poskytovatelé
internetových služeb a další ...

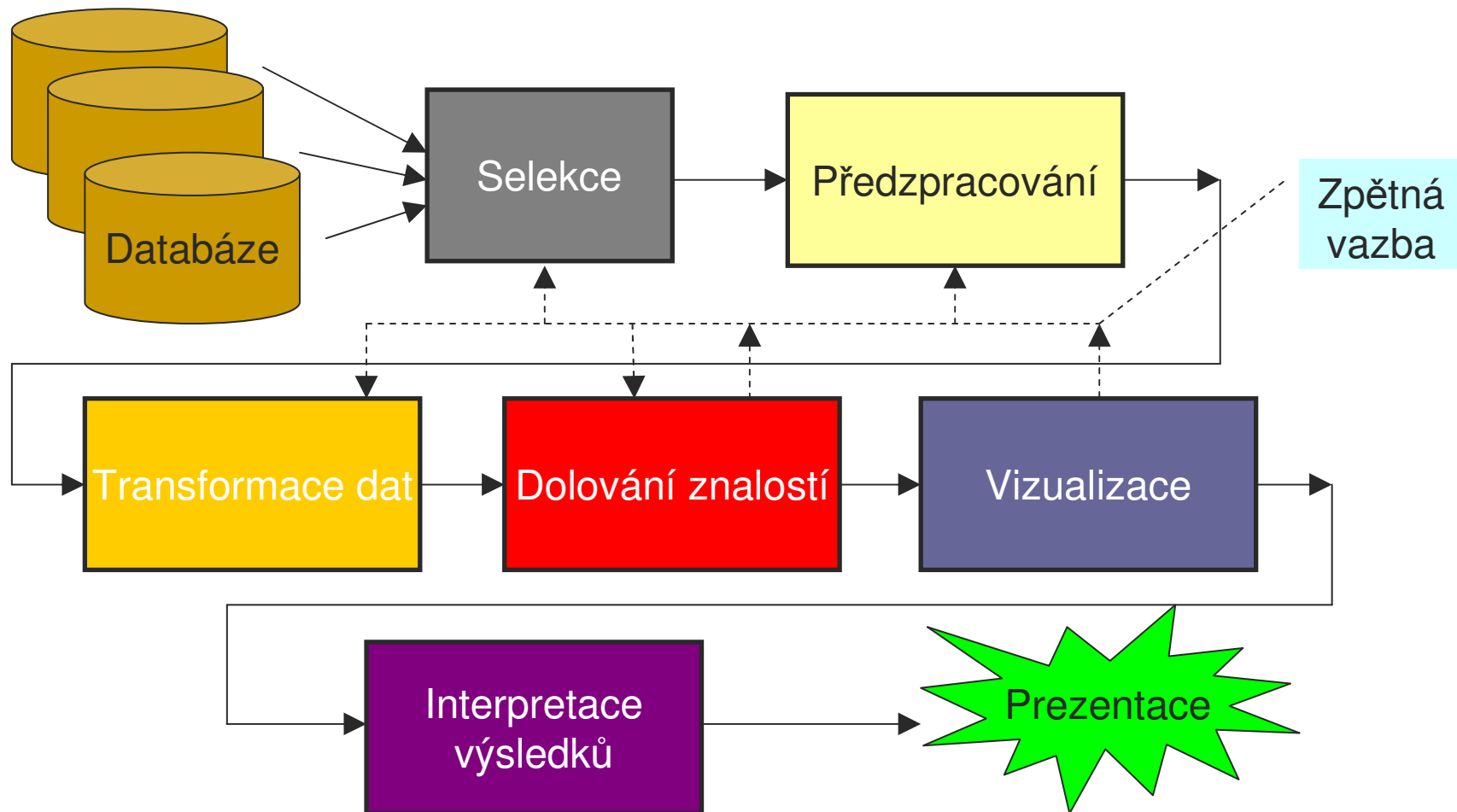
[Data - cenný zdroj informací]



[Knowledge Discovery in Databases (KDD)]

- Dobývání znalostí z databází
- Multi-disciplinární obor zahrnující
 - Databáze
 - Statistiku
 - Umělou inteligenci
- Cílem je automatické vyhledávání zákonitostí v rozsáhlých souborech dat
- V současné době je to proces interaktivní (neobejde se bez experta), současné výzkumy směřují k plné automatizaci

Proces dobývání znalostí



[Selekce dat]

- Výběr relevantní podmnožiny z dostupných dat (relevance má přímou souvislost se zadáním)
- Může být složitý problém
 - Data v různých databázích
 - Data v různých formátech
 - Různý charakter dat (záznamy v databázi, textové dokumenty)
 - Data nelze jednoduše pospojovat do jedné tabulky

[Předzpracování]

- Příprava dat pro další zpracování
- Může zahrnovat
 - Čištění dat od odlehlých hodnot
 - Doplnování chybějících hodnot
 - Agregace dat
 - Extrakci příznaků
 - Detekce závislých atributů
 - Odstranění offsetů a trendů
- Významný krok procesu zpracování, který může významně ovlivnit výsledek analýzy (negativně i pozitivně)

[Transformace dat]

- Nezbytné transformace dat podle potřeb použitých analytických metod
- Může obsahovat
 - Selekcí atributů (feature selection)
 - Vážení atributů (feature ranking)
 - Normalizace atributů
 - Funkční transformace a doplňování atributů vypočtenými hodnotami

[Dolování znalostí]

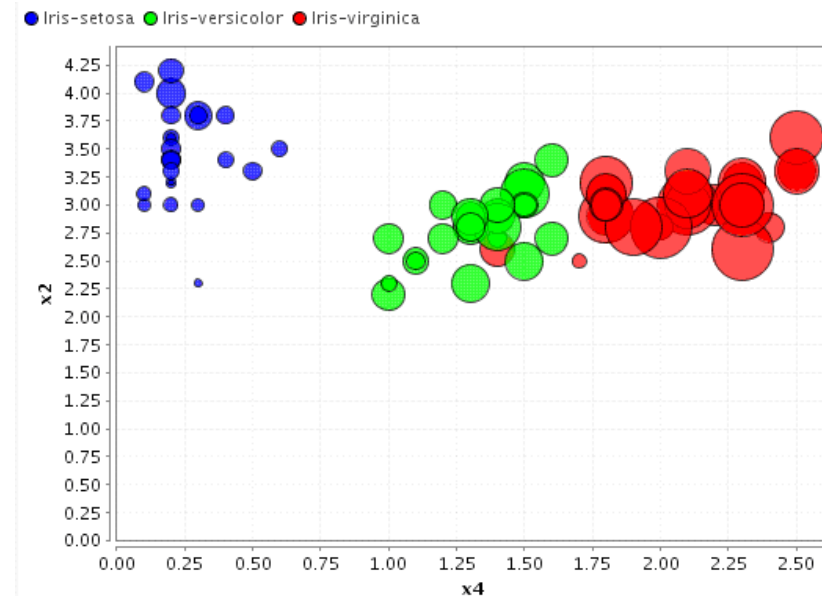
- Využívá metody umělé inteligence, metody založené na strojovém učení
- Využívá metod shlukové analýzy
- Využívá metod modelování a automatické tvorby modelu
- Využívá širokou škálu klasifikátorů

[Dolování znalostí]

- Založeno na
 - Modelování závislostí v datech
 - Klasifikaci dat do tříd
 - shlukové analýze
- Je to iterativní a interaktivní proces, který je řízen expertem
- Současný výzkum směřuje k plné automatizaci tohoto procesu


Vizualizace

- Klíčový nástroj pro interpretaci výsledků
- Využívá širokou škálu grafů
 - Scatter
 - Scatter Matrix
 - Bubble
 - A další
- Řeší problém zobrazení vícerozměrných dat (člověk se přirozeně orientuje pouze v grafech max. 3D)
- Vícerozměrné veličiny různě mapovány např. na tvar, rozměr a barvu objektů



[Interpretace dat a reportování]

- Výsledky analýzy jsou opět čísla, musí se proto převést do srozumitelné řeči (formulace zákonitostí, vizualizace grafy, komentář)
- Při interpretaci výsledků má hlavní slovo expert
- Výstupy analýzy se prezentují ve formě zpráv (reportů)
- Současný výzkum v oblasti směřuje k automatizaci generování reportů



Data mining

UAI/691 Přednáška 2

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda]

- Zdroje dat
- Datová matice a její reprezentace
- Selektce dat z různých zdrojů
- Zpracování dokumentů nebo textových datových souborů

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

[Zdroje dat]

- Databáze (SQL)
- Textové dokumenty
 - Plain text (ASCII, CP1250, ISO8851, ISO8852, UNICODE, UTF-8)
 - HTML, XML
 - Specifické formáty (PDF, RTF, DOC)
- Data v souborech
 - Plain text, CSV
 - XML
 - Specifické formáty (XLS – MS Excel, ODF – OpenOffice Calc)

[Datová matice]

(pojem ze statistiky)

- Základní datová struktura pro uložení dat
- Sloupce se označují jako atributy (proměnné)
- Řádky reprezentují jednotlivé případy. Ve statistice se řádky nazývají jako případy, v data miningu se obvykle označují termínem vzory
- Na vstupu procesu dolování dat se očekává datová matice obsahující relevantní data

[Příklad datové matice]

Atributy

Vzory (případy)

Datum	Příjmení	Jméno	Příchod	Odchod
12.5.2009	Vomáčka	Josef	7:30	16:15
12.5.2009	Novák	Pavel	9:20	14:35
13.5.2009	Vomáčka	Josef	6:15	18:20
13.5.2009	Malá	Jiřina	9:00	16:30

[Metainformace datové matice]

- Názvy atributů (sloupců)
- Datové typy atributů
- Platné hodnoty nominálních atributů
- Statistické údaje charakterizující atributy (střední hodnota, rozptyl, atd.)

Datová matice v textovém formátu

komentář

Plain text, atributy odděleny mezerou

```
#datum příjmení jméno příchod odchod  
12.5.2009 Vomáčka Josef 7:30 16:15  
12.5.2009 Novák Pavel 9:20 14:35  
13.5.2009 Vomáčka Josef 6:15 18:20  
13.5.2009 Malá Jiřina 9:00 16:30
```

Jiné druhy oddělovačů: středník, čárka, tabulátor, svislá čára ..., cokoliv, co se neobjeví v datech

CSV (Comma Separated Values, česká verze, oddělovač středník)

```
datum;prijmeni;jmeno ;prichod;odchod  
12.5.2009;Vomáčka;Josef;7:30;16:15  
12.5.2009;Novák;Pavel;9:20;14:30  
13.5.2009;Vomáčka;Josef;6:15;18:20  
13.5.2009;Malá;Jiřina;9:00;16:30
```

[Datová matice v XML]

(pouze dva řádky)

```
<?xml version="1.0"?>
<Worksheet Name="Datova matice">
  <Table>
    <Row>
      <Cell><Data Type="String">datum</Data></Cell>
      <Cell><Data Type="String">prijmeni</Data></Cell>
      <Cell><Data Type="String">jmeno </Data></Cell>
      <Cell><Data Type="String">prichod</Data></Cell>
      <Cell><Data Type="String">odchod</Data></Cell>
    </Row>
    <Row>
      <Cell><Data Type="Date">2009-05-12</Data></Cell>
      <Cell><Data Type="String">Vomáčka</Data></Cell>
      <Cell><Data Type="String">Josef</Data></Cell>
      <Cell><Data Type="Time">07:30:00.000</Data></Cell>
      <Cell><Data Type="Time">16:15:00.000</Data></Cell>
    </Row>
  </Table>
</Worksheet>
```

XML tag (otevírací)

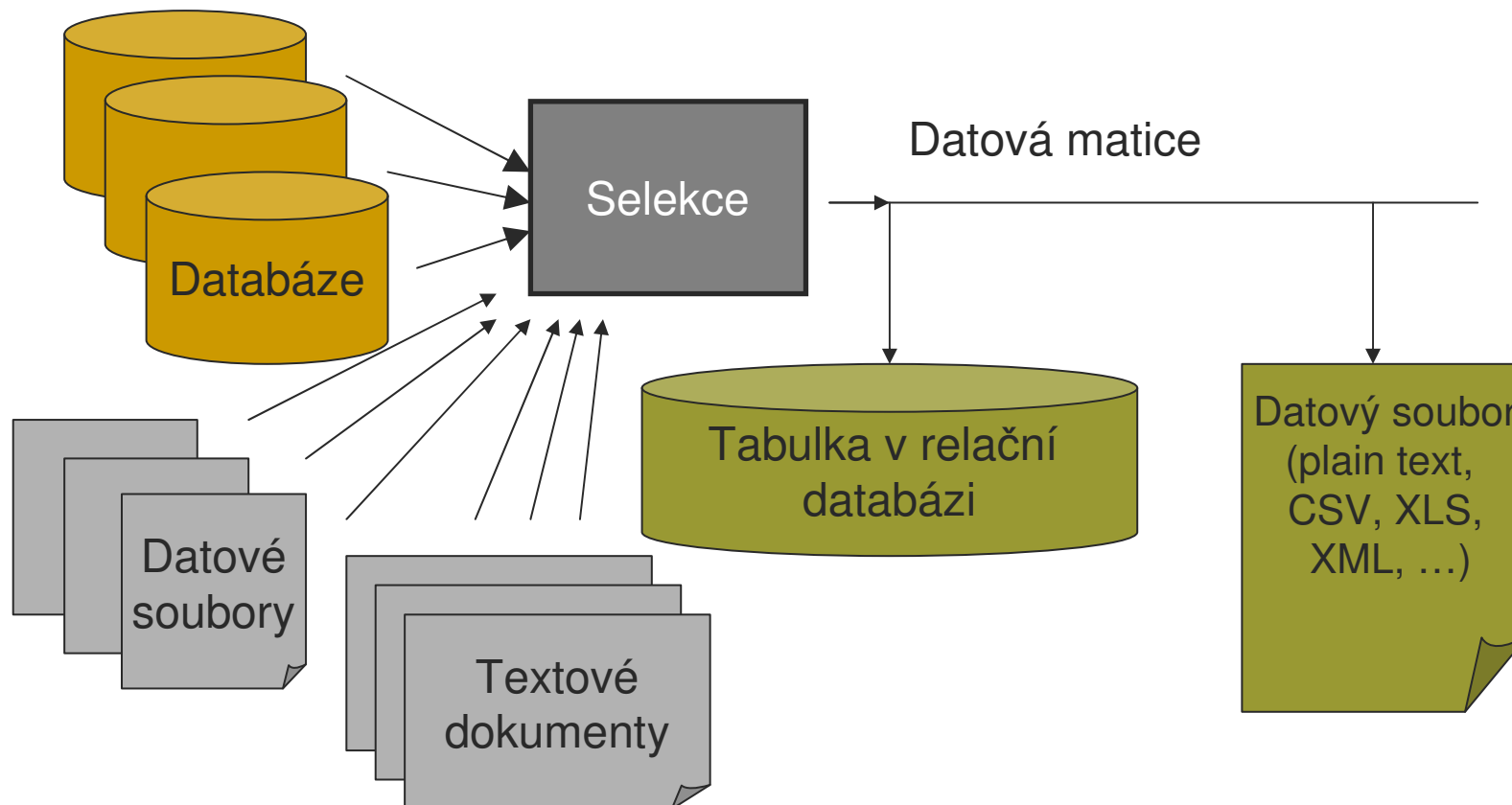
XML atribut

Hodnota
atributu

Hodnota
atributu

XML tag (zavírací)

[Selekce dat]



[Selekce dat]

- Výběr relevantní podmnožiny z dostupných dat (relevance má přímou souvislost se zadáním)
- Může být složitý problém
 - Data v různých databázích
 - Data v různých formátech
 - Různý charakter dat (záznamy v databázi, textové dokumenty)
 - Data nelze jednoduše pospojovat do jedné tabulky

Získání datové matice dotazem v SQL relační databázi

Příkaz pro výběr podmnožiny dat z databáze

Požadované atributy

```
select datum, prijmeni, jmeno, prichod, odchod  
from odchodyprichody  
where datum > {d '2009-11-22' }
```

Pozn: pro složitější databázi s více tabulkami je třeba použít spojování tabulek (join), viz. znalosti z předmětu databáze.

Jméno tabulky
v relační
databázi

Omezení
počtu řádků
je na ty od
data
22.11.2009

Využití příkazů operačního systému (awk, gawk-Linux)

Program awk (gawk) čte textový soubor po řádcích. Každý řádek na základě oddělovače (implicitně mezera) rozseká a jednotlivé segmenty řádku přiřadí v pořadí z leva do prava do proměnných \$1, \$2, Argumentem příkazu je sekvence příkazů, která se opakovaně provede pro každý řádek, a ve které se můžeme odkazovat na jednotlivé proměnné \$1, \$2, ...

Vybere z původního souboru sloupce 1 a 3 a vytiskne je jako dva sloupce v novém souboru

```
awk '{ print $1,$3}' data.txt > datova_malice.txt
```

Sečte čísla v prvním a druhém sloupci a uloží je datové matice (jeden řádek, dvě čísla)

```
awk 'BEGIN{s1=0;s2=0}{s1+=$1;s2+=$2}END{print  
s1,s2}' data.txt > datova_malice.txt
```

Provede se před
zpracováním prvního
řádku

Provede se pro každý řádek

Provede
se po zpracování
posledního řádku

Programové zpracování textových datových souborů v C

fgets() v kombinaci s scanf (sscanf):

fgets přečte řádek, převod na hodnoty zajistí scanf.

Vhodné pro jednoduché, mezerou oddělené atributy.

```
char s[256];  
int rok_naroz;  
char prijmeni[64], jmeno[64];  
// opakuj dokud není konec souboru  
    fgets(s, 256, vstupni_soubor);  
    sscanf(s, "%s %s %d", &prijmeni, &jmeno,  
&rok_naroz);
```

Pro složitější formáty souborů je třeba použít čtení po znacích a použít například lexikální a následně syntaktické analyzátory (flex, bison), viz. předmět Teoretická informatika

Programové zpracování textových datových souborů v Javě

BufferedReader v kombinaci s split, StringTokenizer, StreamTokenizer, Scanner, java.util.regex.Pattern nebo java.util.regex.Matcher :

Třída `BufferedReader` poskytuje funkci pro čtení textového souboru po řádcích (`readLine`). Funkce `split` (třída `String`) rozdělí řetězec do pole řetězců na základě regulárního výrazu (viz. Teroteická informatika). Tokenizery navíc poskytují převody základních datových typů (`int`, `float`, ...) na binární hodnoty.

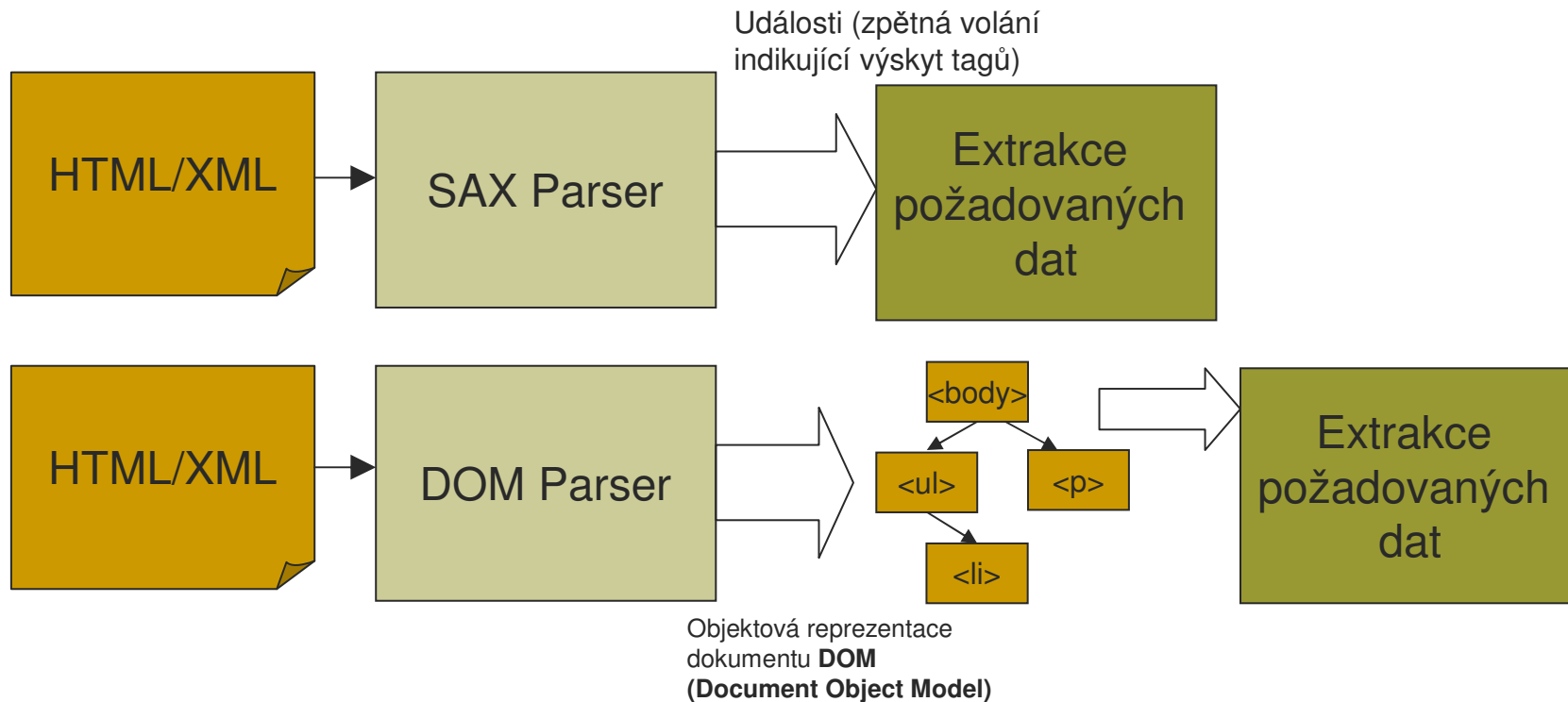
```
BufferedReader br = new BufferedReader(  
    new FileReader("xxx.txt"));  
  
// dokud není konec souboru  
    String line = br.readLine();  
    String[] polozky = line.split(",");
```

Zpracování HTML dokumentů a XML datových souborů

Pro tento typ dokumentů lze doporučit knihovní funkce pro analýzu HTML/XML

Java: třída `javax.xml.parsers.SAXParser` (SAX)

`javax.xml.parsers.DocumentBuilder` (DOM)



[Parsing dokumentu]

```
DocumentBuilderFactory dbf = DocumentBuilderFactory.newInstance();
try {
    dbf.setNamespaceAware(false);
    dbf.setValidating(false);
    dbf.setFeature("http://xml.org/sax/features/namespaces", false);
    dbf.setFeature("http://xml.org/sax/features/validation", false);
    dbf.setFeature(
        "http://apache.org/xml/features/nonvalidating/load-dtd-grammar", false);
    dbf.setFeature(
        "http://apache.org/xml/features/nonvalidating/load-external-dtd", false);

    DocumentBuilder db = dbf.newDocumentBuilder();

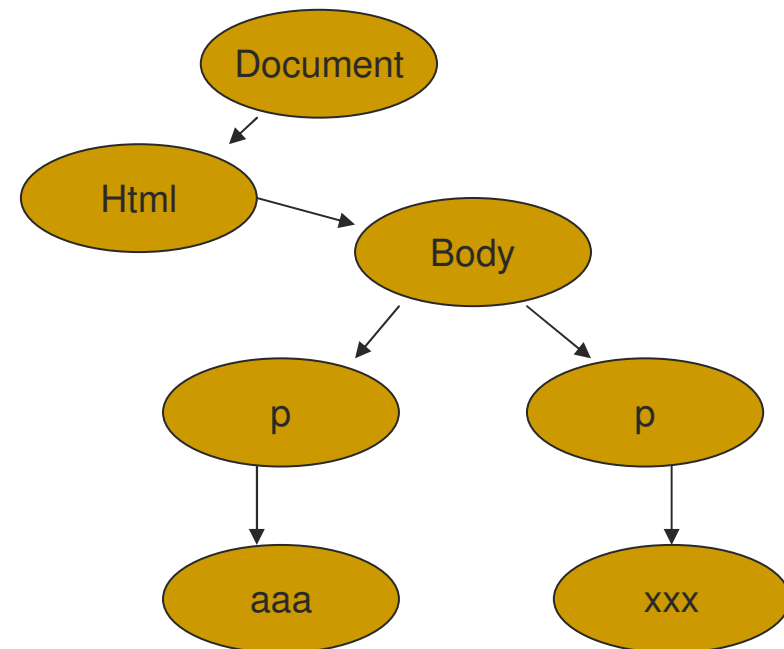
    doc = db.parse(new BufferedInputStream(new FileInputStream(inpf), 1024));
    ... zpracování dat ...
} catch (Exception ex) {
}
```

[DOM]

DOM má stromovou strukturu a skládá se z uzlů (Node) a hran (odkazy na uzly).

Uzly odpovídají HTML tagům

```
<html>  
  <body>  
    <p>aaa</p>  
    <a>xxx</a>  
  </body>  
</html>
```



[Vypis struktury dokumentu]

Rekurzivní metoda pro výpis DOMu

Výpis


```
private static void printDOM(String prefix,
    Node node, PrintStream out) {

    out.println(prefix + node.getNodeName() + " [" +
        ((node.getNodeValue() != null) ?
            node.getNodeValue().trim() : "") + "]" );
    NodeList nodes = node.getChildNodes();
    for(int i = 0; i < nodes.getLength(); i++) {
        printDOM(prefix + "  ", nodes.item(i),
out);
    }
}
```

```
#document []
html []
  #text []
  body []
    #text []
    p []
      #text [aaa]
    #text []
    a []
      #text [xxx]
    #text []
  #text []
```

Nalezení specifického uzlu

```
NodeList nodes = getElementsByTagName("html");
```



Data mining

UAI/691 Přednáška 3

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda

- Statistické metody

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- Jan Hendl: Přehled statistických metod zpracování dat. 2 vydání. Portál, Praha 2006
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010.http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

Střední hodnota náhodné veličiny

Střední hodnota náhodné veličiny X se označuje $E(X)$ nebo μ .

Pro diskrétní
náhodnou veličinu

$$E(X) = \sum_{i=1}^N x \cdot p(x)$$

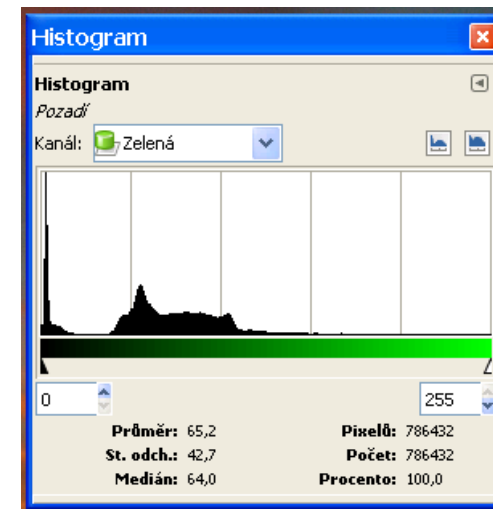
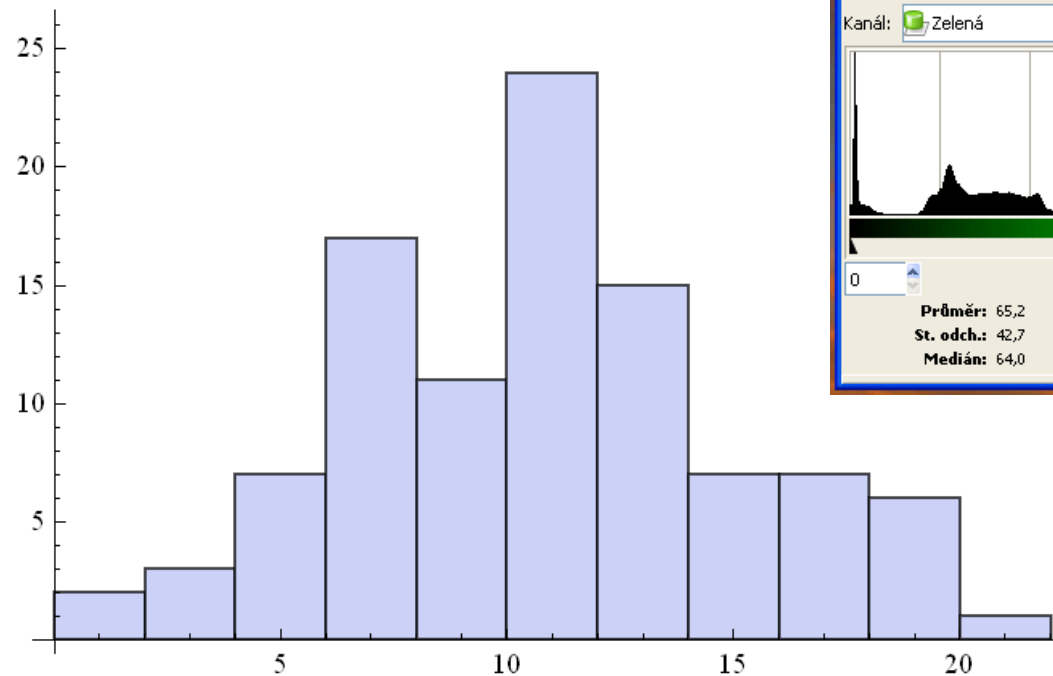
Pro spojitou náhodnou
veličinu

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Hustota
pravděpodobnosti

[Histogram]

Četnost
(nebo
relativní
četnost)



[Konstrukce histogramu]

Pro diskrétní hodnoty spočteme počty výskytu jednotlivých hodnot (četností) v souboru dat, případně vypočteme relativní četnosti tj. poměr četnosti k celkovému objemu dat.

Pro spojité náhodné veličiny stanovíme nejprve intervaly a pak počítáme četnosti hodnot spadajících do daného intervalu. Obdobně jako u diskrétní náhodné veličiny počítáme i relativní četnosti.

Příklad. Je-li náhodná veličina v rozsahu 0-5, stanovíme například intervaly $x \leq 0.5$, $0.5 < x \leq 1$, $1 < x \leq 1.5$, ..., $x > 4.5$. Počet intervalů stanovíme s ohledem na objem dat.

[Popisné statistiky]

- Velké objemy dat lze redukovat, nahrazujeme-li některé množiny nebo podmnožiny dat popisnými statistikami
- Popisná statistika je číselná charakteristika, která popisuje určitý aspekt dat
- Velmi často se užívají
 - Míry centrální tendence (nebo také jinak střední hodnoty, míry střední hodnoty a míry polohy)
 - Míry rozptýlenosti
 - Šiknost, špičatost a další
- Popisné charakteristiky mají silnou vazbu na histogram a de-fakto popisují jeho tvar

Míry centrální tendence (střední hodnoty)

Aritmetický průměr

$$\bar{x} = \frac{1}{n} \sum_{i=1}^N x_i$$

Medián

$$m : P(x \leq m) = 0,5 \wedge P(x \geq m) = 0,5$$

Výpočet: posloupnost čísel seřídíme vzestupně dle hodnoty. Medián m je hodnota, která leží uprostřed seříděné posloupnosti. Pokud posloupnost obsahuje lichý počet prvků vypočteme průměrnou hodnotu prvků přilehlých středu posloupnosti.

Modus

hodnota s největší relativní četností, pokud má histogram více vrcholů (multimodální rozdělení), pak se uvádí více hodnot.

[K zamyšlení - úkol]

	Student 1	Student 2	Student 3	Student 4
	1	2	4	1
	1	2	5	5
	1	2	5	2
	4	1	2	3
	1	2	1	4
	2	2	5	1
	5	4	4	5
	1	2	4	3
	1	1	1	1

Prohlédněte si výsledky studentů během semestru a snažte se bez počítání stanovit jeho známku. Svůj odhad slovně zdůvodněte.

[K zamyšlení - výsledek]

	Student 1	Student 2	Student 3	Student 4
	1	2	4	1
	1	2	5	5
	1	2	5	2
	4	1	2	3
	1	2	1	4
	2	2	5	1
	5	4	4	5
	1	2	4	3
	1	1	1	1
Průměr	1,9	2,0	3,4	2,8
Medián	1,0	2,0	4,0	3,0
Modus	1,0	2,0	4,0	1,0

Neoznámená (přepadová) písemka

Tohle asi trojkař (3,4 zaokr. na 3,0) určitě nebude !

Příliš lehká písemka

Není to omyl ? Tohle přece není jedničkář.

Je to opravdu dvojkař, nebo jedničkář, který občas zalajdačí ?

[Kdy užít aritmetický průměr ?]

- Nelze použít pro kategoriální (nominální) data
- Data musí být z určitého číselného intervalu
- Rozdělení dat je symetrické (= histogram je symetrický)
- Data neobsahují výrazně odlehlé hodnoty
- Pokud budou použity statistické testy

[Kdy užít medián ?]

- Množina hodnot, které se v datech nachází musí být minimálně uspořádaná (toto samozřejmě splňují číselné hodnoty, ale mohou to být i kategoriální data, kde je možné stavovit uspořádání čísla např. bot, oděvů S, M, L, XL, XXL).
- Chceme znát střed rozdělení dat
- Pokud data obsahují odlehlé hodnoty
- Pokud je rozdělení dat silně zešikmené

[Kdy užít modus ?]

- Pro multi-modální rozdělení (více vrcholů)
- Pokud nám stačí základní přehled
- Pokud nás právě zajímá nejčastější hodnota

[Míry rozptýlenosti]

- Míry rozptýlenosti charakterizují jak jsou data rozptýlena
- Příklad: nejlepším sportovním střelcem je ten, který má střední hodnotu zásahu ve středu terče a malý rozptyl střelby (tj. všechny zásahy v ploše desítky, případně devítky). Střelec, který má sice střední hodnotu ve středu terče, ale zásahy rozptýleny po celé ploše terče tedy i v bílých polích, jistě nevyhraje.
- Nejjednodušší charakteristikou je varianční rozpětí $R = x_{\max} - x_{\min}$, ale které je silně citlivé na odlehlé hodnoty

Rozptyl a směrodatná odchylka

Rozptyl

$$\sigma^2 = D(X) = E(X - E(X))^2$$

Rozptyl (základního souboru, populaci)

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Výběrový rozptyl

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Výběrová směrodatná odchylka

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Kdy použít rozptyl nebo směrodatnou odchylku

- Použít za stejných podmínek jako aritmetický průměr
- Obojí je citlivé na odlehlá data
- Nevhodné pro silně zešikmená rozdělení

[Varianční koeficient]

$$VK = \frac{s}{\bar{x}}$$

Vhodný pro porovnání například různých měření s různými průměry, kdy lze předpokládat, že se rozptyl roste lineárně se střední hodnotou veličiny.

[Empirický kvantil]

Empirický kvantil je hodnota, pod kterou leží určité procento údajů

$$x_q; 0 < q < 1$$

Empirický
kvantil

Hladina

Příklad: $x_{0,3}=150$ cm, což znamená, že 30% žáků naší školy je menších než 150 cm.

Podobně jako u modusu je podmínka uspořádanosti množiny hodnot.

[Specifiké kvantily]

- Q_I dolní kvartil $q=0.25$ (25%)
- Q_{II} medián $q=0.5$ (50%)
- Q_{III} horní kvartil $q=0.75$ (75%)
- Percentily okrajů rozdělení
 - $q=2,5\%$ nebo $q=97,5\%$
 - $q=5\%$ nebo $q=95\%$

[Mezikvartilové rozpětí]

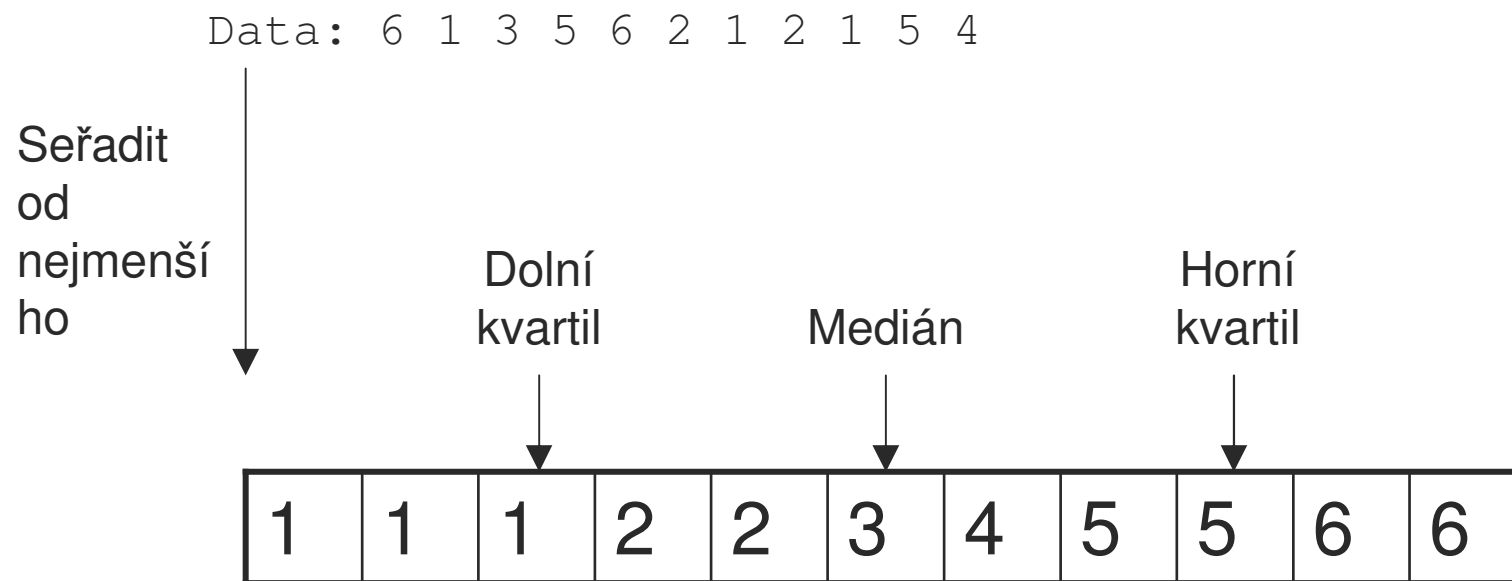
$$Q = Q_{III} - Q_I$$

Určíme horní kvartil $Q_{III} = x_{0,75}$ a dolní kvartil $Q_I = x_{0,25}$ a hodnoty odečteme.

Na rozdíl od směrodatné odchylky není mezikvartilové rozpětí citlivé na odlehlé hodnoty. To znamená, že použijeme-li medián na místo aritmetického průměru, tak můžeme použít mezikvartilové rozpětí místo směrodatné odchylky.

Mezikvartilové rozpětí říká, se v intervalu nachází 50% všech hodnot.

[Kvartily a medián prakticky]



[Centrální momenty]

$$m_k = E(X - E[X])^k$$

$$m_k = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})^k]$$

Centrální momenty
charakterizují tvar
rozdělení
pravděpodobnosti.

a pro $\bar{x} = 0$

Rozptyl

$$m_1 = 0$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n x_i^2$$

$$m_3 = \frac{1}{n} \sum_{i=1}^n x_i^3$$

$$m_4 = \frac{1}{n} \sum_{i=1}^n x_i^4$$

[Šikmost, špičatost]

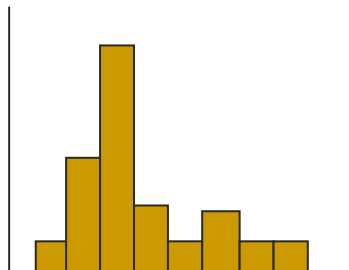
Šikmost

$$\gamma_1 = \frac{m_3}{\sigma^3}$$

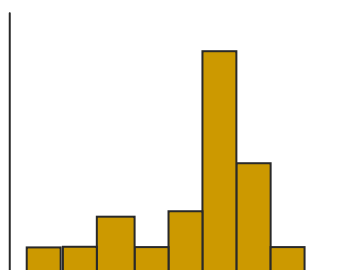
Špičatost

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

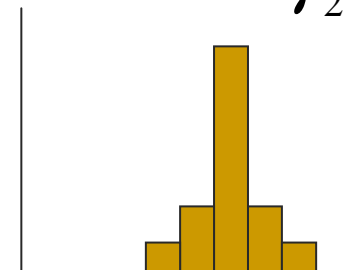
$\gamma_1 > 0$



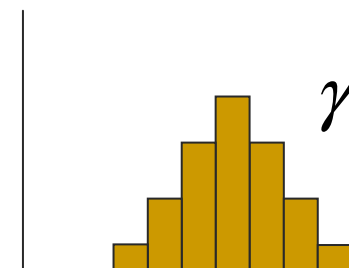
$\gamma_1 < 0$



$\gamma_2 > 0$



$\gamma_2 < 0$



Pozn.: referencí pro špičatost je normální rozdělení. Pro kladné hodnoty je špičatější, pro záporné méně špičaté než normální rozdělení

[Šikmost a špičatost v Excelu]

Vzorečky na tomto slides nemusíte umět zpaměti

Funkce SKEW (Šikmost)

$$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Funkce KURT (Špičatost)

$$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Srovnejte s předchozími vztahy pro velká n

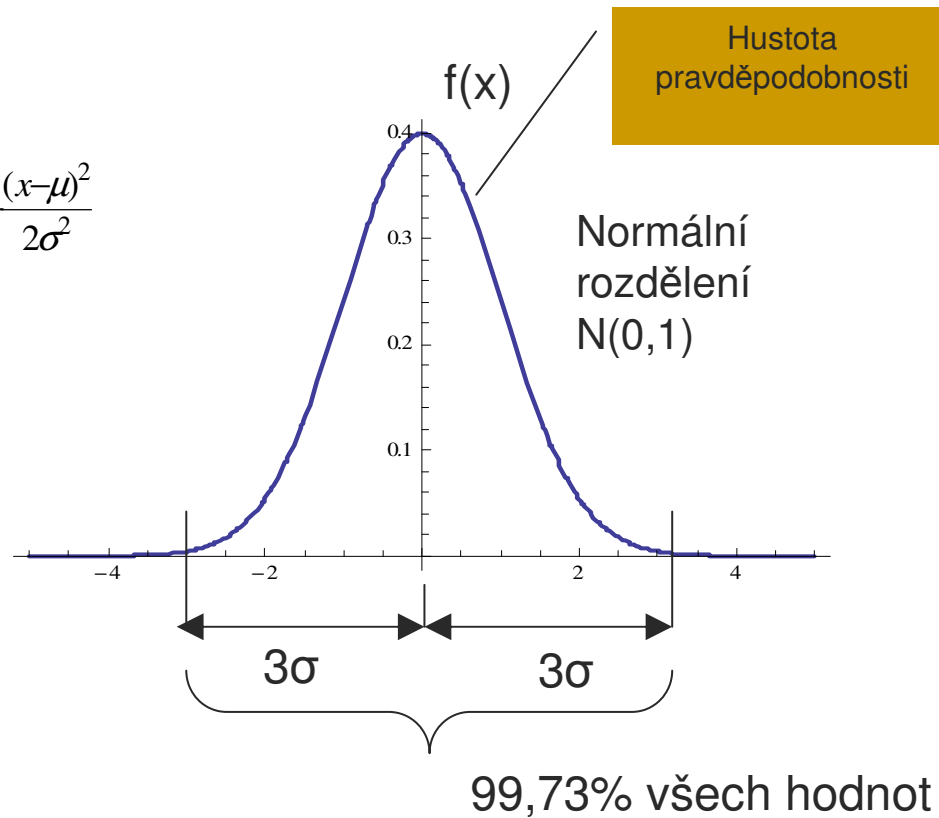
[Použití momentů]

- Momenty charakterizují rozdělení pravděpodobnosti diskrétní nebo spojitě náhodné veličiny.
- Pokud máme rozsáhlý soubor dat, tak jej můžeme vizualizovat v podobě histogramu.
- Pokud ale data chceme dále číselně zpracovávat, můžeme celý soubor dat nahradit centrálními momenty a tím dosáhnout značné redukce dat pro následné zpracování.
- Příklad: v rozpoznávání se centrální momenty používají jako charakteristiky tvaru objektu. Na základě těchto charakteristik se objekt rozpoznává. Centrální momenty jsou invariantní vůči posunutí.

[Normální rozdění]

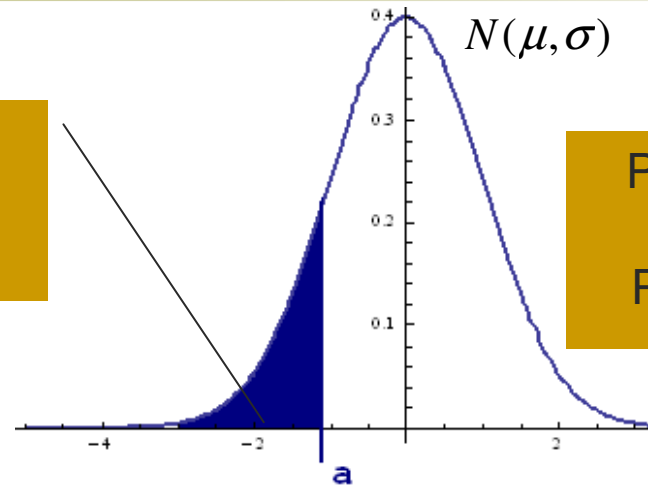
$$N[\mu, \sigma] = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normální rozdění patří mezi spojitá rozdění

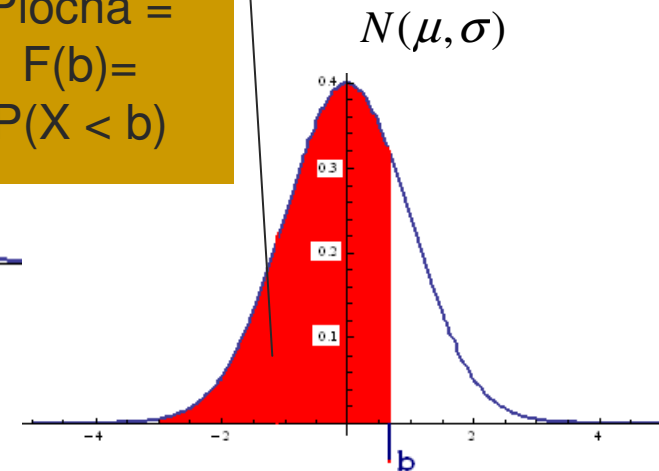


Konstrukce histogramu pro spojité veličiny s normálním rozdělením

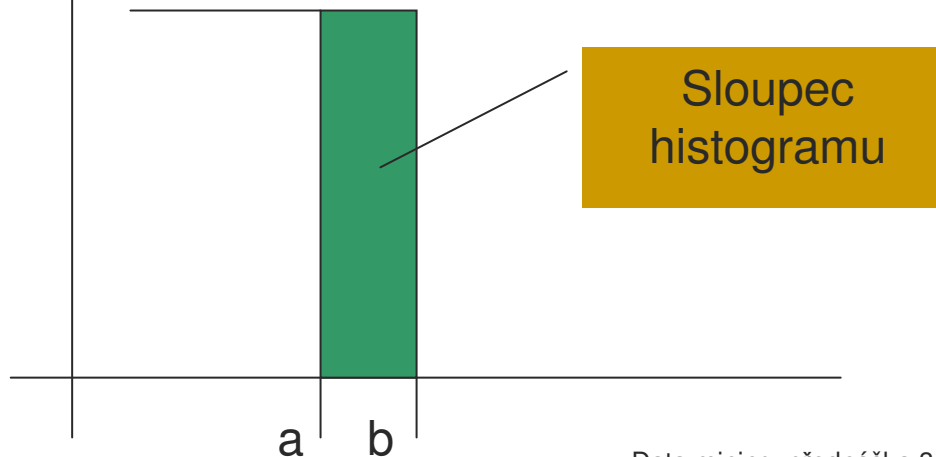
Plocha =
 $F(a) = P(X < a)$



Plocha =
 $F(b) = P(X < b)$



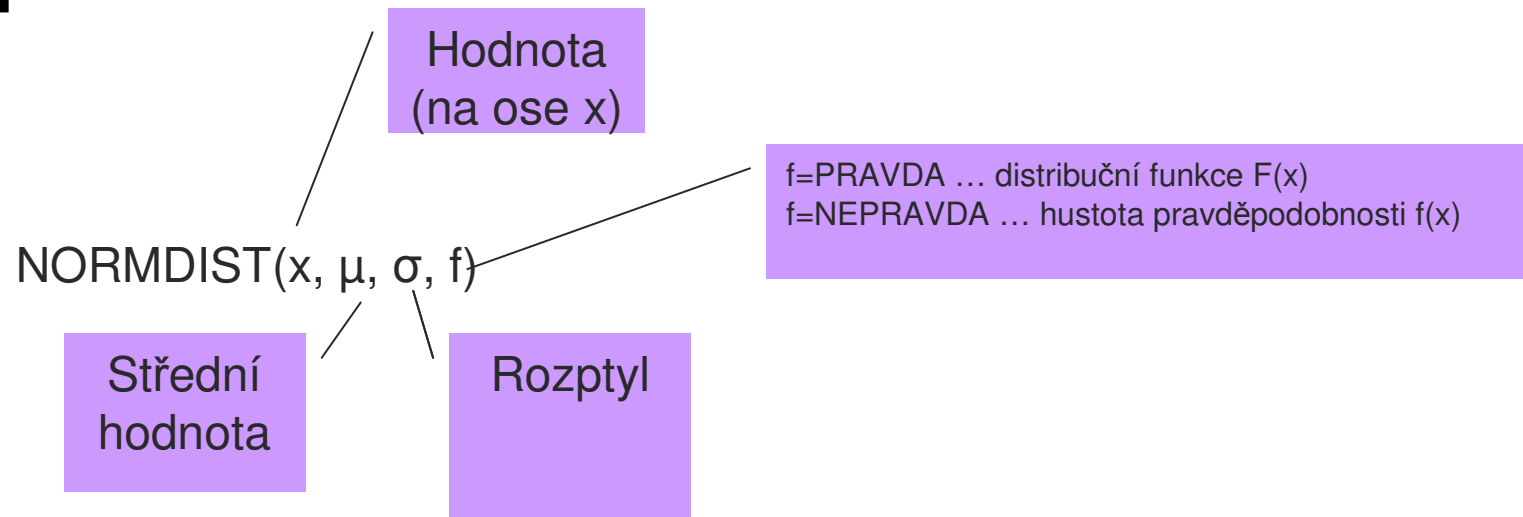
$P(a \leq X \leq b) = F(b) - F(a)$



$F(x)$ je tzv. distribuční funkce

$$f(x) = \frac{dF(x)}{dx}$$

[Normální rozdělení v Excelu]



$$P(a \leq X \leq b) = \text{NORMDIST}(b, 0, 1, \text{PRAVDA}) - \text{NORMDIST}(a, 0, 1, \text{PRAVDA})$$

Vypočteno pro normální rozdělení $N(0,1)$

[Korelace]

$$\rho_{x,y} = \frac{E(X.Y) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2} \sqrt{E(Y^2) - (E(Y))^2}}$$

Korelace vyjadřuje míru závislosti dvou náhodných veličin.

Pro statisticky nezávislé veličiny je korelace rovna nule.
Mluvíme o veličinách, které nejsou korelované.

Pozor ! korelace reflektuje pouze lineární vztah mezi veličinami

Pro nulové střední hodnoty

$$\rho_{x,y} = \frac{E(X.Y)}{\sqrt{E(X^2)} \sqrt{E(Y^2)}}$$

Výpočet korelace

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n a_i$$

$$\bar{B} = \frac{1}{n} \sum_{i=1}^n b_i$$

A	B	$A - \bar{A}$	$B - \bar{B}$	$(A - \bar{A})(B - \bar{B})$
1	5	-1,5	-0,25	0,375
2	-4	-0,5	-9,25	4,625
3	9	0,5	3,75	1,875
4	11	1,5	5,75	8,625

$$\rho_{AB} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^n (a_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (b_i - \bar{B})^2}}$$

Korelace v Excelu

	A	B	C
1	0,640632	-0,0781	-0,07431
	0,849731	0,549193	0,164519
	0,235557	-1,29333	0,000272
	0,478534	-0,5644	-0,02263
	0,987972	0,963917	-0,03496
	0,618721	-0,14384	0,23606
	0,545767	-0,3627	-0,30772
	0,916829	0,750488	0,380135
	0,544866	-0,3654	0,075874
	0,230149	-1,30955	-0,4638
	0,382359	-0,85292	-0,00105
	0,857963	0,573889	-0,43948
13	0,689791	0,069373	0,472912

$$\rho_{AB}=1$$

Korelované

$$\rho_{BC}=0,3$$

Nekorelované

$$\rho_{AC}=0,3$$

Nekorelované

Poměrně vysoká hodnota korelace 0,3 u nekorelovaných atributů je způsobena malým vzorkem dat.

Příklad funkce v Excelu `CORREL(A1:A13;B1:B13)`

[Autokorelace]

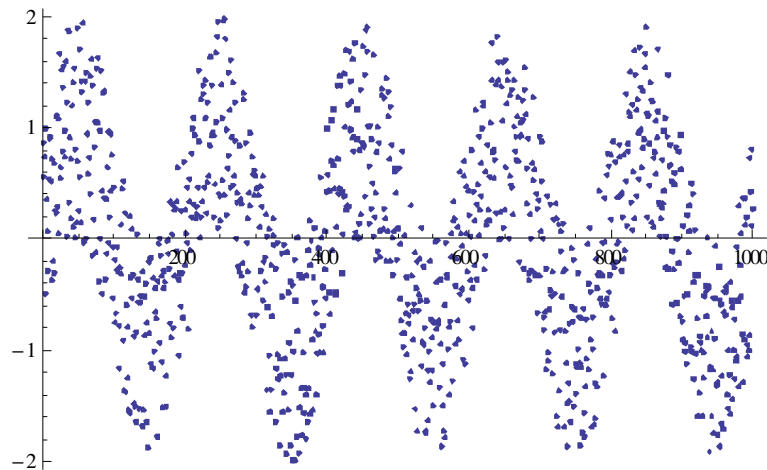
Autokorelace

$$R(\tau) = \frac{E((X_t - E(X))(X_{t+\tau} - E(X)))}{\sigma^2}$$

Umožní identifikovat periodické děje v datech nebo signálu

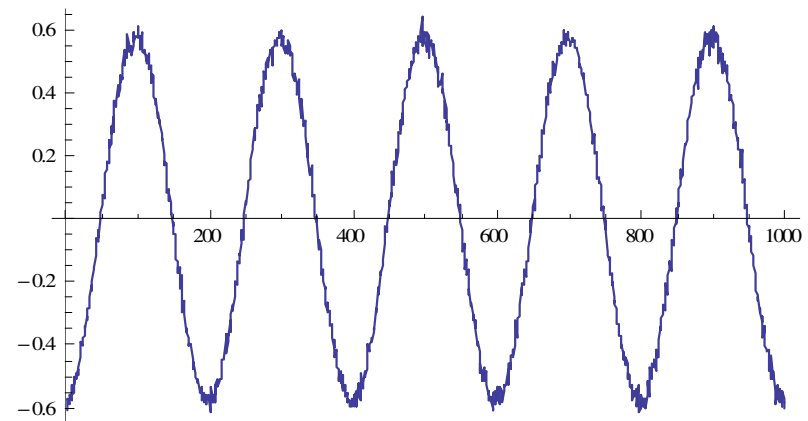
Periodický děj pro určité τ se projeví kladnou nebo zápornou hodnotou na grafu $R(\tau)$. Pokud perioda není přítomna, je $R(\tau)$ blízké nule.

Aplikace autokorelace



Vstupní
zašuměná data
s periodickou
složkou

Autokorelační
funkce $R(\tau)$



Kontingenční tabulka

Nominální atributy

X	Y
0	A
0	N
1	A
1	A
0	N
0	N
0	A
0	N
1	N
0	N

Tabulka obsahuje četnosti

	A	N	
0	2	5	7
1	2	1	3
Celkem	4	6	10

Vyhodnocení kontingenční tabulky dává odpověď zda jsou X a Y statisticky nezávislé.

Této tabulce se říká čtyřpolní (má 4 pole)

Kontingenční tabulka - vyhodnocení

Pro nezávislé X a Y
platí, že

$$a_{ij} = \frac{r_i s_j}{n}$$

Reálná data se budou
lišit, proto počítáme
chybu

$$\chi^2 = \sum_{i=1}^{N_r} \sum_{j=1}^{N_s} \frac{(a_{ij} - \frac{r_i s_j}{n})^2}{\frac{r_i s_j}{n}}$$

Chí-kvadrát

	A	N	
0	20 <small>a₁₁</small>	50 <small>a₁₂</small>	70 <small>r₁</small>
1	20 <small>a₂₁</small>	10 <small>a₂₂</small>	30 <small>r₂</small>
Celkem	40 <small>s₁</small>	60 <small>s₂</small>	100

V tabulce určíme nebo funkcí v Excelu spočítáme hodnotu rozdělení chí-kvadrát pro požadovanou hladinu významnosti a $(N_r-1)(N_s-1)$ stupni volnosti. Pokud je vypočtená hodnota větší, hypotézu o nezávislosti zamítneme. Funkce CHITEST.

V Excelu lze celý test svěřit funkci CHITEST

Tento test je použitelný, pokud pro všechna i, j platí $r_i * s_j / n \geq 5$, jinak se doporučuje Fischerův test [Berka, 2003]

Kontingenční tabulka – vyhodnocení

Skutečné hodnoty

	A	N	
0	20	50	70
1	20	10	30
Celkem	40	60	100

Očekávané hodnoty pro nezávislost

	A	N	
0	28	42	70
1	12	18	30
Celkem	40	60	100

Hodnoty


$$\frac{(a_{ij} - \frac{r_i s_j}{n})^2}{\frac{r_i s_j}{n}}$$

	A	N	
0	2.3	1.5	70
1	5.3	3.5	30
Celkem	40	60	100

$$\chi^2 = 2.3 + 1.5 + 5.3 + 3.5 = 12.6$$

$$\chi^2_{(1)}(0.05) = 3.84 < \chi^2 = 12.6$$

Hypotézu o nezávislosti zamítáme, mezi veličinami je závislost



Data mining

UAI/691 Přednáška 4

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda

- Úvod do programu RapidMiner

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

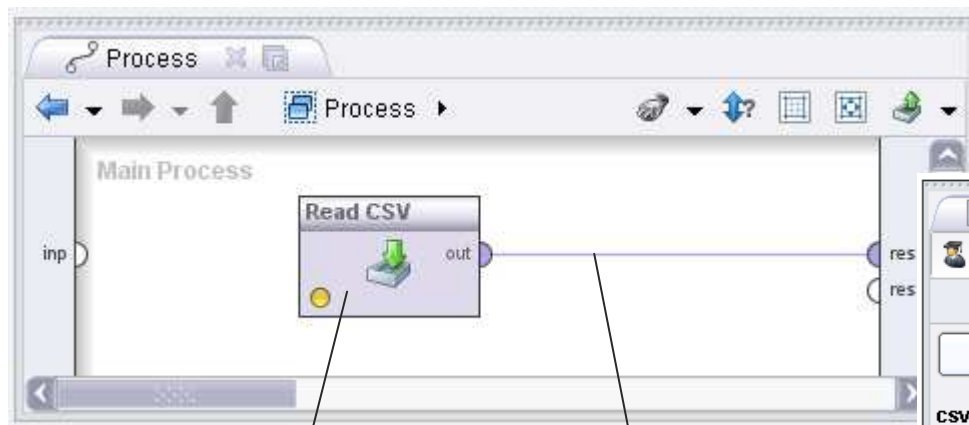
[RapidMiner]

- Nástroj pro zpracování, modelování a vizualizaci dat
- Integruje velké množství algoritmů z oblasti databází, statistiky a umělé inteligence
- Disponuje grafickým designérem pro návrh schémat procesu zpracování dat

[Import dat (CSV)

Soubor CSV

```
x1,x2,x3,x4  
1.3,2,3.3,-11.1  
5.7,4.4,-6,4
```



Operátor
načtení dat z
CSV
souboru

Odeslání
dat do
prohlížeče
výsledků

Parameters

Read CSV

Import Configuration Wizard...

csv file: C:\data\iris.csv

column separators: ,

☐ trim lines

☒ use quotes

quotes character: "

escape character for quotes: \

Data v prohlížeči výsledků

Datová matice
(datová množina)

Meta data

Result Overview ExampleSet (Sample (Stratified))

☐ Meta Data View ☒ Data View ☐ Plot View ☐ Annotations

ExampleSet (15 examples, 1 special attribute, 4 regular attributes)

Row No.	y	x1	x2	x3	x4
1	Iris-setosa	4.700	3.200	1.300	0.200
2	Iris-setosa	4.800	3.400	1.600	0.200
3	Iris-setosa	5.500	4.200	1.400	0.200
4	Iris-setosa	4.900	3.100	1.500	0.100
5	Iris-setosa	4.500	2.300	1.300	0.300
6	Iris-versicolc	6.500	2.800	4.600	1.500
7	Iris-versicolc	5	2	3.500	1

Result Overview ExampleSet (Sample (Stratified)) ExampleSet (Multiply)

☒ Meta Data View ☐ Data View ☐ Plot View ☐ Annotations

ExampleSet (15 examples, 1 special attribute, 4 regular attributes)

Role	Name	Type	Statistics	Range
label	y	polynomial	mode = Iris-setosa (5), least = Iris-setosa (5)	Iris-setosa (5), Iris-versicolor (5), Iris-virginica (5)
regular	x1	real	avg = 5.760 +/- 1.020	[4.500 ; 7.700]
regular	x2	real	avg = 2.887 +/- 0.536	[2.000 ; 4.200]
regular	x3	real	avg = 3.813 +/- 2.037	[1.300 ; 6.900]
regular	x4	real	avg = 1.187 +/- 0.849	[0.100 ; 2.500]

[Typy atributů]

Typ	Popis
nominal	Kategorická proměnná
numeric	Číselné hodnoty
integer	Celočíselné hodnoty
real	Reálná čísla
binominal	Kategorická proměnná se dvěma kategoriemi (zvláštní případ nominal)
polynominal	Kategorická proměnná s více než dvěma kategoriemi (zvláštní případ nominal)
date_time	Časové razítko – datum a čas
date	Datum (pouze)
time	Čas (pouze)

[Role atributů]

Role	Popis
regular	Data (typicky vstupy modelů)
label	Požadovaná požadovaný výstup modelu (odezva modelu)
outlier	Odlehlá hodnota
id	Identifikátor záznamu
weight	váha
cluster	shluk

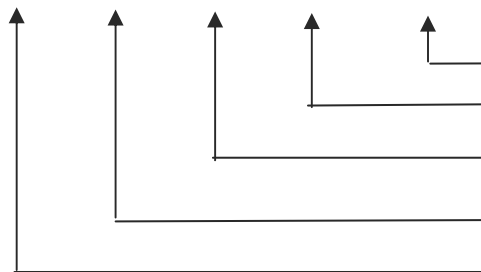
[Datová množina IRIS]

Nejpopulárnější databáze užívaná k testování algoritmů vůbec.

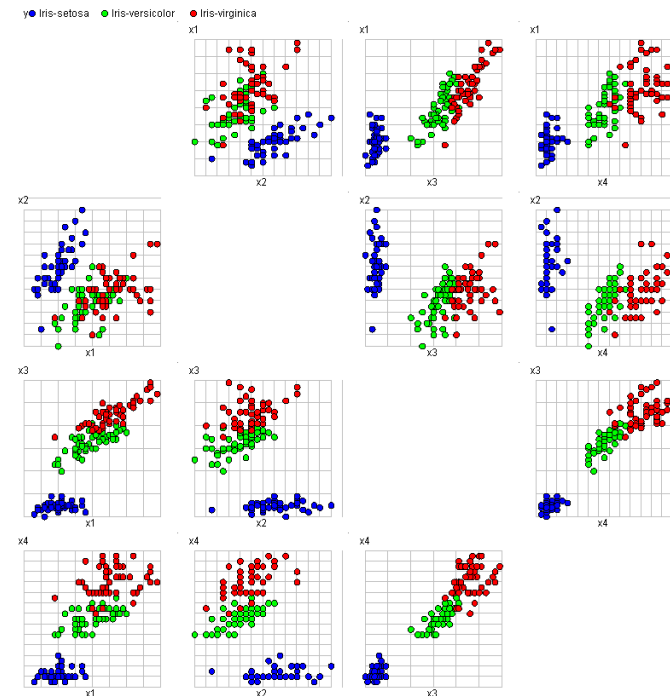
Zdroj dat: UCI database (<http://archive.ics.uci.edu/ml/datasets>)

Fisher, R.A. "The use of multiple measurements in taxonomic problems. Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950)..

5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
7.7, 2.6, 6.9, 2.3, Iris-virginica
6.0, 2.2, 5.0, 1.5, Iris-virginica



Druh kosatce
X4 petal width in cm
X3 petal length in cm
X2 sepal width in cm
X1 sepal length in cm

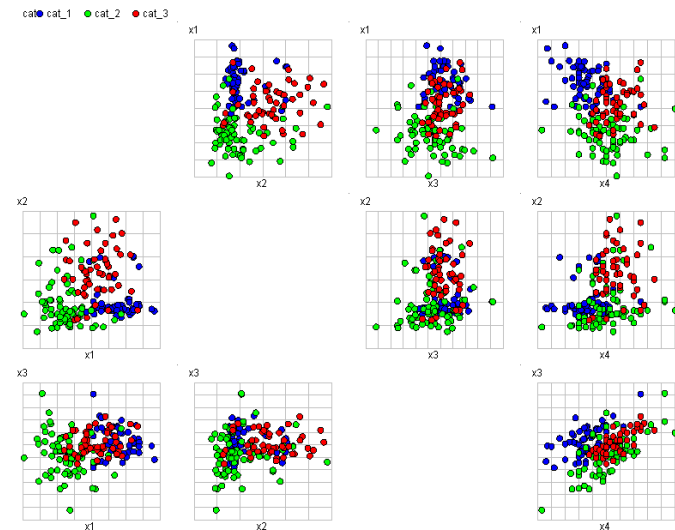


Datová množina - Wine

Zdroj dat: UCI database (<http://archive.ics.uci.edu/ml/datasets>)

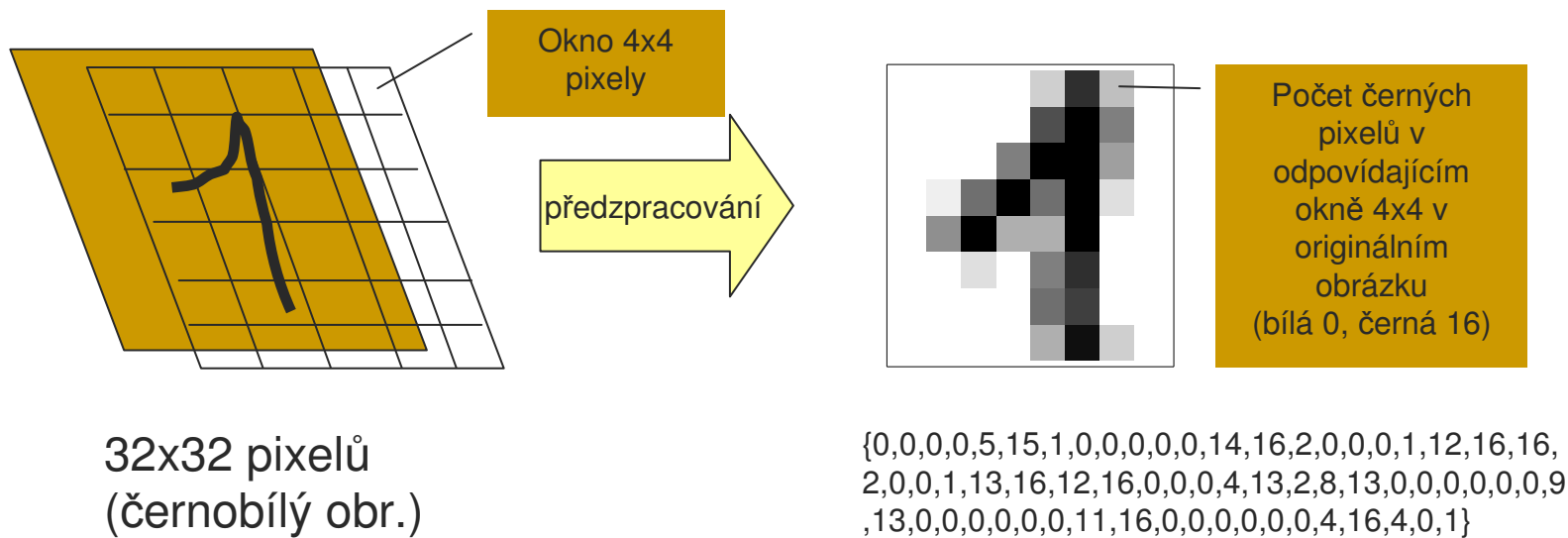
Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.

- 1) Alcohol
- 2) Malic acid (kyselina jablečná)
- 3) Ash (popel)
- 4) Alcalinity of ash (zásaditost popela)
- 5) Magnesium (Hořčík)
- 6) Total phenols (fenoly)
- 7) Flavanoids (flavonoidy, vitamin P)
- 8) Nonflavanoid phenols
- 9) Proanthocyanins (třída flavonoidů)
- 10) Color intensity
- 11) Hue (Odstín)
- 12) OD280/OD315 of diluted wines (zředěná vína)
- 13) Proline (druh aminokyseliny)



[Datová množina - Digits]

Rozpoznávání ručně psaných číslic - praktická aplikace: třídění obálek na poště



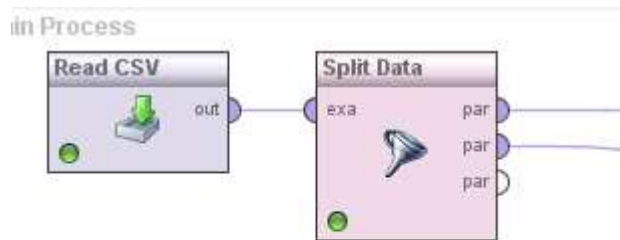
Pro tato můžeme vytvořit klasifikátor

Zdroj dat: UCI database

(<http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits>)

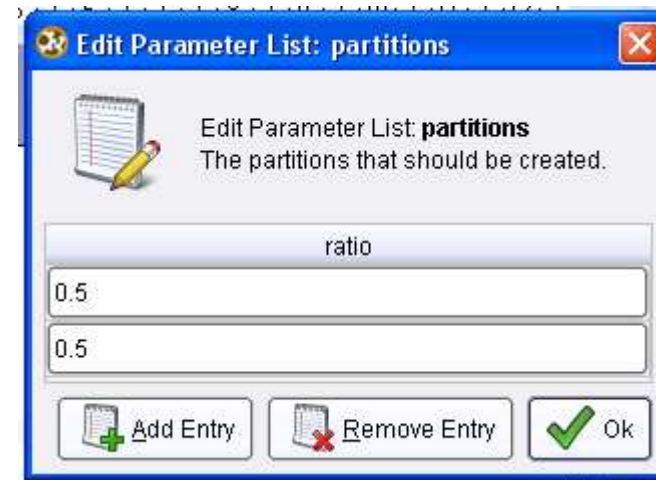
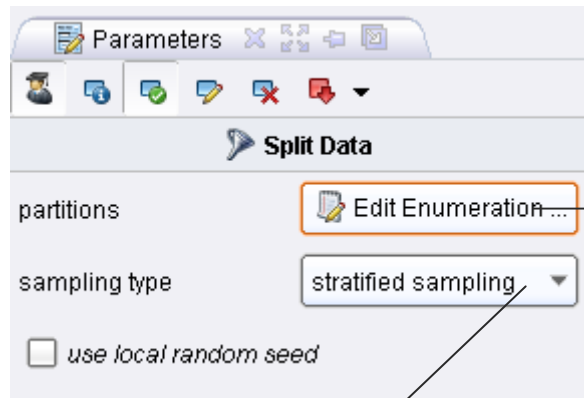
E. Alpaydin, C. Kaynak, Department of Computer Engineering, Bogazici University, 80815 Istanbul Turkey, alpaydin@boun.edu.tr, July 1998

Vytvoření trénovací a testovací množiny




Trénovací
Množina
(ExampleSet)

Testovací
Množina
(ExampleSet)



Způsob výběru vzorů z množiny
(stratified = rovnoměrné zastoupení
vzorů ze všech kategorií)



Data mining

UAI/691 Přednáška 5

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda]

- Normalizace dat
- Měření podobnosti

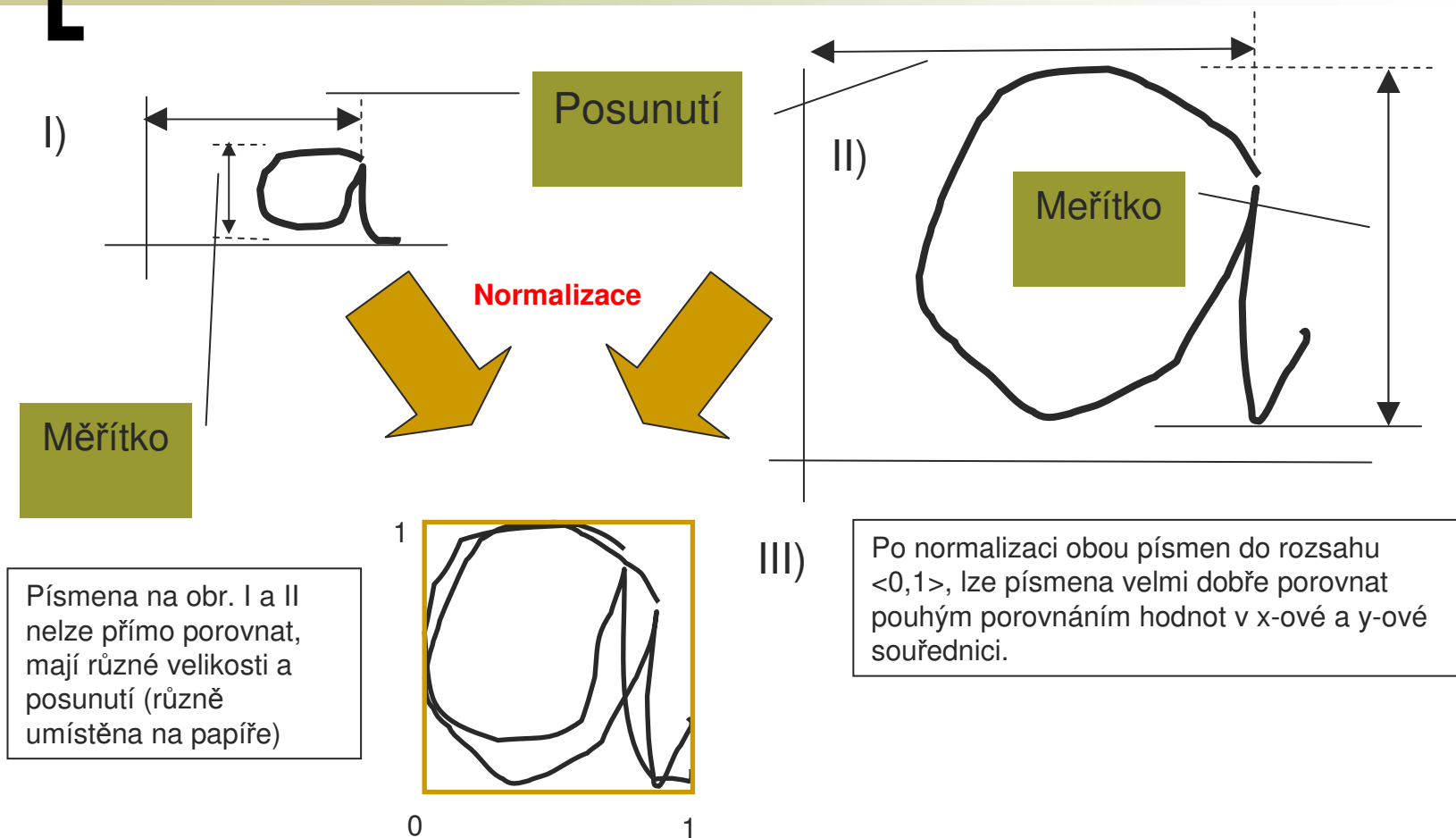
[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

[Normalizace dat]

- Transformace dat do požadované rozsahu
- Zavádí invarianci (nezávislost) vůči
 - Posunutí
 - Měřítku
- Řada algoritmů používaných v dataminingu vyžaduje normalizovaná data

Normalizace - motivace



[Norma min-max]

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} (y_{\max} - y_{\min}) + y_{\min}$$

Kde x je proměnná, kterou normalizujeme (atribut, který normalizujeme) a y je normalizovaná proměnná. x_{\min}/x_{\max} je minimální/maximální hodnota v datech a y_{\min}/y_{\max} je minimální a maximální hodnota normalizované proměnné. y_{\min} a y_{\max} určuje rozsah, do kterého proměnnou x transformujeme.

Příklad: normalizujte do rozsahu $<-1,1>$, víte-li (nebo jste zjistili), že data jsou v rozsahu $<-4, 8>$.

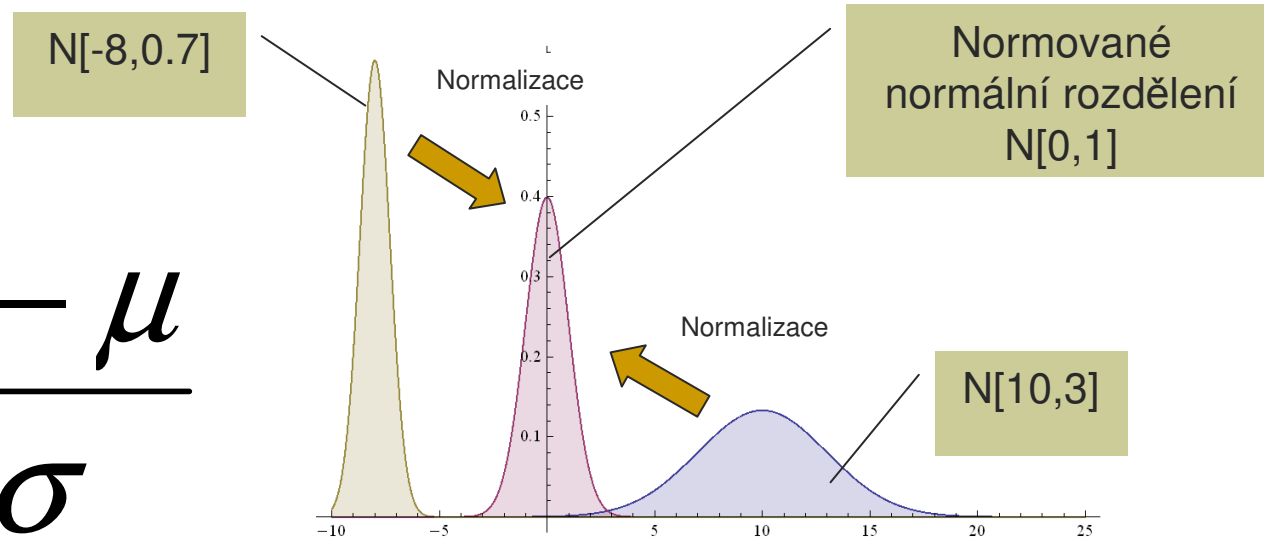
$x_{\min} = -4$, $x_{\max} = 8$, $y_{\min} = -1$, $y_{\max} = 1$

$y = (x - (-4)) / (8 - (-4)) (1 - (-1)) + (-1) = (x + 4) / 12 * 2 = (x + 4) / 6 - 1$

$x = 2, y = 0$; $x = -1, y = -1/2$; $x = 5, y = 1/2$; $x = -4, y = ?$; $x = 8, y = ?$

[Z-scores]

$$z = \frac{x - \mu}{\sigma}$$



Kde x je proměnná, kterou normalizujeme (atribut, který normalizujeme) a z je normalizovaná proměnná. μ je střední hodnota x a σ je rozptyl x . Normalizovaná proměnná z má střední hodnotu nula a rozptyl 1.

[Euclidean norm (L2)]

Používá se pro normalizaci vektorů. Neaplikuje na jednotlivé atributy, tj. sloupce datové matice, jako Min-Max a Z-scores, ale na řádky, a to na všechny nebo na vybranou podmnožinu atributů. Řádek datové matice nebo vybraná podmnožina atributů se považuje za vektor.

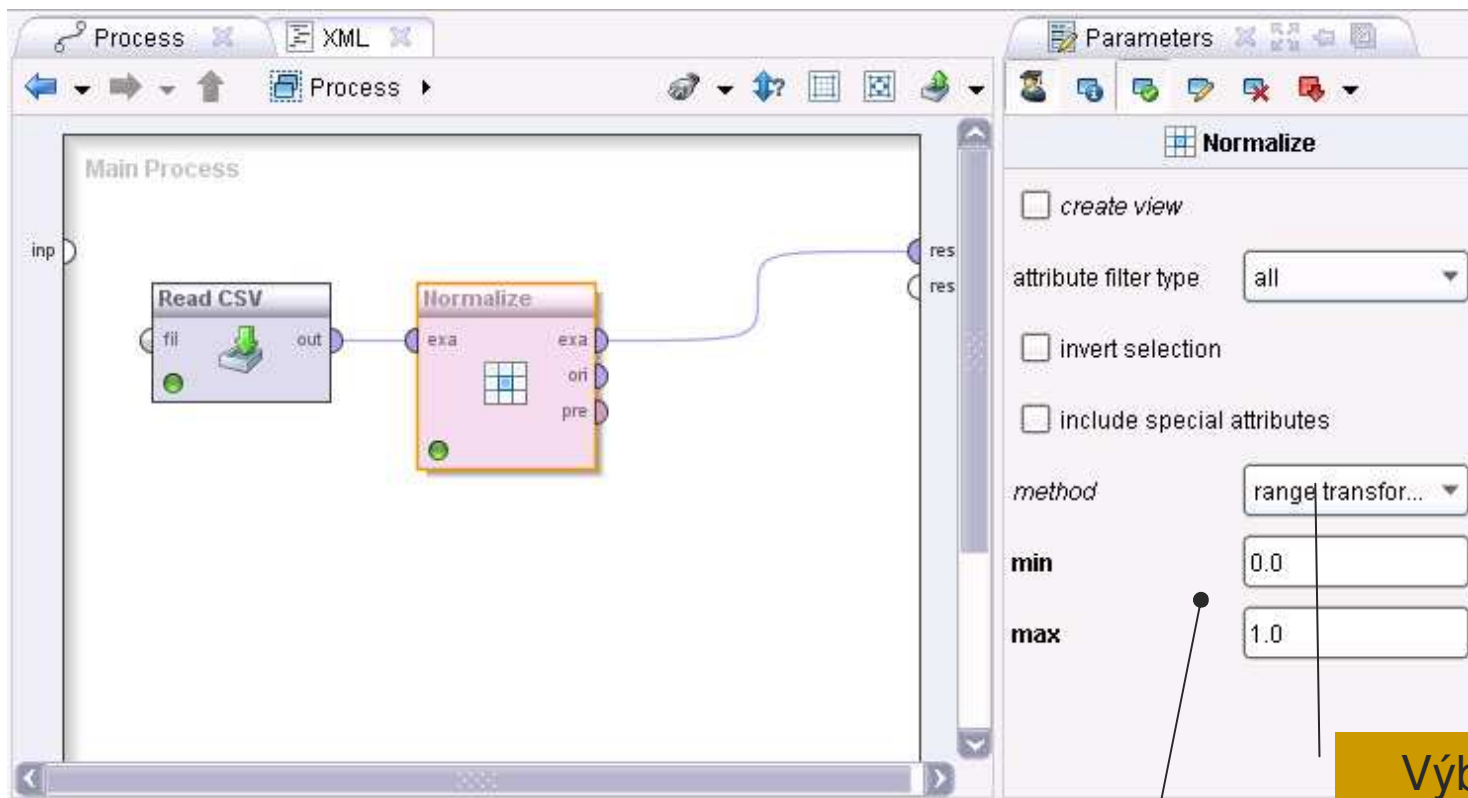
$$\vec{y} = \frac{\vec{x}}{\|\vec{x}\|}, y_i = \frac{x_i}{\sqrt{\sum_{i=1}^N x_i^2}}$$

x1	x2	x3	 x 	y1	y2	y3
1	4	-5	6.48	0.15	0.62	-0.77
-2	3	7	7.87	-0.25	0.38	0.89

Kde x je řádkový vektor v datové matici a y je normalizovaný vektor. Normou je velikost vektoru $\|\vec{x}\|$. N je dimenze vektoru (počet složek).

Normovaný vektor je invariantní (nezávislý) vůči velikosti vektoru,
ale zachovává směr vektoru

Normalizace v RapidMineru

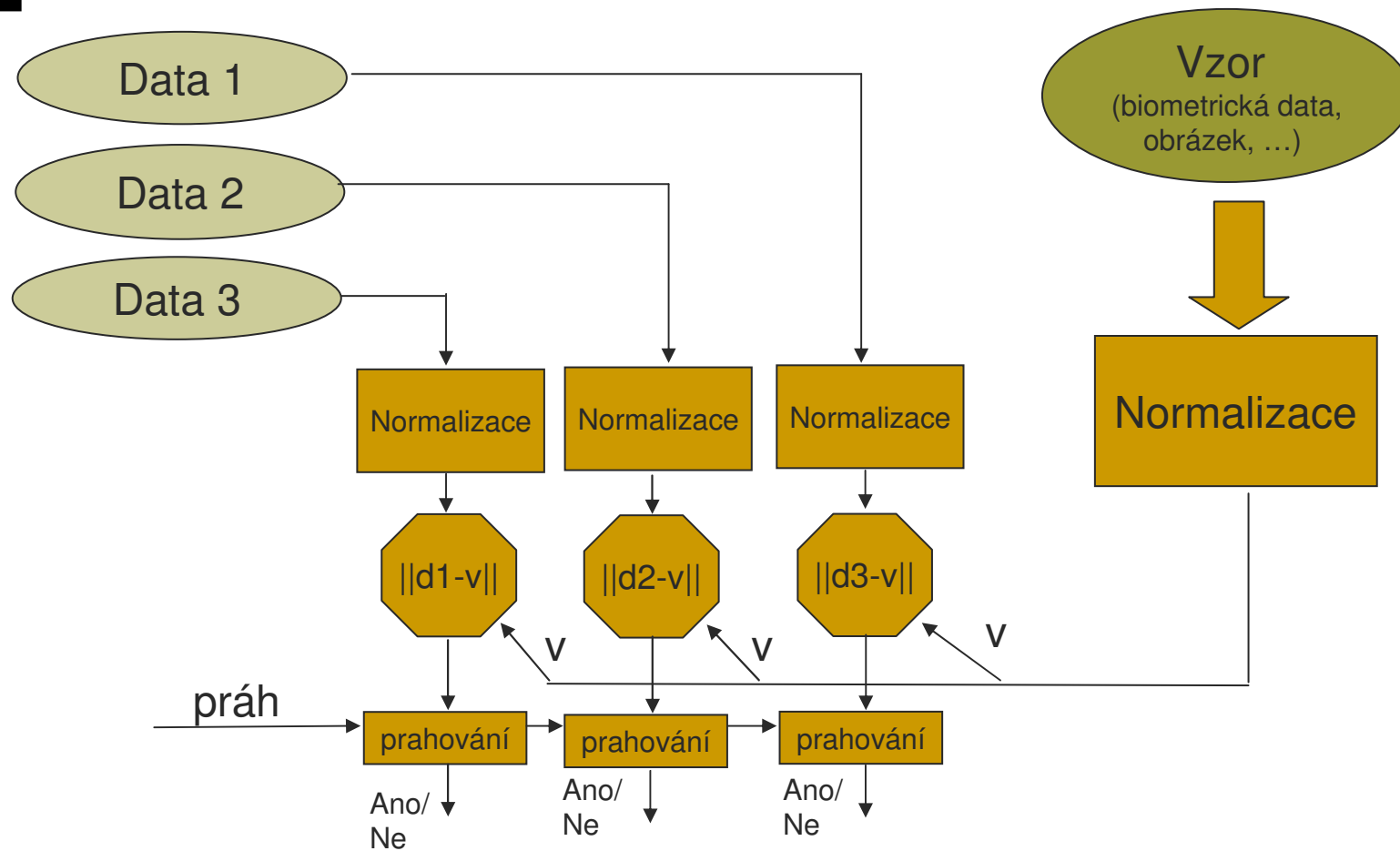


Range transformation ... Norma Min-Max
Z-transformation ... Z-scores

Určení rozsahu
(Pouze pro Min-Max normu)

Výběr metody

Použití normalizace při klasifikaci dat



[Míry podobnosti]

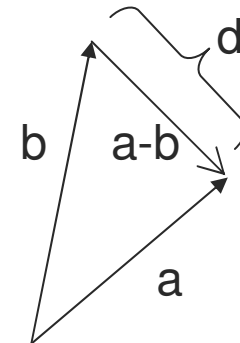
- V data miningu je často třeba data porovnávat (např. při shlukové analýze)
- Porovnávají se
 - Vektory
 - Funkční závislosti (typicky na čase)
 - Řetězce
 - Soubory dat (množiny)
- Skutečná podobnost (Karel je na fotografii podobný Pavlovi, osoba na videu se pohybuje podobně jako náš zločinec, ...) je těžko matematicky popsatelná, řeší se kombinací jednoduchých matematických metod (více či méně úspěšně)

[Euclidean distance]

Vzdálenost dvou vektorů

$$d = \|\vec{a} - \vec{b}\| = \sqrt{\sum_{i=1}^N (a_i - b_i)^2}$$

N je počet složek vektoru (dimenze)



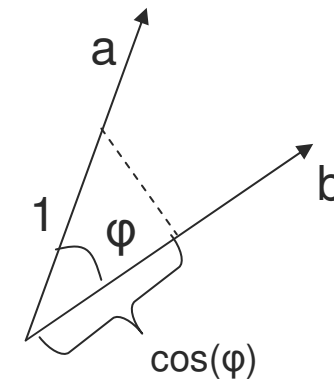
Jedna z nejčastěji užívaných metrik. Shodné vektory mají vzdálenost nula. Více odlišné vektory vykazují větší vzdálenost.

V porovnání obou vektorů hraje významnou roli velikost rozdílů složek těchto vektorů.

[Skalární součin]

Není metrikou v matematickém slova smyslu, lze však použít jako míru podobnosti.

$$\cos(\varphi) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|} = \frac{\sum_{i=1}^N a_i b_i}{\sqrt{\sum_{i=1}^N a_i^2} \sqrt{\sum_{i=1}^N b_i^2}}$$



Skalární součin vektorů $\vec{a} \cdot \vec{b}$ dělený velikostmi vektorů udává cosinus úhlu mezi těmito vektory. Shodné vektory vykazují hodnotu jedna a vektory kolmé (nejodlišnější) vykazují hodnotu 0.

V porovnání se nebere zřetel na velikost vektorů, ale jejich vzájemný úhel.

[Hammingova vzdálenost]

Uvažujeme binární vektory.

$$d = \sum_{i=1}^N a_i \oplus b_i$$

Hammingova vzdálenost je počet složek, ve kterých se dva vektory liší. Plus v kroužku označuje operaci XOR.

1**0**1110**1**0**1**0**1**00**1**

111110000000000

vzdálenost: 5

Levensteinova vzdálenost - výpočet

S1="aoj", S2="ahoj"

j	3	!=	!=	!=	==
o	2	!=	!=	==	!=
a	1	==	!=	!=	!=
	0	1	2	3	4
		a	h	o	j

		d(i,j)				
i	j	3	2	2	2	1
i=3	o	2	1	1	1	2
	a	1	0	1	2	3
i=0		0	1	2	3	4
		a	h	o	j	j
		j=0	j=4			

Výsledná vzdálenost t

Tento výpočet provádíme

pro $i = 1, i \leq \text{delka}(s1); i++$

pro $j = 1, j \leq \text{delka}(s2); j++$

Řetězy jsou indexovány od jedničky !

Předpokládáme, že je předem do tabulky vyplněno $d(0,j)=j$, $d(i,0)=i$ a $d(0,0)=0$.

$$d(i, j) = \min \begin{cases} d(i, j-1) + 1 \\ d(i-1, j) + 1 \\ d(i-1, j-1) + (s1(i) \neq s2(j)) \end{cases}$$


[Levensteinova vzdálenost]

Slouží k porovnání dvou řetězců. Jedná o tzv. editační vzdálenost.

Řetězci se postupuje zleva doprava. Pokud je na daném místě shoda ve znacích, vzdálenost se nemění. Je-li nutné pro dosažení shody přidat (nebo odebrat znak), zvětšuje se vzdálenost o jedničku.

Vyhodnocení se provádí metodou dynamického programování.

D=0	D=1	D=1
Ahoj	Aoj	Ahooj
Ahoj	Ahoj	Ahoj



Data mining

UAI/691 Přednáška 5-6

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

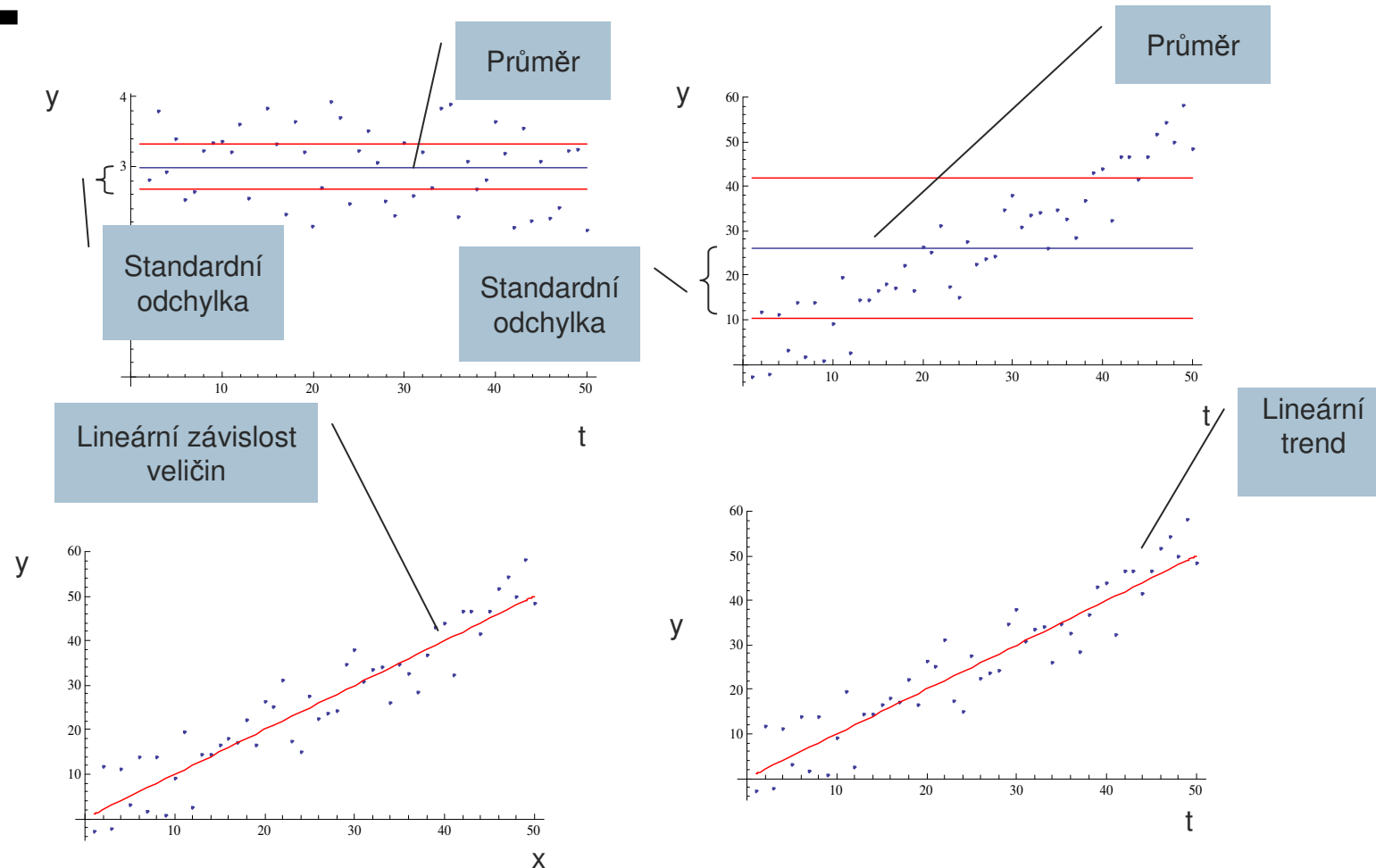
[Agenda

- Lineární regrese

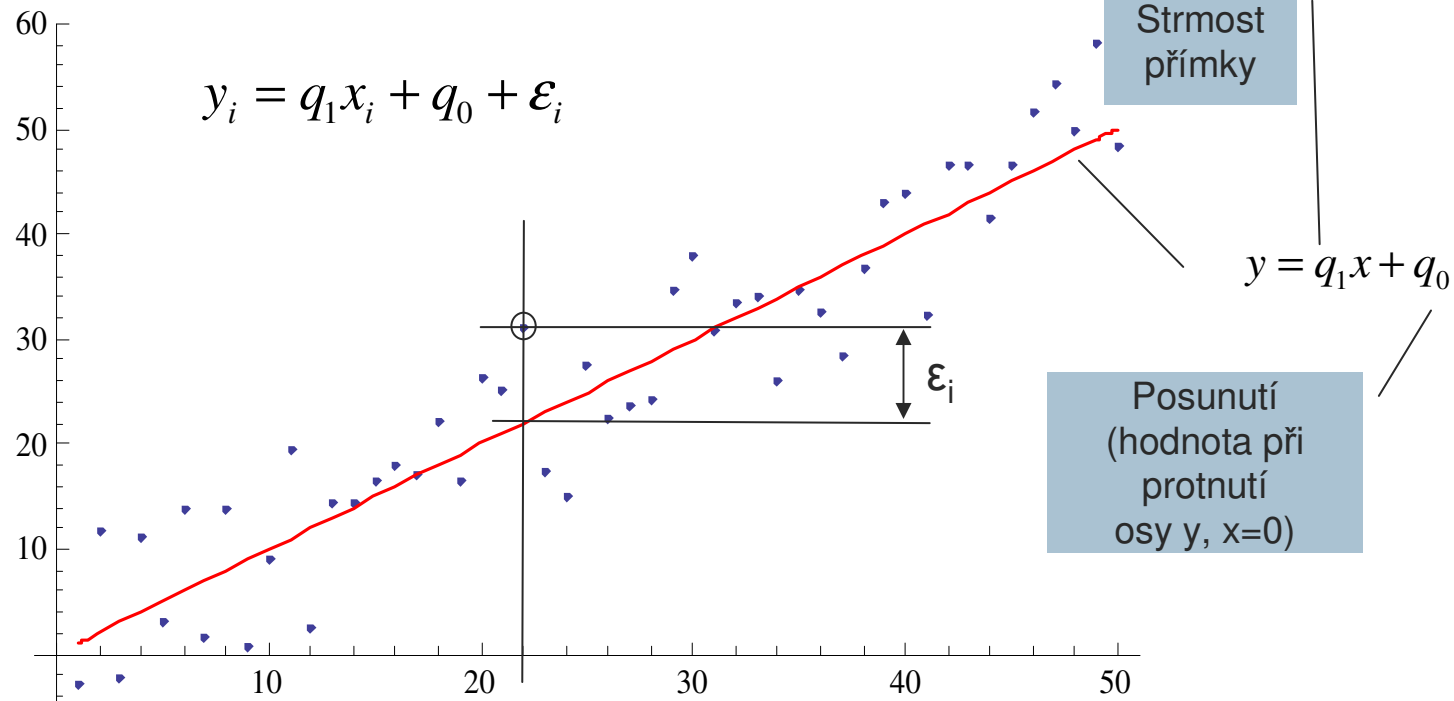
[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

Lineárně závislé veličiny a veličiny s lineárním trendem



y [Lineární regrese]



Cílem lineární regrese je nalézt takové parametry q_0 a q_1 , aby součet všech odchylek ε_i přes všechna data byl minimální.

Jedná se o jednoduchou optimalizační úlohu, kterou lze řešit metodou nejmenších čtverců

Matematika: parciální derivace

Mějme funkci $f(x_1, x_2, x_3, \dots, x_n)$ více proměnných. Tato funkce má v bodě $B=(b_1, b_2, b_3, \dots, b_n)$ parciální derivaci podle x_i , pokud existuje limita

$$\frac{\partial f(x_1, x_2, \dots, x_n)}{\partial x_i} = \lim_{h \rightarrow 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}$$

Pro výpočet parciální derivace využíváme stejná pravidla, jako v případě funkce jedné proměnné, přičemž všechny proměnné vyjma té, podle které derivujeme, považujeme za konstanty (tj. jejich derivace jsou rovny nule).

Příklad:

$$f(x, y, z) = 5x^2 - 8\log(y) + \sin(\omega z)$$

$$\frac{\partial f(x, y, z)}{\partial x} = 10x, \quad \frac{\partial f(x, y, z)}{\partial y} = -8\frac{1}{y}, \quad \frac{\partial f(x, y, z)}{\partial z} = \omega \cos(\omega z)$$

Výpočet q_1 a q_0 metodou nejmenších čtverců

$$D = \{[x_i, y_i], i \in \langle 1, N \rangle\} \quad E = \sum_{i=1}^N (y_i - q_1 x_i - q_0)^2$$

$$\min(E) \approx \frac{\partial E}{\partial q_0} = 0, \frac{\partial E}{\partial q_1} = 0 \Rightarrow q_0, q_1$$

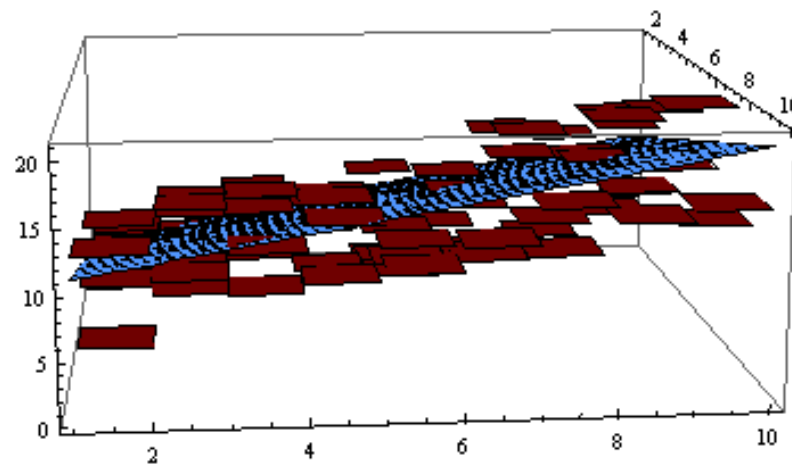
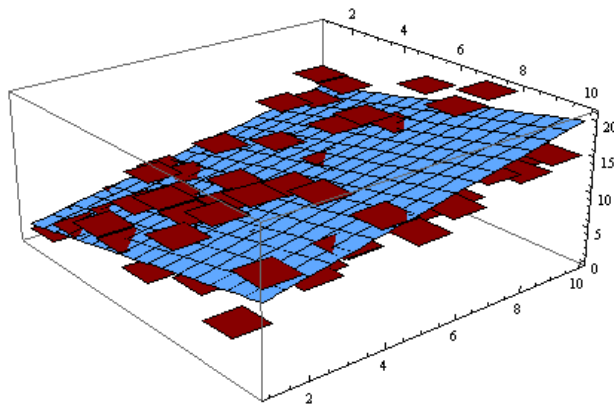
$$\begin{aligned} -2 \sum_{i=1}^N y_i + 2q_1 \sum_{i=1}^N x_i + 2q_0 n &= 0 \\ -2 \sum_{i=1}^N y_i x_i + 2q_1 \sum_{i=1}^N x_i^2 + 2q_0 \sum_{i=1}^N x_i &= 0 \end{aligned}$$

$$q_0 = \frac{\sum_{i=1}^N y_i \sum_{i=1}^N x_i^2 - \sum_{i=1}^N x_i y_i \sum_{i=1}^N x_i}{n \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2}$$

$$q_1 = \frac{n \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{n \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2}$$

Vícerozměrná lineární regrese

Data prokládáme rovinou a hledáme takovou orientaci roviny v prostoru, aby se minimalizoval součet chyb pro všechny řádky datové matice.



Vícerozměrná lineární regrese

$$y_i = q_0 + q_1 x_{i1} + q_2 x_{i2} + q_3 x_{i3} + q_4 x_{i4} + \dots + q_m x_{im} + \varepsilon_i$$

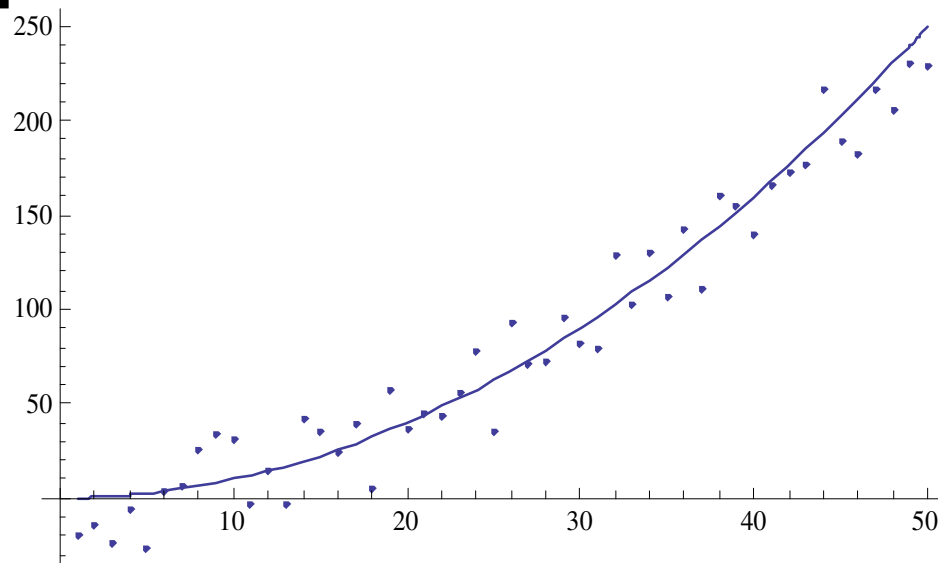
$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} q_1 \\ \dots \\ q_m \end{bmatrix}$$

$$\mathbf{q} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

x_{i1}	x_{i2}	x_{i3}	y
1	0.3	-1.2	5
0.2	-5.1	0.2	2
-2.3	1.4	2.2	-3
1.1	0	-0.8	3

Převzato z: Petr Berka: Dobývání znalostí z databází. Nakladatelství
ACADEMIA, 2003. ISBN 80-200-1062-9

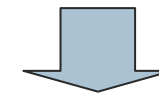
Linearizace nelineárních závislostí



$$y = q_0 + q_1x + q_2x^2$$

Podobně můžeme přidávat další nelineární členy x^3, x^4

x	y
1	-19.7
2	-13.9
3	-23.4
...	



Přidáme x^2

x	x^2	y
1	1	-19.7
2	4	-13.9
3	9	-23.4
...		

Pak již řešíme lineární úlohu

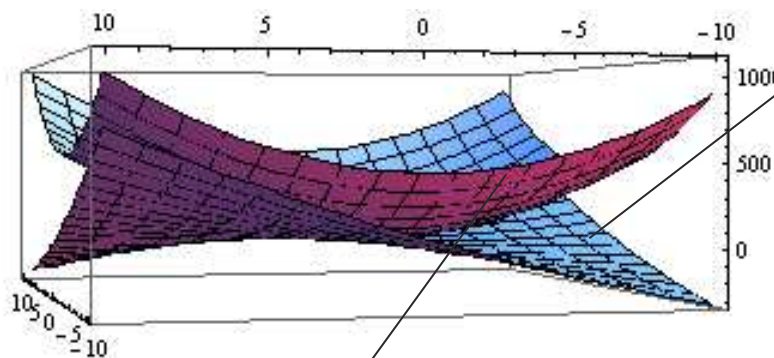
[Interakce]

Příklad pro dvě nezávislé proměnné

interakce

$$y = q_0 + q_1x_1 + q_2x_2 + q_3x_1x_2 + q_4x_1^2 + q_5x_2^2$$

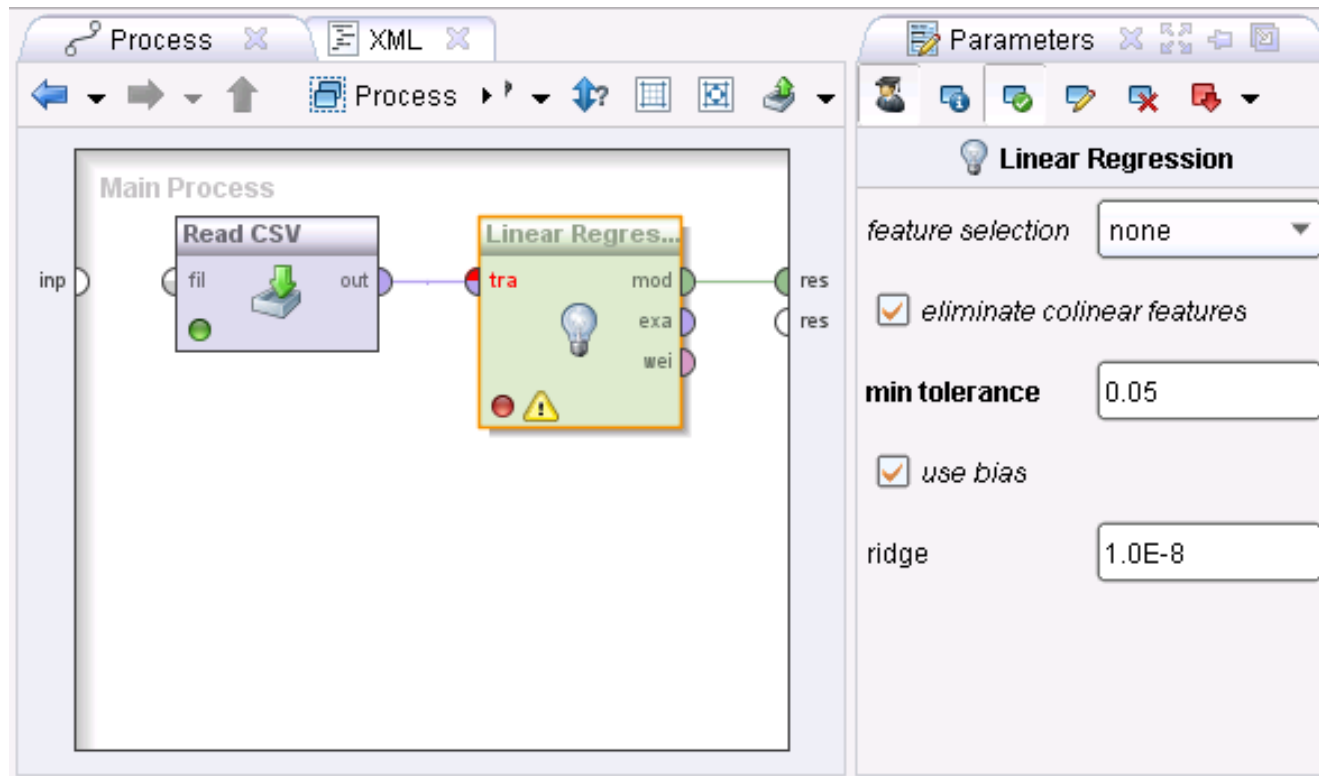
Přílišný počet interakcí může vést k špatnému modelu



Bez interakce

S interakcí

Lineární regrese v Rapid Mineru

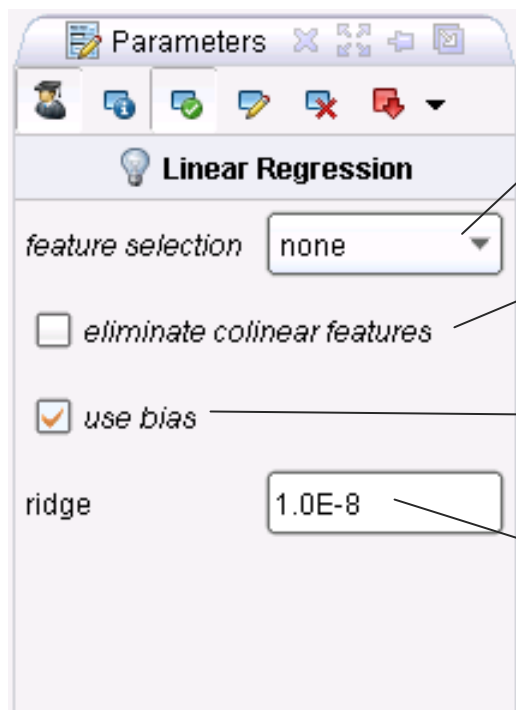


Modeling → Function Fitting → Linear Regression

[Datová matice pro lineární regresi]

- Numerické atributy (1 a více)
- Jeden numerický atribut v roli label
- Nominální atributy je třeba konvertovat na numerické

[Základní funkcionality]



Automatický výběr atributů je vypnut – všechny atributy vstupují do regrese


Vyřazení lineárně závislých atributů

Zařazení členu q_0 (posunutí)

Pokud je $\det(X^T X)$ je blízký nule, pak se použije modifikovaný algoritmus, který využívá tento parametr.

[Rozšířená funkcionality]

- Automatický výběr atributů
 - Speciálním algoritmem (viz. parametr: feature selection) se vybere taková podmnožina atributů, pro které je lineární regrese nejpřesnější.
- Vyřazení závislých atributů (Eliminate colinear features). Závislé atributy nenesou žádnou novou informaci a způsobují problémy při výpočtu vektoru q . Parametr minimum tolerance určuje míru závislosti pro vyřazení
- Vynechání posunutí, tj. koeficientu q_0 , z modelu (use bias nezaškrtnuto)



Data mining

UAI/691 Přednáška 5-6

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda

- Shluková analýza

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

[Shluková analýza]

- Důležitá metoda analýzy dat
- Řeší problém nalezení podobností v neznámých datech
- Podobná data typicky leží v prostoru blízko sebe (tvoří shluky)
- Zajímají nás parametry shluku jako je střed (těžiště) a velikost (rozptyl). Vzory v okolí středu můžeme považovat za reprezentanty shluku (typické hodnoty)

Příklad

Máme data z vyšetření od skupiny pacientů a máme zjistit, zda existují takové podskupiny skupiny pacientů, kteří mají podobné výsledky vyšetření. V dalším kroku pak zkoumáme, zda tyto skupiny náležejí k zdravým, či nemocným, případně s jakou závažností nemoci, případně které nemoci.

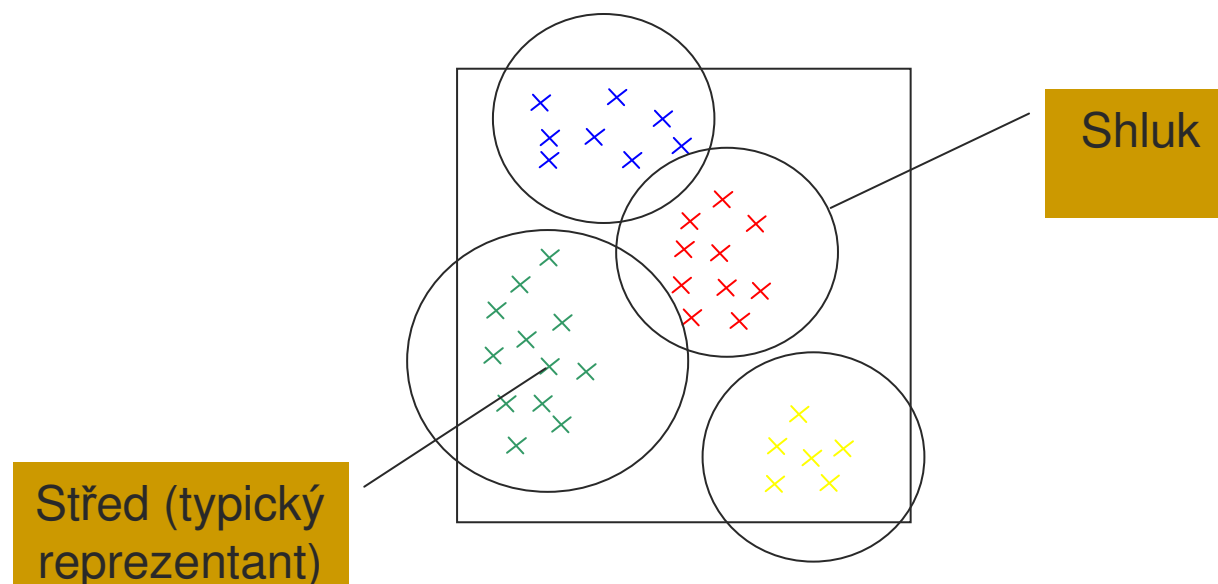
Data pacientů s obdobnými příznaky nemoci budou podobná a budou tvořit shluky. Pacienti, ležící ve středu shluku reprezentují typického pacienta s danou formou nebo závažností nemoci.

Očekáváme, že najdeme vztah mezi výsledky vyšetření a úsudkem lékaře (když to lékař pozná musí takový vztah existovat).

Pokud nenajdeme vztah shluků ke zdraví pacientů, pak patrně bude chybná diagnostická metoda, která dostatečně nevypovídá o diagnostikované nemoci. Předpokládáme, že známe spolehlivě stav pacientů z úsudku lékaře, který používá jinou diagnostickou metodu.

Jiným příkladem pak bude zkoumání dat o zákaznících mobilních operátorů, kde nám může např. tato metoda pomoci nalézt novou cílovou skupinu zákazníků. Podobně se můžeme zaměřit na uživatele Internetu, e-maily (spam), hledání typického chování hackerů, apod.

Příklad shluků v datech



[Shlukování dat]

- Hledání podmnožin podobných vzorů
- Definice podobnosti na rozdíl od shodnosti je nejasná, používají se různé metriky
 - Hammingova
 - Euklideovská
 - Čebyševova
- Výsledkem jsou sobě podobné množiny vzorů, interpretace (co znamenají, kdo jsou ti v množině) je na expertovi a výsledcích dalších analýz

[Hierarchické shlukování]

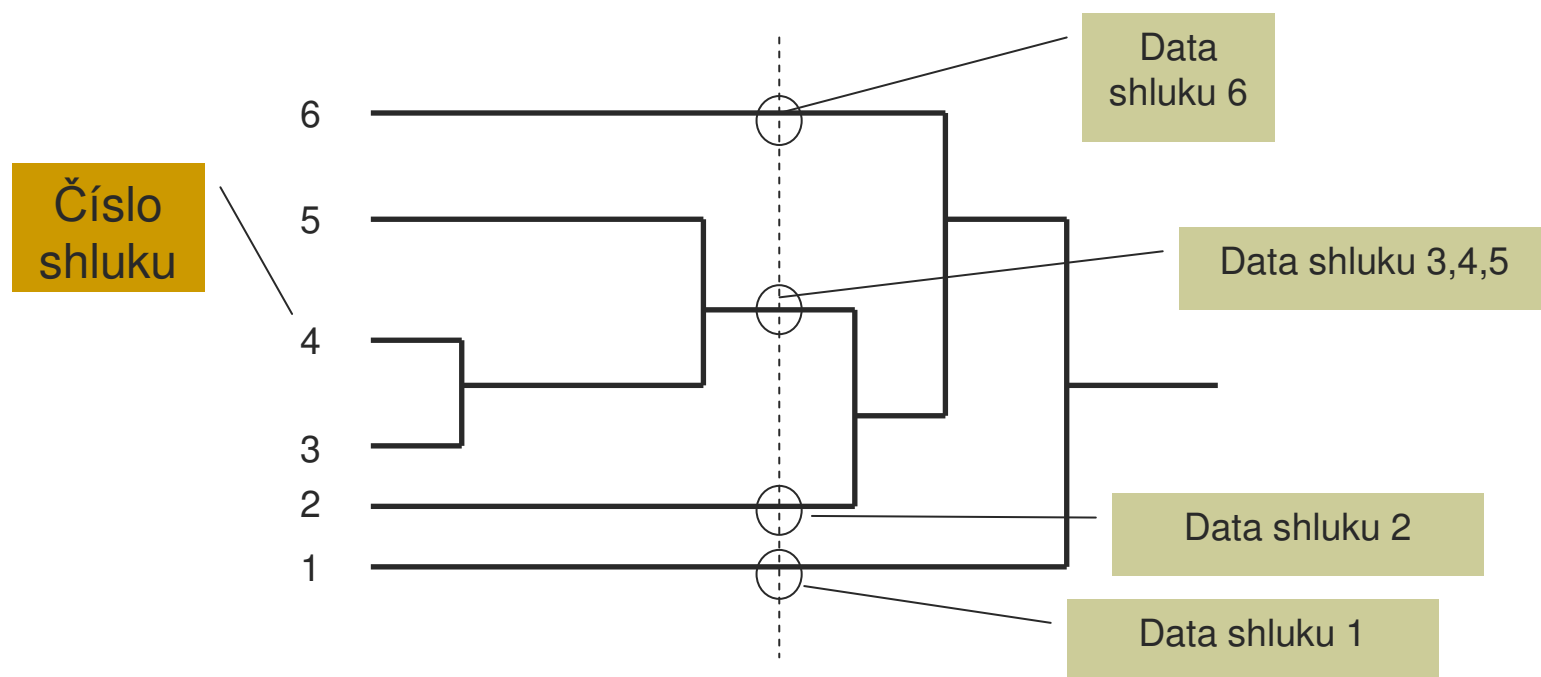
- Metoda shlukování metodou zdola nahoru
- Na začátku každý vzor (případ, example, řádek datové matice) je umístěn do samostatného shluku (co vzor, to shluk)
- Pak opakovaně shluky spojujeme, až získáme jeden shluk, obsahující všechny vzory.
- Spojujeme blízké shluky, spojování si poznamenáváme ve formě stromu.

Určení vzdálenosti mezi shluky (pro předem stanovenou metriku)

- Metoda nejbližšího souseda
 - Vzdálenost mezi shluky A a B je dána minimem vzdálenosti mezi vzory shluků A a B
- Metoda nejvzdálenějšího souseda
 - Vzdálenost mezi shluky A a B je dána maximem vzdálenosti mezi vzory shluků A a B
- Metoda průměrné vzdálenosti
 - Vzdálenost mezi shluky A a B je dána průměrnou vzdáleností mezi vzory shluků A a B
- Metoda centroidní
 - Vzdálenost mezi shluky je dána vzdáleností středů shluku

[Dendrogram]

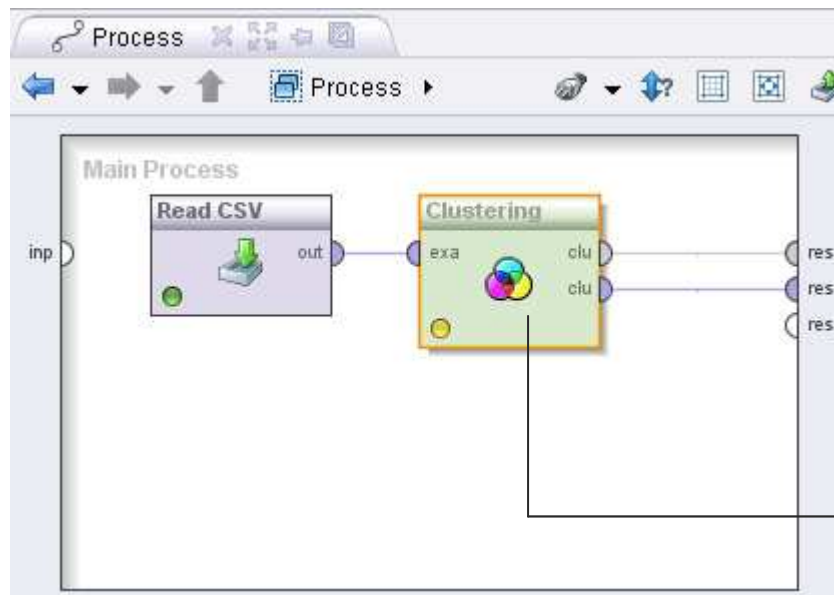
Zachycuje hierarchii shluků.



[K-means algoritmus]

1. Odhadneme počet shluků.
2. Náhodně vygenerujeme středy shluků. Počet středů je roven počtu shluků, střed je reprezentován vektorem s stejnou dimenzí, jako je dimenze vektorů, které tvoří shlukovaná data (tj. počet atributů ve vzoru).
3. Určíme, k jakému shluku patří jaký vzor. Daný vzor patří do toho shluku, k jehož středu je nejbližší (minimální Euclidean vzdálenost (nebo jiná vzdálenost) od daného středu)
4. Z dat, která náleží shluku vypočteme nový střed (postupně po attributech počítáme průměr přes všechny vzory ve shluku) a nahradíme jím původní střed shluku. Toto provedeme pro všechny shluky.
5. Body 3 a 4 opakujeme do ustálení (tj. do doby, kdy se středy posunou o méně než je zadaná hodnota)
6. Ukončíme algoritmus


Shlukování algoritmem K-means



Parameters

Clustering (k-Means)

- ☒ add cluster attribute
- ☐ add as label
- ☐ remove unlabeled
- k: 3
- max runs: 10
- max optimization steps: 100
- ☐ use local random seed



Data mining

UAI/691 Přednáška 10-11

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda]

- Modelování
- Klasifikace
- Rozhodovací stromy
- Vyhodnocení modelu

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

[Modelování]

- Významná součást data miningu, zejména procesu dobývání znalostí
- Vstupem modelování jsou předzpracovaná data
- Výstupem modelování jsou znalosti ve formě
 - Reprezentantů (etalonů ve formě hodnot nebo vektorů)
 - Funkcí (jednotlivých nebo směsí funkcí)
 - Pravidel
- Modely mohou být
 - Jednoduché
 - Komplexní
- Současný trend v modelování zahrnuje
 - Kombinování modelů
 - Volbu modelu na základě meta informací o datech a modelovacích metodách

[Rozhodovací stromy]

- Znalosti jsou reprezentovány v podobě stromů
 - Uzly reprezentují určitou třídu dat
 - Větvení reprezentuje strukturu dané třídy dat
- Postup metodou rozděl a panuj
 - Trénovací množina se postupně dělí tak, aby v každé množině převládala data jedné třídy.
- Algoritmus je vhodný pro kategorická data, po úpravách i pro numerická data
- Rozhodovací stromy se používají pro
 - Regresi (regresní stromy)
 - Klasifikaci (klasifikační stromy)

[Základní algoritmus]

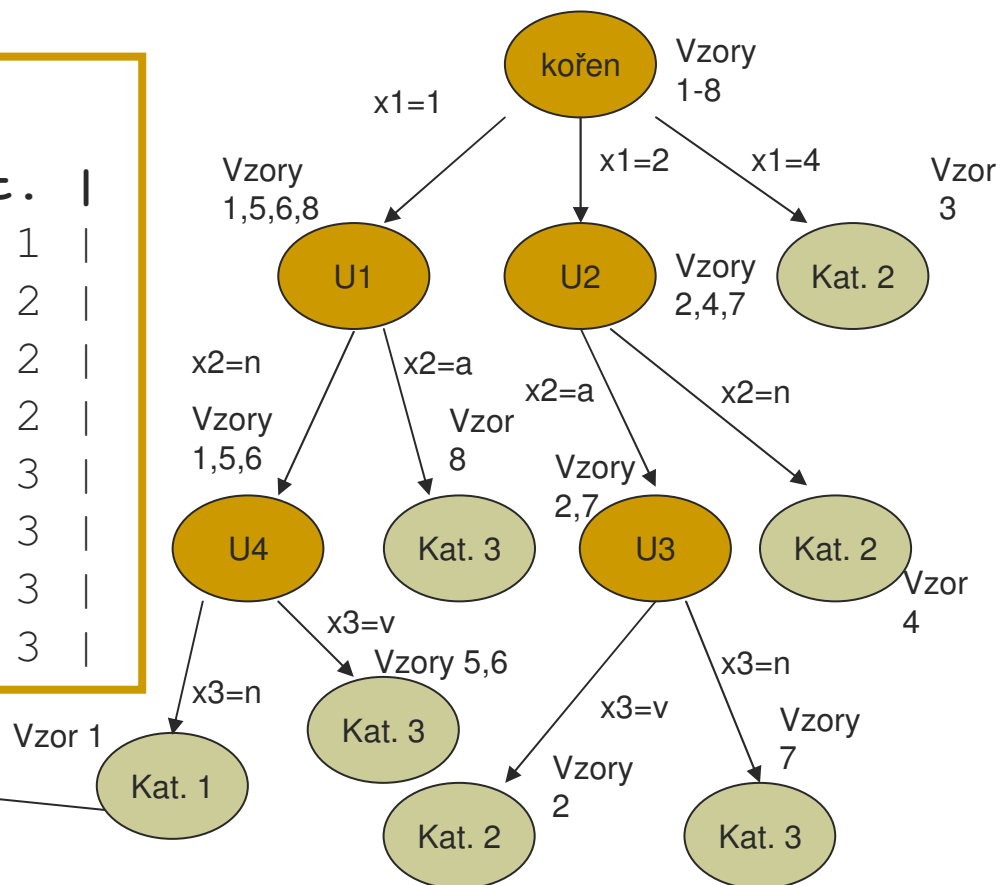
1. Zvolit jeden atribut (kořen podstromu)
2. Data rozdělit podle hodnot zvoleného atributu na podmnožiny a každé podmnožině přiřadit nový uzel stromu
3. Pokud existuje uzel, kde všechna data nepatří do téže třídy, pak zpět na bod 1
4. Ukonči algoritmus

Zdroj: Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9

Příklad

DATA				
č.	x1	x2	x3	kat.
01	1	n	n	1
02	2	a	v	2
03	4	a	v	2
04	2	n	n	2
05	1	n	v	3
06	1	n	v	3
07	2	a	n	3
08	1	a	n	3

Všechny vzory spadají do kategorie 3, není nutno dále dělit



[Volba atributu pro větvení]

- Cílem je vybrat atribut, který nelépe odliší příklady různých tříd, tj.
 - Maximalizovat počet vzorů téže třídy v podmnožinách vzniklých rozdělením množiny daným atributem.
- Používaná kritéria
 - Entropie (minimalizace entropie – neurčitosti - náhodnosti v podmnožinách)
 - Informační zisk (maximalizace redukce entropie při použití zvoleného atributu. Vztahuje se na entropii počítanou pro daný atribut pro celá data)
 - Poměrný informační zisk (informační zisk dělený počtem větvení, zohledňuje počet hodnot atributu)
 - Gini index (vychází z počtu příkladů dané třídy zjišťované na nějaké množině nebo podmnožině)
 - Chí kvadrát

[Výběr atributu dle entropie]

Výběr atributu dle entropie:

1. Pro všechny atributy A a hodnoty v , které atribut nabývá, spočti entropie takto:
 1. Pro všechny třídy $t = \{1, 2, \dots, N_t\}$ spočti pravděpodobnost, že je třída t pokryta atributem hodnoty v , tj. spočítej četnost příkladů, které spadají do třídy t a mají vybraný atribut A roven hodnotě v a poděl počtem příkladů s hodnotu atributu A rovnou v .
 2. Na základě pravděpodobností spočti entropii pro daný atribut a jeho konkrétní hodnotu
2. Pro každý atribut A spočti střední entropii $H(A)$ přes všechny možné hodnoty v atributu. Entropii násobíme poměrem počtu příkladů atributu A s hodnotou v k celkovému počtu příkladů a sečteme.
3. Vybereme atribut s nejmenší střední entropií.

[Informační zisk]

$$Z(A)=H(C)-H(A)$$

Kde $H(A)$ je vypočtená hodnota z minulého slidu a $H(C)$ je entropie tříd atributu reprezentujícím třídy.

$$H(C) = -\sum_{t=1}^T p_t \log p_t \qquad p_t = \frac{n_t}{n}$$

[Testování klasifikačních modelů]

- Testování na trénovacích datech
- Křížová validace
- Leave-one-out
- Bootstrap
- Testování na testovacích datech

Testování na trénovacích datech

- Má omezenou vypovídací schopnost
 - Říká nám, jak přesně se model přiblížil trénovacím datům
 - Nepostihuje schopnost modelu zevšeobeňovat (tj. reagovat na neučená data)
- Neodhalí přeučení modelu
- Nedostatečná metoda pro vytvoření kvalitního modelu

[Křížová validace]

- Dostupná data se rozdělí na n části (např. $n=10$)
- 9/10 dat se použije pro učení
- 1/10 dat se použije pro testování
- Proveďte se celkem n testů a výsledky se zprůměrují

[Leave-one-out]

- Obdoba křížové validace
- Máme-li n vzorů, $n-1$ učíme a jeden použijeme na testování
- Provedeme tedy n testů a výsledky zprůměrujeme
- Metoda může být časově náročná

[Bootstrap]

- Vzory do trénovací množiny z dostupných dat vybíráme tak, že se některé vzory mohou opakovat
- Zbylé vzory použijeme pro testování
- Počty vzorů jsou
 - Přibližně 63% trénovacích
 - Přibližně 37% testovacích
 - Tyto hodnoty platí za předpokladu, že trénovací množina má stejný počet vzorů jako má datová množina, ze které trénovací množinu vytváříme.

[Náhodný výběr]

- Z dostupných dat vybereme
 - 75% pro trénování
 - 25% pro testování
- Vzory se neopakují
- Testování se provede jen jednou

[Matice záměn (confusion matrix)]

- Matice, kde řádky odpovídají odpovědím modelu a sloupce správným odpovědím
- Pro bezchybný klasifikátor jsou nenulové hodnoty pouze na hlavní diagonále
- Součet všech hodnot v matici je roven počtu vzorů

[Příklad pro 2 hodnoty]

	Model Ano	Model Ne
Správně Ano	True Positive (TP)	False Negative (FN)
Správně NE	False Positive (FP)	True Negative (TN)

[Vyhodnocení]

$\text{Acc} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$ Celková správnost (accuracy)


$\text{Err} = (\text{FP} + \text{FN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$ Celková chyba

$\text{Přesnost} = \text{TP} / (\text{TP} + \text{FP})$

$\text{Úplnost} = \text{TP} / (\text{TP} + \text{FN})$

F - míra

$F = 2 \cdot \text{Přesnost} \cdot \text{Úplnost} / (\text{Přesnost} + \text{Úplnost})$



Data mining

UAI/691 Přednáška 10-11

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda

- Asociační pravidla

[Literatura]

- Petr Berka: Dobývání znalostí z databází. Nakladatelství ACADEMIA, 2003. ISBN 80-200-1062-9
- RapidMiner 5.0 - User Manual, Technická dokumentace k programu RapidMiner. Rapid-I GmbH, 2010. http://garr.dl.sourceforge.net/project/rapidminer/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf

[Asociační pravidla]

Pravidlo typu

IF ... THEN ...

IF x1==ano && x2==stredne THEN x3==ne

atribut

hodnota

Předpokládají se kategorické hodnoty, numerické hodnoty se musí diskretizovat.

Z hodnot atributů se vytvářejí konjunkce

např: x1(ano), x2(stredne), x3(ne), x1(ano) & x2(stredne), x1(ne) & x3(ne), x2(malo) & x3(ano), x1(ano) & x2(malo) & x3(ano)

a počítají se četnosti výskytu výše uvedených kombinací pro konkrétní hodnoty atributů.

[Hodnocení pravidel]

IF Ancestor **THEN** Successor

Ancestor – předpoklad

Successor - závěr

Čtyřpolní tabulka

	Successor (pravdivý)	Successor (nepravdivý)
Ancestor (pravdivý)	a	b
Ancestor (nepravdivý)	c	d

a,b,c,d jsou četnosti výskytu

a+b+c+d je počet příkladů
(počet řádků datové matice)

Podpora (support) =
 $a/(a+b+c+d)$

Spolehlivost(confidence) =
 $a/(a+b)$

Příklad

X1(A,N)	X2(1,2,3,4)	X3(S,M,L)
A	1	S
N	2	S
N	3	M
A	2	S
N	4	L
N	2	L
N	3	L

Počet příkladů $n = 7$

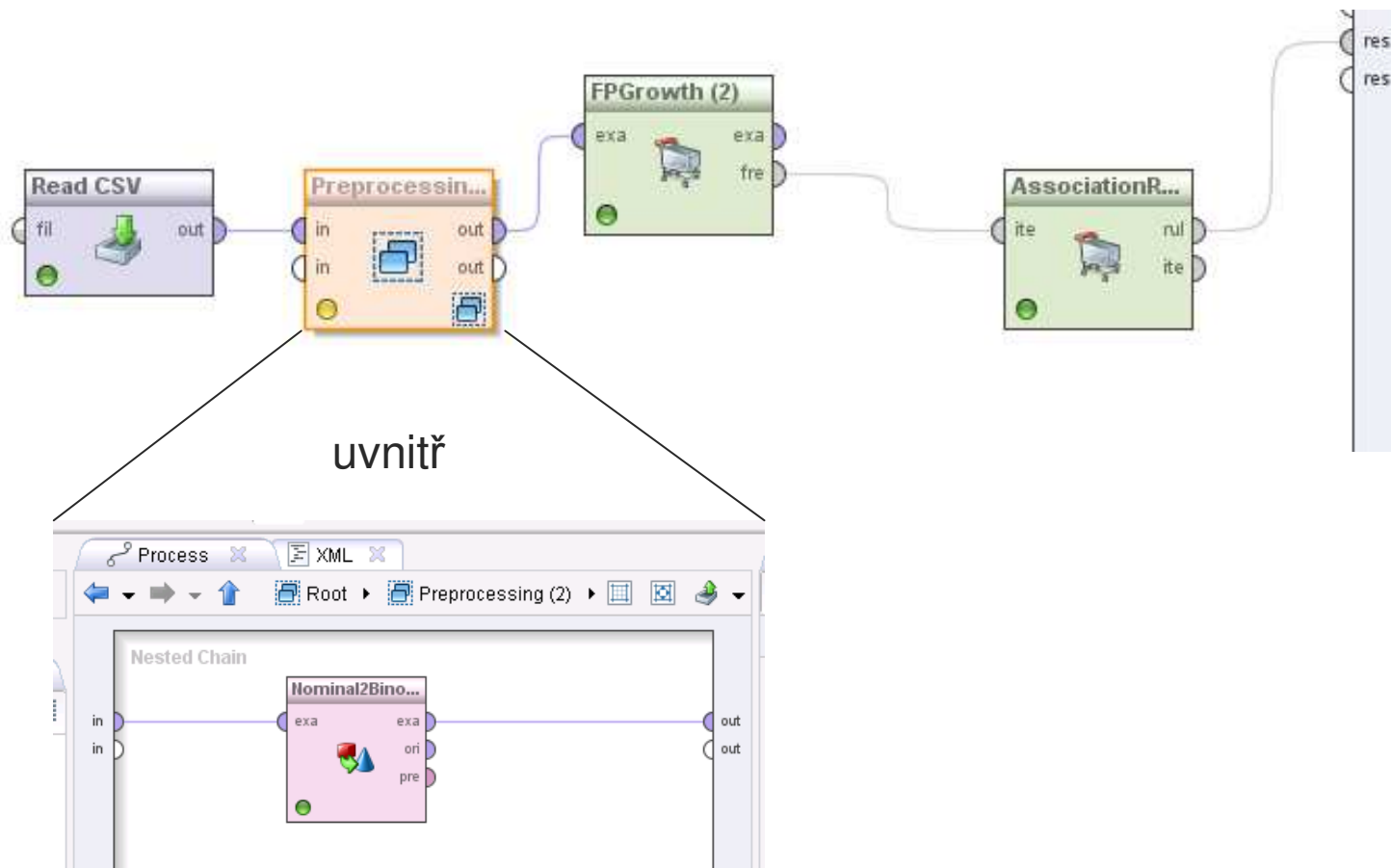
Kombinace	četnost	podp.	spol. X3==S


X1 (A)	2	2/7	1
X1 (N)	5	5/7	1/5
X2 (1)	1	1/7	1
X2 (2)	3	3/7	2/3
X2 (3)	2	2/7	0
X2 (4)	1	1/7	0
X3 (S)	3	3/7	1
X3 (M)	1	1/7	0
X3 (L)	3	3/7	0
X1 (A) X2 (1)	1	1/7	1
X1 (N) X2 (2)	2	2/7	1/2
X1 (N) X3 (L)	3	3/7	0
X1 (N) X2 (3) X3 (L) ...	1	1/7	0

[Algoritmus apriori]

- Nejznámější algoritmus pro hledání asociačních pravidel (Agrawal, 1996)
- Založen na hledání kombinací atributů s vysokou četností
- Postupuje se od kombinací délky $k=1$ výše
- Kombinace délky k vzniknou spojením dvojic kombinací délky $k-1$.
- Po nagenеровání kombinací délky k se provádí prořezání (prunning). Odstraňují se kombinace, které po spojení nemají $k-2$ shodných kategorií a ty, které nemají některou z podkombinací délky $k-1$ obsaženou v seznamu kombinací délky $k-1$

[V Rapid Mineru]





Data mining

UAI/691 Přednáška 14

Miroslav Skrbek
mskrbek@prf.jcu.cz

*Ústav aplikované informatiky
Přírodovědecká fakulta
Jihočeské univerzity v Českých Budějovicích*

[Agenda]

- Kombinování modelů
- Visualizace
- Analýza textu

[Kombinování modelů]

- Nejlepších výsledků nelze dosáhnout jedním modelem
- Lepších výsledků se dosahuje kombinováním modelů
- Metody
 - Bagging
 - Boosting

[Bagging]

- Všechny modely mají stejnou váhu
- Z trénovacích dat se vytvoří podmnožiny (náhodný výběr s opakováním) a na každou množinu se naučí jeden model
- Vytvořené modely hlasují o výsledku

[Boosting]

- Modely se vytvářejí postupně
- Novější modely mají větší váhu hlasu
- Při učení se nově vytvořený model zaměřuje na data, která byla špatně klasifikována

[Grafické znázornění dat]

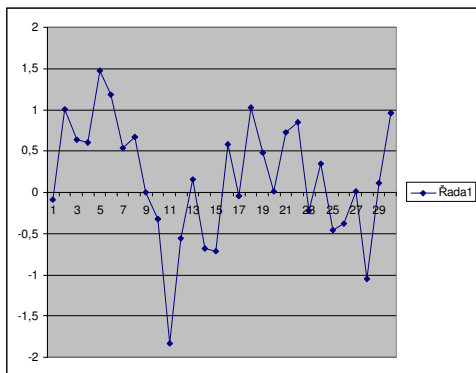
- Poskytuje velmi cenné informace, které je možné užít pro volbu předzpracování nebo modelu
- Problémem je omezení do dimenze max. 3
- Pro vyšší dimenze je nutno použít method pro redukci dimenze např. PCA (Principal Component Analysis)
- Existuje velké množství zobrazovacích metod (typů grafů)
- Důležitý pro prezentaci výsledků analýzy

[Základní grafy]

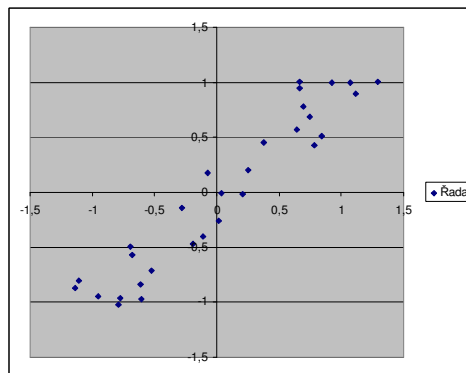
- Bodový, spojnicový
- Sloupcový (2D, 3D)
- Sloupcový kumulativní (2D, 3D)
- XY (2D), XYZ (3D)
- Koláčový
- Polární souřadnice

[Základní grafy]

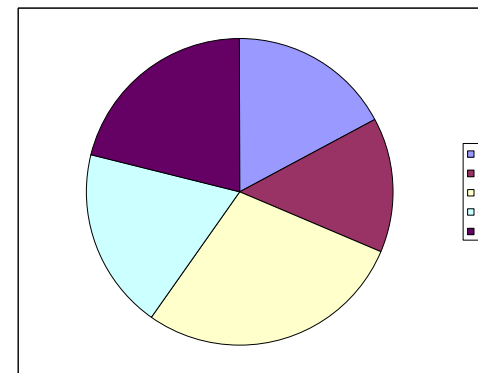
Spojnicový



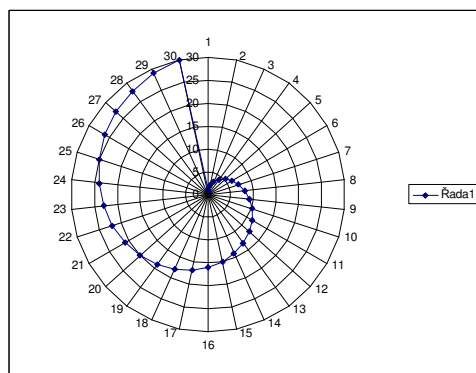
XY



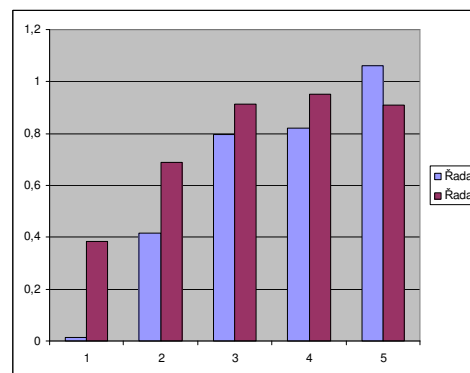
Koláčový



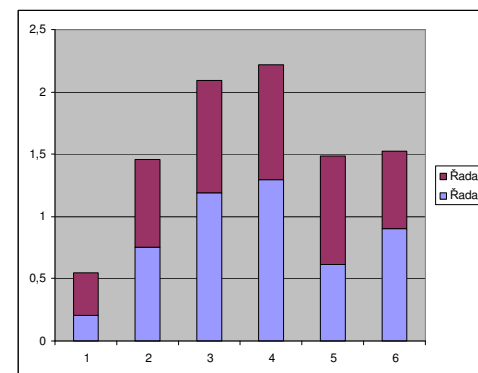
Polární



Sloupcový



Sloupcový kumulativní

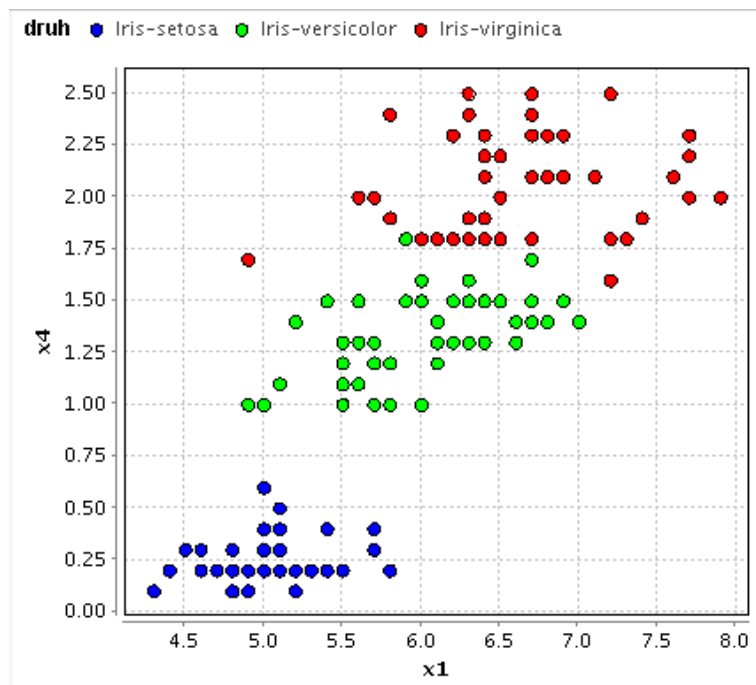


[Pokročilé grafy]

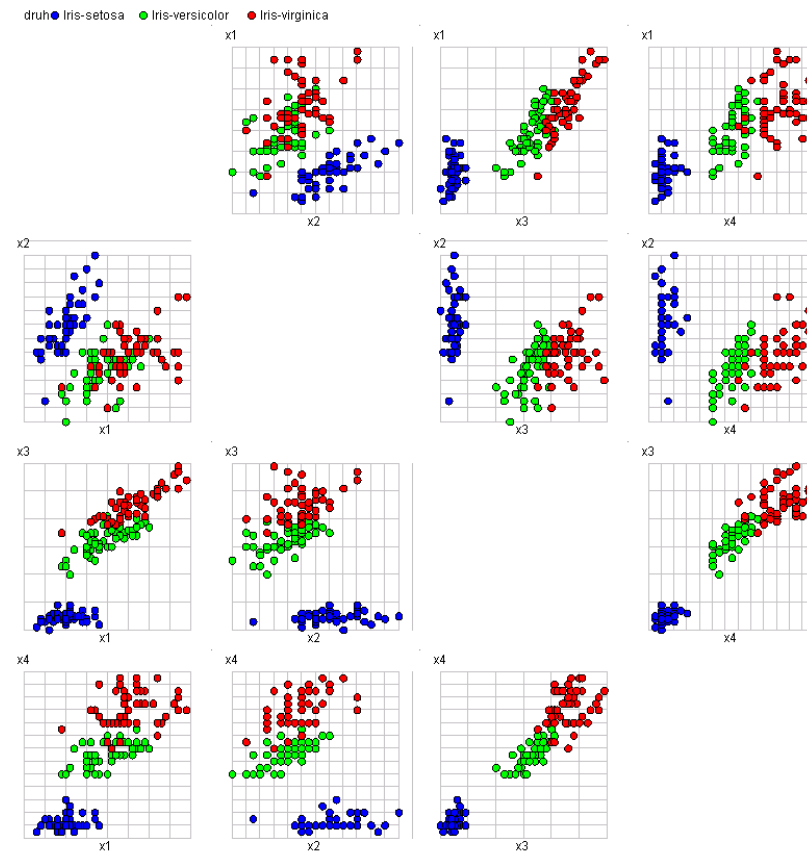
- Scatter plot
- Paralelní
- Graf (uzly, hrany, orientovaný či neorientovaný)
 - Zobrazení vztahu (silou čáry míru síly) mezi entitami
- Matice záměn (barvou chyby/správně, odstínem velikost)
- ROC křivka
 - Hodnocení klasifikátoru
- Spektrogram

[Scatter plot]

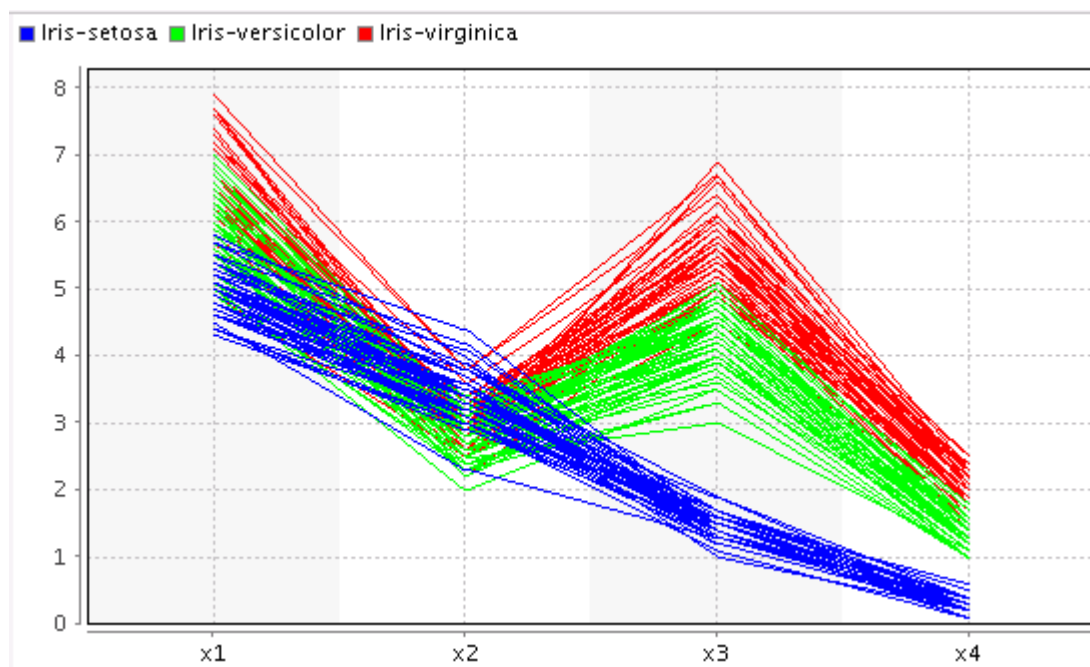
Scatter



Scatter Matrix



[Paralelní graf]

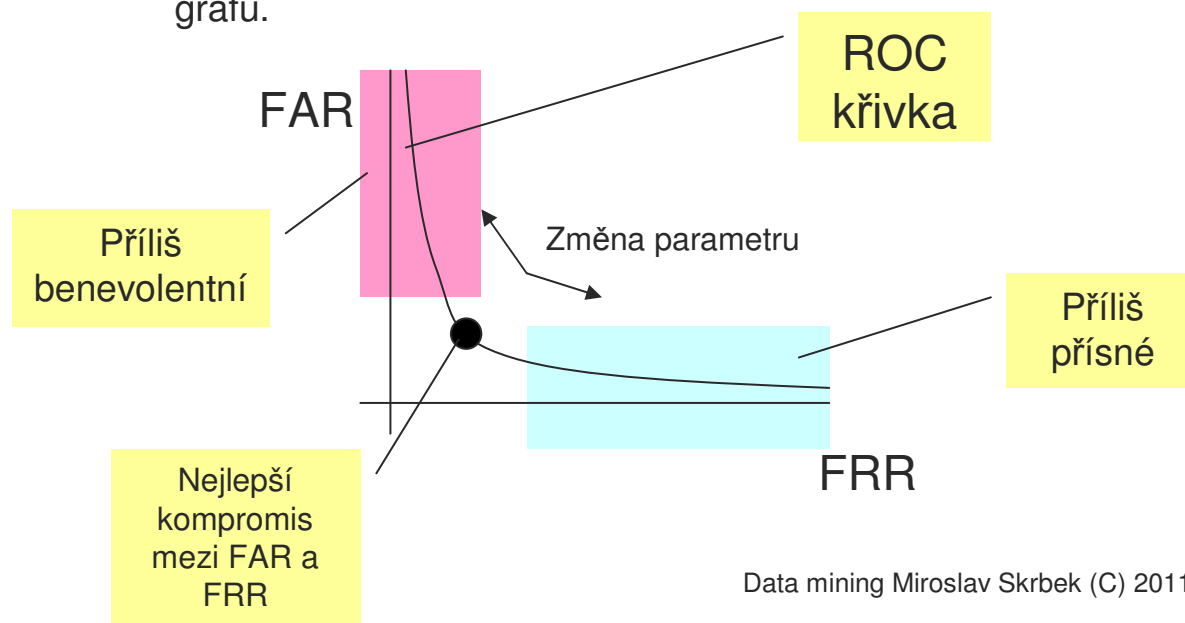


ROC křivka

ROC (Receiver Operating Curve) se využívá u hodnocení klasifikátorů

Konkrétní příklad: čtečka otisků prstů – výstup akceptovat/neakceptovat otisk

Pro testovací množinu sestojíme sadu čtyřpolních tabulek pro některý parametr rozpoznání (typicky prahovou hodnotu) a určíme False Acceptance Rate (FAR = chyba **false negative**) a False Rejection Rate (FRR = chyba **false positive**). Získané hodnoty FRR a FAR vyneseme do grafu.



[Zpracování textu]

- Slovníky
- N-gramy
 - Bi-gramy
 - Trigramy
- Stemming
 - Nahrazení slova jeho základem
- Lematization
 - Obdobné jako stemming, ale s ohledem na kontext