

Základní analytické metody

Měření a zpracování dat (MDS)

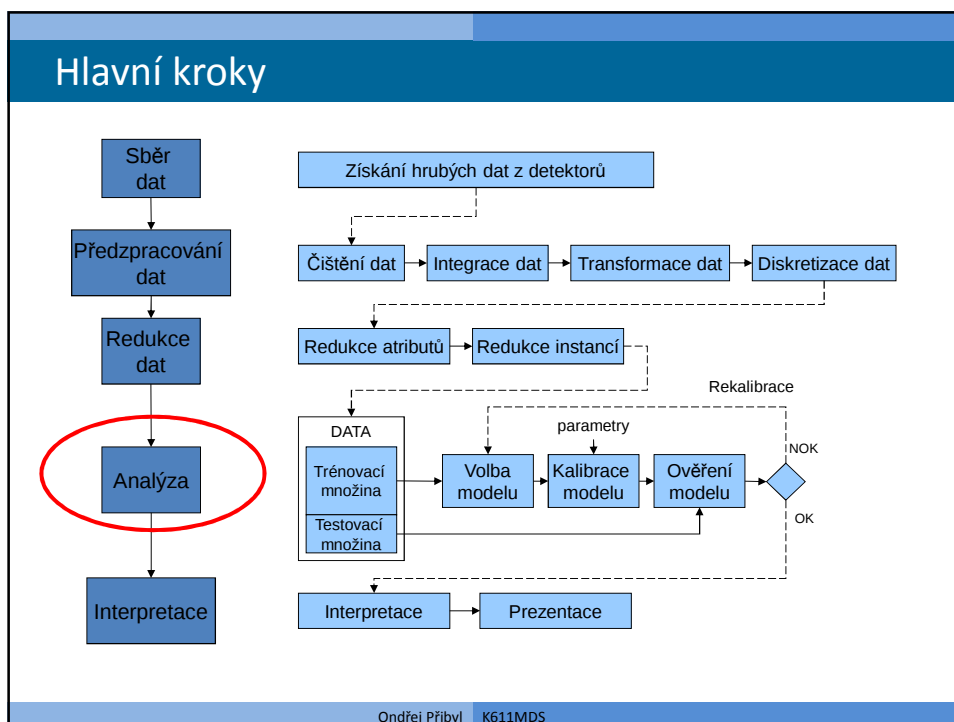
Ondřej Přibyl

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní



Obsah prezentace

- Úvod do analýzy dat
- Shlukování
- Rozhodovací stromy
- Lineární regrese



Dělení analytických nástrojů

dle toho jak nastavují či modifikují své parametry (fáze učení či kalibrace):

- **Učení s učitelem** (i v české literatuře se často používá anglický termín „supervised learning“) kdy **existuje informace** o správné kategorii či o chybě predikce pro každý vzor.
- **Učení bez učitele** (unsupervised learning) kdy **neznáme** správnou kategorii ani informaci o výsledné chybě predikce a systém sám vytváří přirozené shluky na základě podobnosti mezi objekty.

Jak složitý model použít?

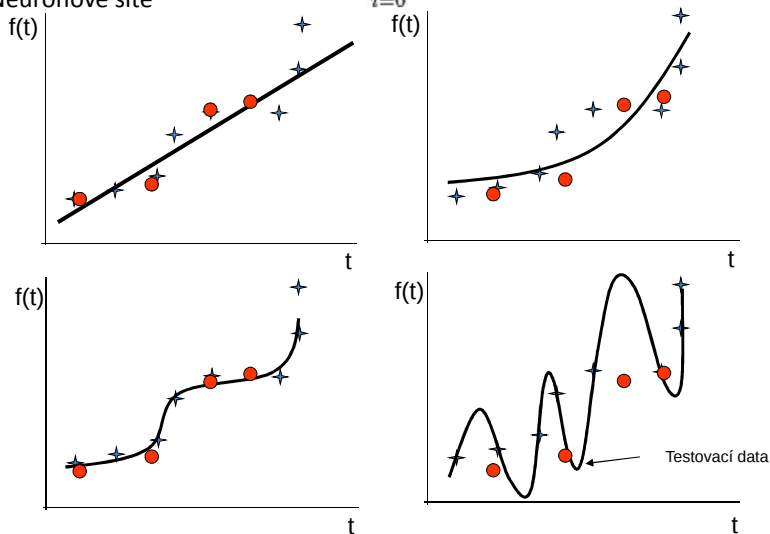
- **příliš jednoduchý model**
 - není schopen popsat složitější procesy v datech.
- **příliš komplexní model**
 - dojde k jeho nakalibrování nejen s ohledem na obecné trendy v trénovací množině, ale i všech **náhodných procesů** a složek zachycených v těchto datech.
 - Toto je ekvivalentní procesu učení adaptivních systémů, například neuronových sítí, proto se používá výraz přetrénování.

Ondřej Přibyl K611MDS

Přetrénování (Overfitting)

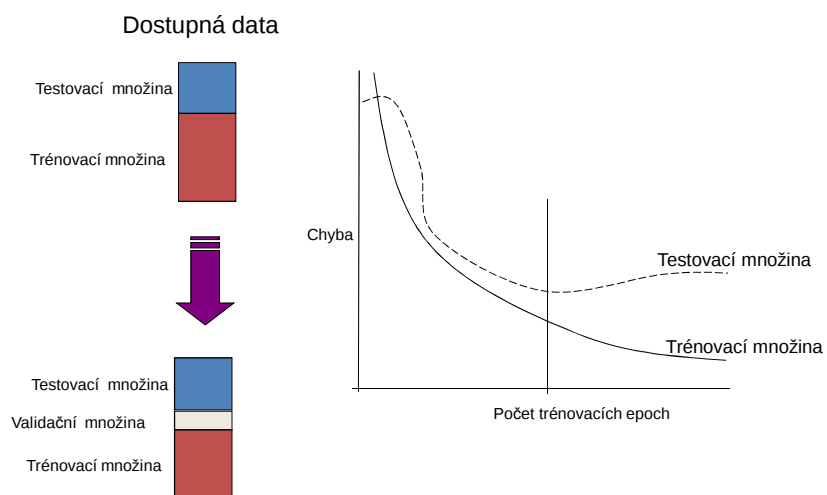
- Polynom n-tého řádu
- Neuronové sítě

$$p(x) = \sum_{i=0}^n a_i x^i = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$$



Ondřej Přibyl K611MDS

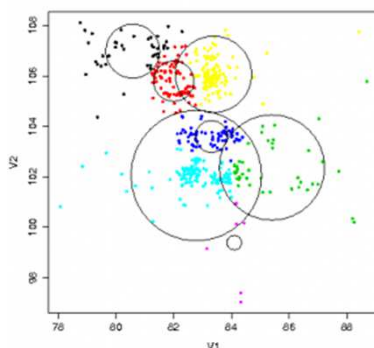
Jak odstranit problém přetrénování?



Ondřej Přibyl K611MDS

Diskuse

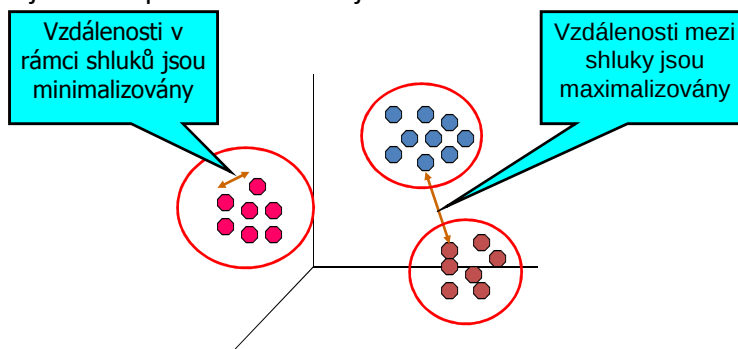
- Co je shlukování?
 - Jedná se o učení s učitelem či bez učitele?



Ondřej Přibyl K611MDS

Shlukování - Cíle

- Nalézt přirozené shluky dat.
- Každý objekt je reprezentován vektorem I atributů.
- Shlukování se uskutečňuje na **základě podobnosti** mezi jednotlivými objekty, tak že objekty v každém shluku jsou si navzájem více podobné než objektům v ostatních shlucích.



Ondřej Přibyl K611MDS

Diskuze

- Co znamená míra podobnosti?
- Jak by jste vyjádřili míru podobnosti pro:
 - Numerická data
 - A (3,10) ;
 - B (6, 6)
 - Binární data
 - A (00101101) ;
 - B (11011111)
 - Speciální data
 - např. intenzita dopravy (časová řada)

Ondřej Přibyl K611MDS

Míra podobnosti objektů

Numerická data

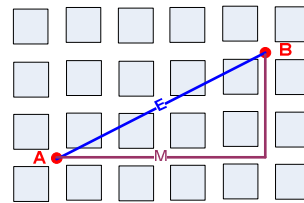
- Minkowského metrika

- E - Euklidovská vzdálenost (Lambda = 2)
- M - Manhattan vzdálenost - městský blok (Lambda = 1)

$$\left(\sum_{k=1}^p (x_k(i) - x_k(j))^{\lambda} \right)^{\frac{1}{\lambda}}$$

Binární data

- Například počtem souhlasných bitů
- $\text{distance}(a,b) = \{\text{Počet } a_i \neq b_i\} / i$



Ondřej Přibyl K611MDS

Dělení metod shlukové analýzy

- Hierarchické metody
- Optimizační metody

Ondřej Přibyl K611MDS

Dělení metod shlukové analýzy

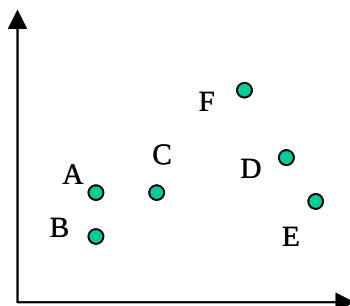
Hierarchické metody

- Pracují **iterativně** pro všechny hodnoty počtu shluků K .
- Existují dva základní principy:
 - **aglomerační** (ze zdola nahoru) - začínají tak, že každý objekt je zároveň samostatným shlukem. V každém kroku pak dojde ke spojení dvou nejpodobnějších shluků a zároveň je vypočítána poloha centra tohoto shluku
 - **rozdělovací** (ze shora dolů) - začínají tak, že všechny objekty patří do jediného shluku. V každém kroku potom dojde k rozdělení tohoto shluku v nejvhodnějším místě.
- Výhody
 - vytvoří shluky pro všechny hodnoty K .

Ondřej Přibyl K611MDS

Příklad hierarchického shlukování

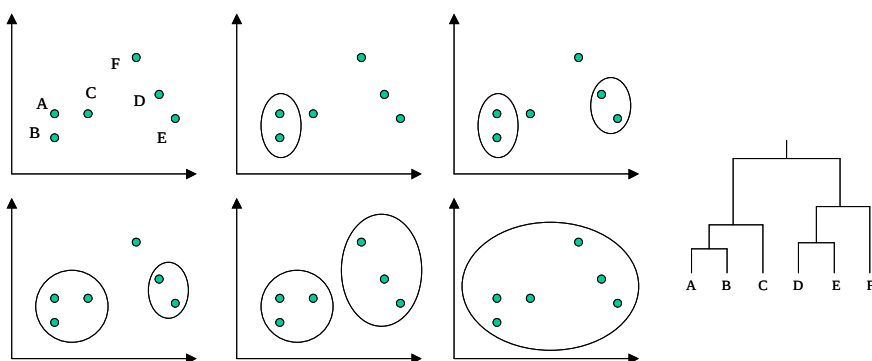
- Zadání
 - Úkolem je najít shluky v příkladě šesti 2D objektů, $\{A, B, C, D, E, F\}$



Ondřej Přibyl K611MDS

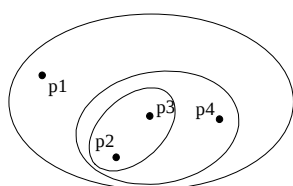
Příklad hierarchického shlukování

- **Dendrogram** - graf na jehož ypsilonové ose jsou vzdálenosti mezi jednotlivými shluky.

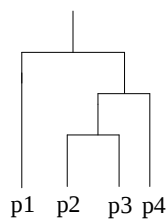


Ondřej Přibyl K611MDS

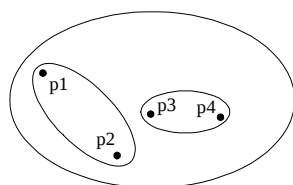
Hierarchické shlukování



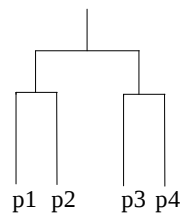
Tradiční shlukování



Tradiční Dendrogram



Alternativní shlukování



Alternativní Dendrogram

Ondřej Přibyl K611MDS

Diskuze

- Jaká jsou omezení hierarchických metod?

Ondřej Přibyl K611MDS

Problémy hierarchického shlukování

- **neoptimalizují žádnou účelovou funkci.**
- **hledají v každém kroku nejlepší možné řešení**, ovšem to nezaručuje dosažení globálního optima.
- neumožňují následně měnit shluky. Jakmile jednou dojde ke spojení (či rozdělení) dvou shluků, jedná se o spojení konečné.
- jsou také poměrně **citlivé na odchylky**, šum a okrajová data.
- V neposlední řadě jsou tyto algoritmy pro větší objemy dat poměrně **náročné na výpočetní čas**.

Ondřej Přibyl K611MDS

Optimalizační metody shlukování

- Na rozdíl od hierarchických metod, **vytváří požadovaný počet shluků** najednou.
- Z toho je zřejmé, že **počet shluků musí být znám** před analýzou a je jedním ze vstupních parametrů těchto metod.
- Nejznámější je metoda zvaná k-means

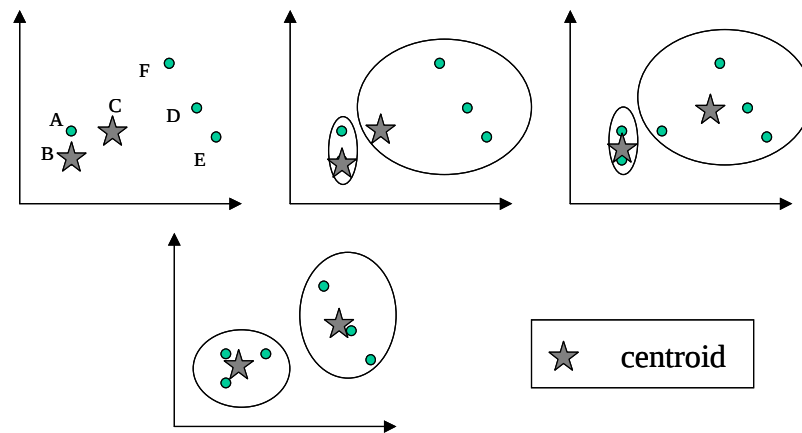
Ondřej Přibyl K611MDS

Metoda K-means

- Iterativní metoda
- Začíná s náhodně generovanými počátečními shluky. Každý shluk je reprezentován svým geometrickým středem nazývaným centroid.
- Během jednotlivých iterací jsou tyto centroidy posouvány tak, aby se zvýšila podobnost objektů v rámci jednotlivých shluků a zvýšila odlišnost objektů náležejícím do různých shluků. Celý proces je popsán následujícími kroky:
 - Vybrat náhodně K objektů (centroidů) které reprezentují jednotlivé shluky
 - Přiřadit každý objekt v databázi k nejbližšímu centroidu.
 - Spočítat nový centroid každého shluku (geometrický průměr)
 - Krok 2 a 3 se opakuje dokud dochází ke změně ve shlucích

Ondřej Přibyl K611MDS

Příklad k-means



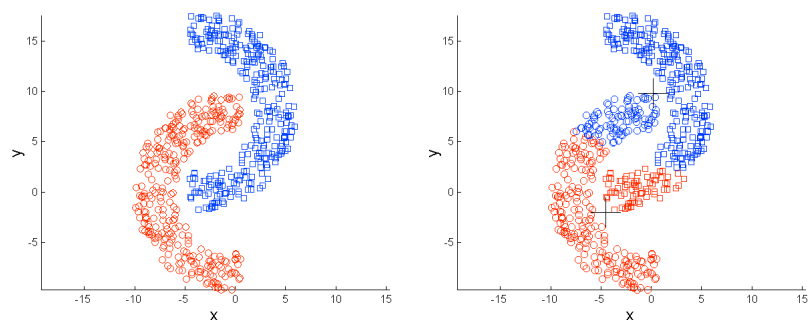
Ondřej Přibyl K611MDS

Diskuze

- What are limitation of a k-means algorithm?
- How to overcome these limitations?

Ondřej Přibyl K611MDS

Omezení k-means: Nekulovité tvary v datech

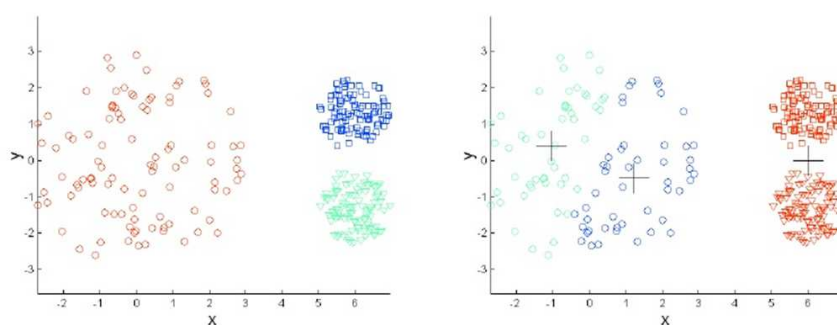


Data

Zdroj: <http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes>

Ondřej Přibyl K611MDS

Omezení k-means: Různá hustota dat



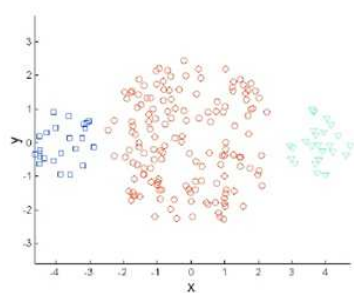
Data

K-means (3 Shluky)

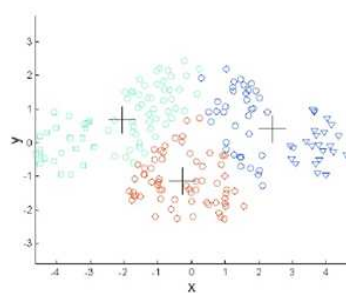
Zdroj: <http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes>

Ondřej Přibyl K611MDS

Omezení k-means: Různá velikost shluků



Data

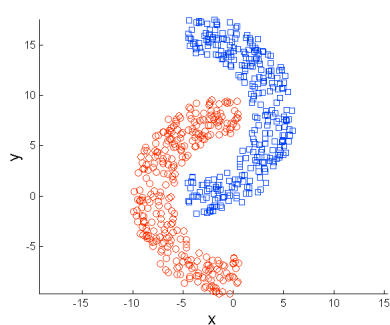


K-means (3 Shluky)

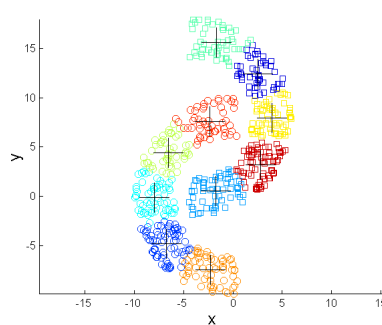
Zdroj: <http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes>

Ondřej Přibyl K611MDS

Overcoming K-means Limitations



Original Points



K-means Clusters

Zdroj: <http://www.scribd.com/doc/6646055/16/Limitations-of-K-means-Non-globular-Shapes>

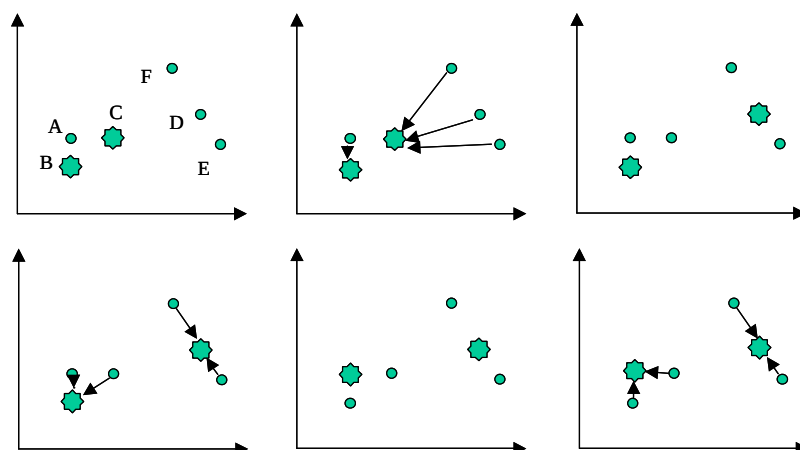
Ondřej Přibyl K611MDS

Problémy k-means algoritmu

- Jedná se o velmi jednoduchý algoritmus, který vždy najde řešení.
- Bohužel je velmi citlivý k volbě prvních centroidů a v závislosti na nich může najít pouze lokální optimum.
- Omezení pro
 - různé hustoty dat
 - různé velikosti shluků
 - nekulovité tvary dat
- Tento algoritmus je také omezen na data v Eukleidovském prostoru. Nemůže být použit pro shlukování kategoričských dat.
 - Existuje řešení, například použití k-medoids algoritmu
 - Již se zde nepočítá geometrické centrum každého shluku, ale algoritmus nalezne objekt, který je mu nejbližší a který potom tento shluk reprezentuje

Ondřej Přibyl K611MDS

Příklad k-medoids algoritmu



Ondřej Přibyl K611MDS

Jak určit optimální počet shluků?

- Počet shluků K musí být zadán před zahájením shlukování.
- Jak jej určit?
- Obvyklým řešením je, že se shlukování provede pro celou řadu počtu shluků. Následně se provede analýza a vybere nejvhodnější řešení.
- Analytické řešení
 - Na základě vzdáleností uvnitř a mezi shluky
 - Silhouette koeficienty

Ondřej Přibyl K611MDS

Silhouette koeficienty

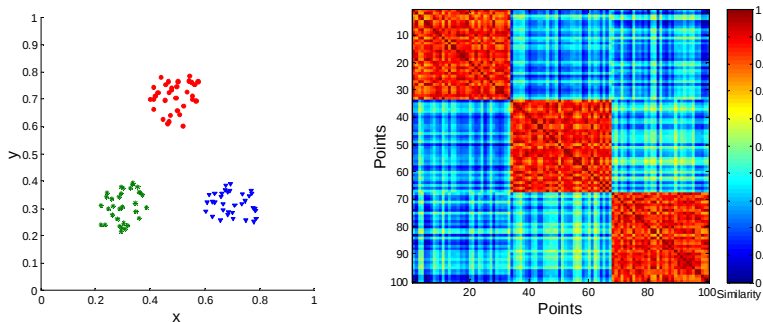
- Pro každý objekt i bude podle následující rovnice spočítána jeho takzvaná silueta, která v podstatě porovnává spojitost každého shluku vůči vzdálenosti mezi shluky
- kde a_i udává průměrnou vzdálenost objektu i vůči ostatním objektům v tomtéž shluku, a b_i udává průměrnou vzdálenost objektu i vůči všem objektům v dalším nejbližším shluku (další nejlepší kandidát pro umístění objektu i).

- Je zřejmé, že je definováno na intervalu $(-1,1)$.
$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

SC	Doporučený význam
0.71 – 1.00	Byla nalezena silná struktura
0.51 – 0.70	Byla nalezena dostatečná struktura
0.26 – 0.50	Nalezená struktura je slabá a doporučuje se vyzkoušet jiný přístup
< 0.25	Žádná podstatná struktura nalezena nebyla

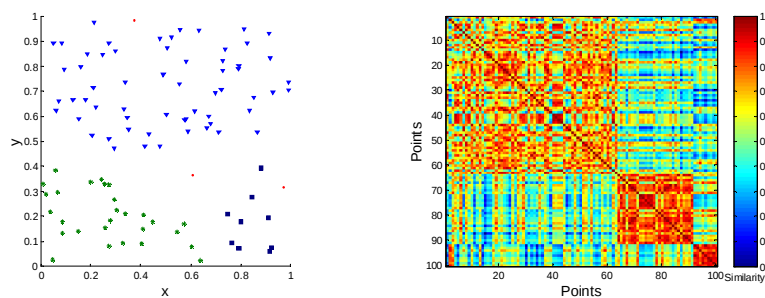
Ondřej Přibyl K611MDS

matice podobnosti pro ověření kvality shluků



Ondřej Přibyl K611MDS

matice podobnosti pro ověření kvality shluků



Ondřej Přibyl K611MDS

ROZHODOVACÍ STROMY

Co jsou rozhodovací stromy?

- Analytický nástroj
 - Popisuje vztah mezi nezávislými a závislými proměnnými
- Podporuje rozhodování
- Používají se často v marketingu – segmentace trhu
 - Jaká cílová skupina kupuje naše zboží?
- Výsledný rozhodovací strom je snadno čitelný pro člověka
 - Rozhodovací pravidla (IF-THEN rules)

Základní dělení

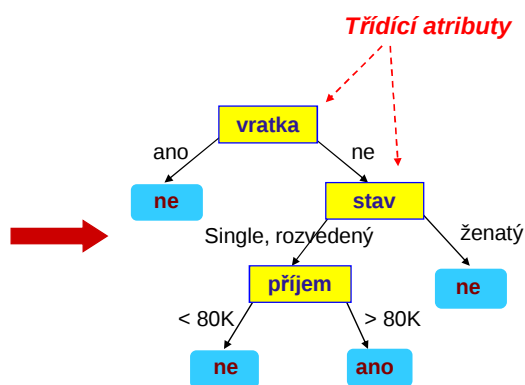
- Binární x Multinomiální
- Spojité x diskrétní x kombinované hodnoty proměnných

Ondřej Přibyl K611MDS

Příklad

Tid	categorical		categorical		continuous		class
	Refund	Marital Status	Taxable Income	Cheat			
1	Yes	Single	125K	No			
2	No	Married	100K	No			
3	No	Single	70K	No			
4	Yes	Married	120K	No			
5	No	Divorced	95K	Yes			
6	No	Married	60K	No			
7	Yes	Divorced	220K	No			
8	No	Single	85K	Yes			
9	No	Married	75K	No			
10	No	Single	90K	Yes			

Trénovací množina



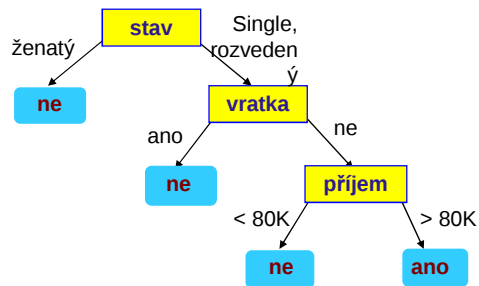
Model: Rozhodovací strom

Ondřej Přibyl K611MDS

Jiný model

categorical
categorical
continuous
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

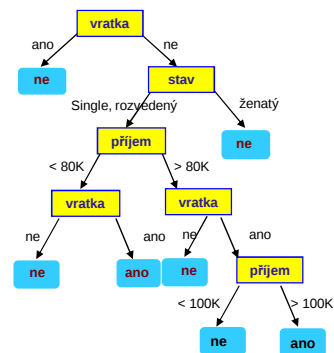
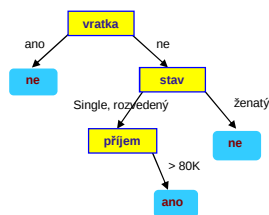


There could be more than one tree that fits the same data!

Ondřej Přibyl K611MDS

Diskuse

Jaký z následujících modelů zvolit?



Okamova břitva:

Pokud pro nějaký jev existuje vícero vysvětlení, je lépe upřednostňovat to nejméně komplikované.

Ondřej Přibyl K611MDS

strana 38

Diskuse

- Jaké jsou metody indukce rozhodovacích stromů?
- Jaká jsou kritéria pro výběr algoritmu rozhodovacího stromu?

Ondřej Přibyl K611MDS

Indukce rozhodovacích stromů

- Mnoho různých algoritmů
 - CART
 - C4.5
 - CHAID
- Jedná se o „**greedy**“ metody
 - Vždy zvolí optimalizaci v každém kroku,
Může vést k nalezení lokálních extrémů
- **Otázky**
 - Jak rozdělit záznamy?
 - Jaké zvolit atributy?
 - Jaké zvolit hranice?
 - Kdy skončit dělení?

Ondřej Přibyl K611MDS

Příklad

Otázka: Bude mi po tomto nápoji špatně či nikoli?

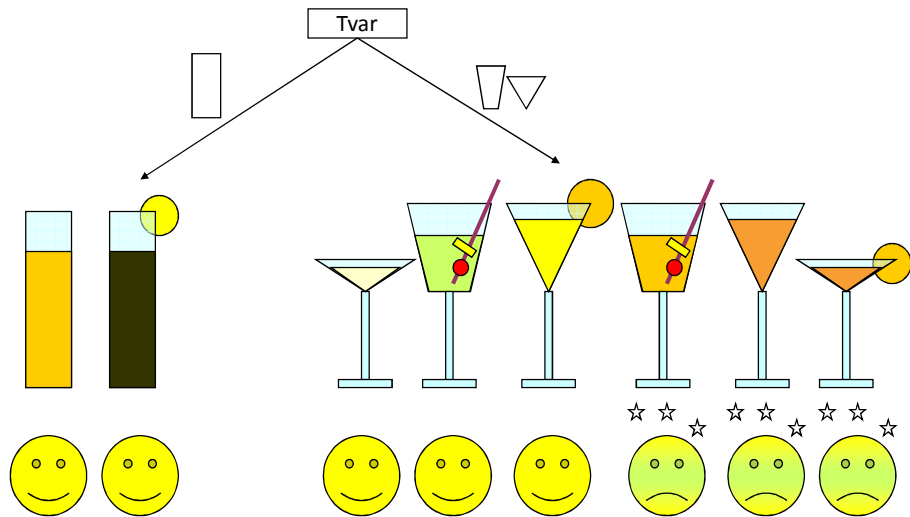
Zkušenosti

Ondřej Přibyl K611MDS Zdroj: Kurt Driessens

Množina: 8 klasifikovaných prvků

Ondřej Přibyl K611MDS Zdroj: Kurt Driessens

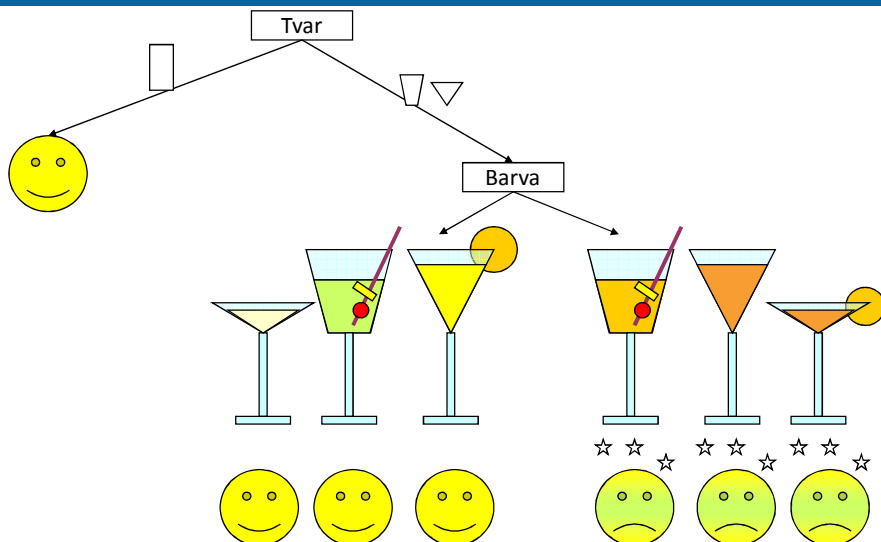
Pozorování 1: TVAR je důležitý



Ondřej Přibyl K611MDS

Zdroj: Kurt Driessens

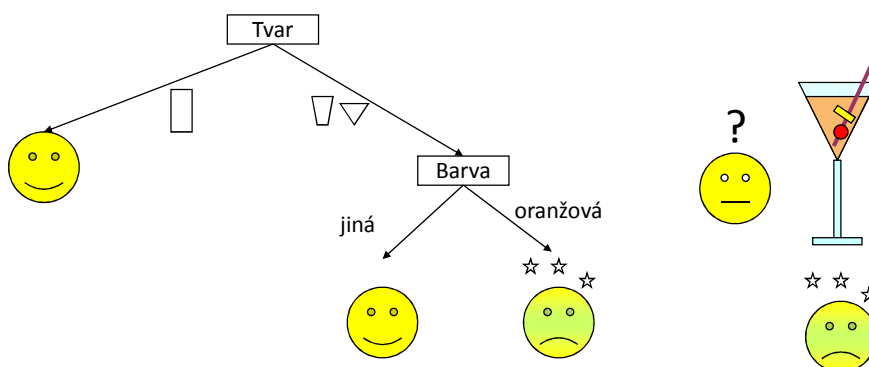
Pozorování 2: Pro některé tvary je důležitá BARVA



Ondřej Přibyl K611MDS

Zdroj: Kurt Driessens

Dedukce



Ondřej Přibyl K611MDS

Indukce rozhodovacích stromů - princip

Indukce:

- Vyber atribut, **který nejvíce popisuje** výstupní atribut (který od sebe nejlépe odliší příklady z různých tříd)
- Rozděl tento atribut tak aby bylo maximalizováno určité kritérium
 - **GINI index** ("Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset")
 - **Entropie** (udává homogenitu atributu)
 - **Počet chybně klasifikovaných vzorků**
 - ...

Ondřej Přibyl K611MDS

Princip algoritmu TDIDT

- TDIDT: (Top Down Induction of Decision Trees)
- Indukce stromů metodou shora dolů (rozděl a panuj)

TDIDT algoritmus

1. zvol jeden atribut jako kořen dílčího stromu,
2. rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu,
3. existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči.

Ondřej Přibyl K611MDS

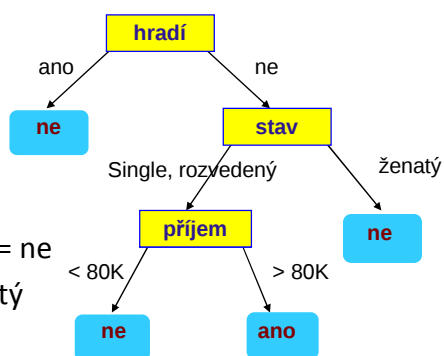
Výstupy

• Rozhodovací strom

- Podvádí či ne?

• Rozhodovací pravidla

- IF hradí = ano THEN podvádí = ne
- IF hradí = ne AND stav = ženatý THEN podvádí = ne
- IF hradí = ne AND stav = Single, rozvedený AND příjem < 80K THEN podvádí = ne
- ELSE podvádí = ano



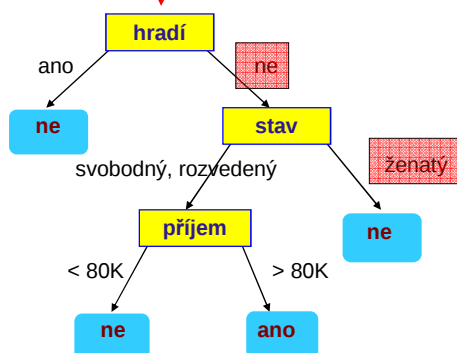
Ondřej Přibyl K611MDS

Aplikování modelu

Začít od kořene stromu

Testovací data

Hradí	Stav	příjem	podvádí
No	Ženatý	80K	?



Podvádí? = ne

Ondřej Přibyl K611MDS

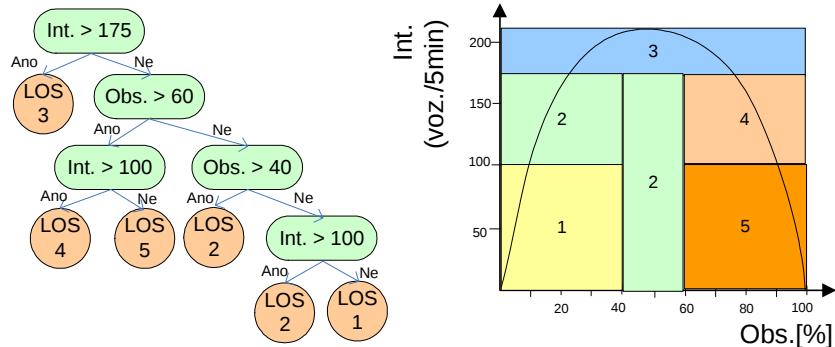
Diskuze

- Jmenujte některé aplikace z oblasti dopravy / ITS.
- Vlastnosti rozhodovacích stromů? (výhody, nevýhody)

Ondřej Přibyl K611MDS

Příklad využití v dopravě

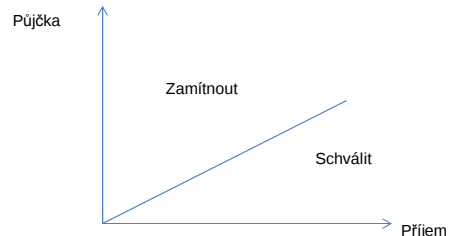
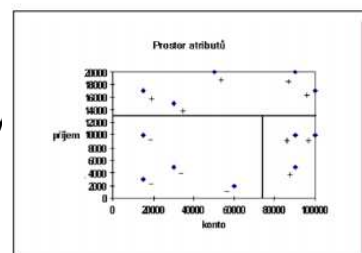
- Automatický klasifikátor stupně dopravy
- Pracuje v grafu závislosti intenzita / obsazenost



Ondřej Přibyl K611MDS

Rozhodovací stromy - vlastnosti

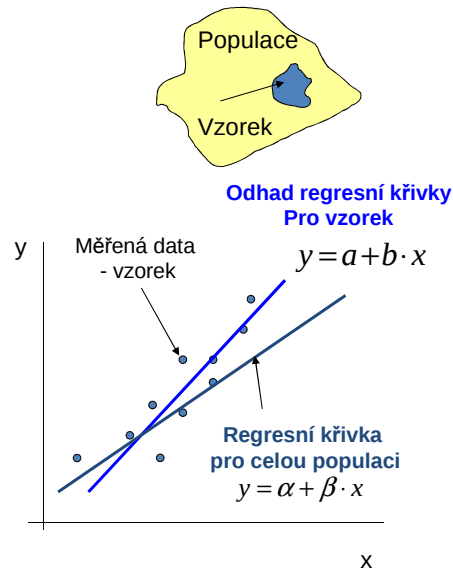
- Výhody
 - Akceptují chybějící hodnoty
 - Akceptují spojité i diskrétní hodnoty
 - Výsledek ve formě snadno pochopitelných pravidel
 - Nalezne důležité proměnné
- Nevýhody
 - Vhodné pro oblasti s obdélníkovými oblastmi
 - Mohou být příliš velké pro některé aplikace
 - Doplnující prořezávání



Ondřej Přibyl K611MDS

Princip metody lineární regrese

- Předpoklad
 - x...nezávislé proměnné
 - y...závislá proměnná
 - koeficienty
- Hledáme odhad těchto koeficientů
 - a, b α, β $y = \alpha + \beta \cdot x$
- Metody odhadu
 - Nejmenší čtverce



Ondřej Přibyl K611MDS

strana 53

Lineární regrese - vlastnosti

- Hledáme vztah, lineární regresní model, který sumarizuje soubor dat
- Jedná se o vyčíslení nejlepších odhadů neznámých parametrů regresního modelu
- Snahou je nalézt odhady b parametrů β , které jsou
 - Nejlepší (Best),
 - Nestranné (Unbiased) a
 - Lineární (Linear).
- a mají asymptoticky normální rozdělení.



(NNLO/BLUE)

Ondřej Přibyl K611MDS

Diskuse

- Jak je možné dosáhnout nalezení NNLO (BLUE) odhadů a mají asymptoticky normální rozdělení?
- Za splnění základních předpokladů:

Ondřej Přibyl K611MDS

Předpoklady metody nejmenších čtverců

- Regresní parametry β mohou nabývat libovolných hodnot.
- Regresní model je lineární v parametrech.
- Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných X má hodnotu rovnou právě m . To znamená, že žádné její dva sloupce x_j , x_k nejsou kolineární, tj. rovnoběžné
- Náhodné chyby g_i mají nulovou střední hodnotu $E(g_i) = 0$.
- Náhodné chyby g_i mají konstantní a konečný rozptyl $E(g_i^2) = \sigma^2$. Také podmíněný rozptyl $D(y/x) = \sigma^2$ je konstantní.
- Náhodné chyby g_i jsou vzájemně nekorelované a platí $\text{cov}(g_i, g_j) = E(g_i g_j) = 0$.
- Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin y .
- Chyby g_i mají normální rozdělení $N(0, \sigma)$.

Ondřej Přibyl K611MDS

Vlastnosti lineární regrese (zjednodušeně)

- Výhody
 - Velmi kvalitně zpracované metody
 - Poměrně jednoduché
 - Výpočetně nenáročné
- Nevýhody
 - Téměř nikdy nemohou být splněny všechny předpoklady
 - Existuje řada opatření a jiných statistických modelů, která se to snaží vyřešit (Logit, Probit, ...)
- Alternativní řešení
 - Použití jiných metod
 - Rozhodovací stromy
 - Soft-computing (FS, ANN, GA, ...)

Ondřej Přibyl K611MDS

Děkuji za pozornost

Ondřej Přibyl K611MDS

Témata ke zkoušce

- Učení s učitelem/bez učitele
- Princip přetrénování a jak jej odstranit
- Shluková analýza
 - Co je shluková analýza
 - Jaké jsou přístupy a jednotlivé metody
 - Míra podobnosti pro různé typy dat
 - Výhody a nevýhody, kdy použít
- Rozhodovací stromy
 - Co jsou rozhodovací stromy
 - Jaké jsou přístupy a jednotlivé metody, postup při indukci
 - Výhody a nevýhody, kdy použít
- Lineární regrese
 - Jak se spočítá
 - Jaká jsou omezení a základní vlastnosti
 - Populace versus vzorek, rozdíly omezení

Zdroje:

- <http://www.public.asu.edu/~kirkwood/DASstuff/decisiontrees/index.html>
- <http://www.cs.cmu.edu/afs/cs/Web/People/awm/tutorials/dtree.html>
- http://www.norusis.com/pdf/SPC_v13.pdf
- <http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf>
- <http://www.cs.cmu.edu/afs/cs/Web/People/awm/tutorials/>