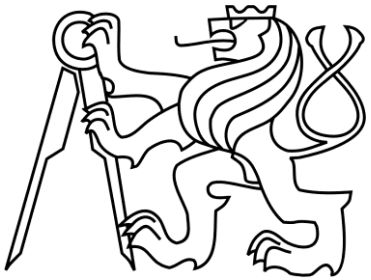
Předzpracování dat

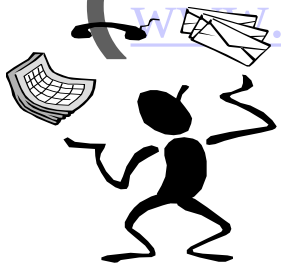
Lenka Vysloužilová

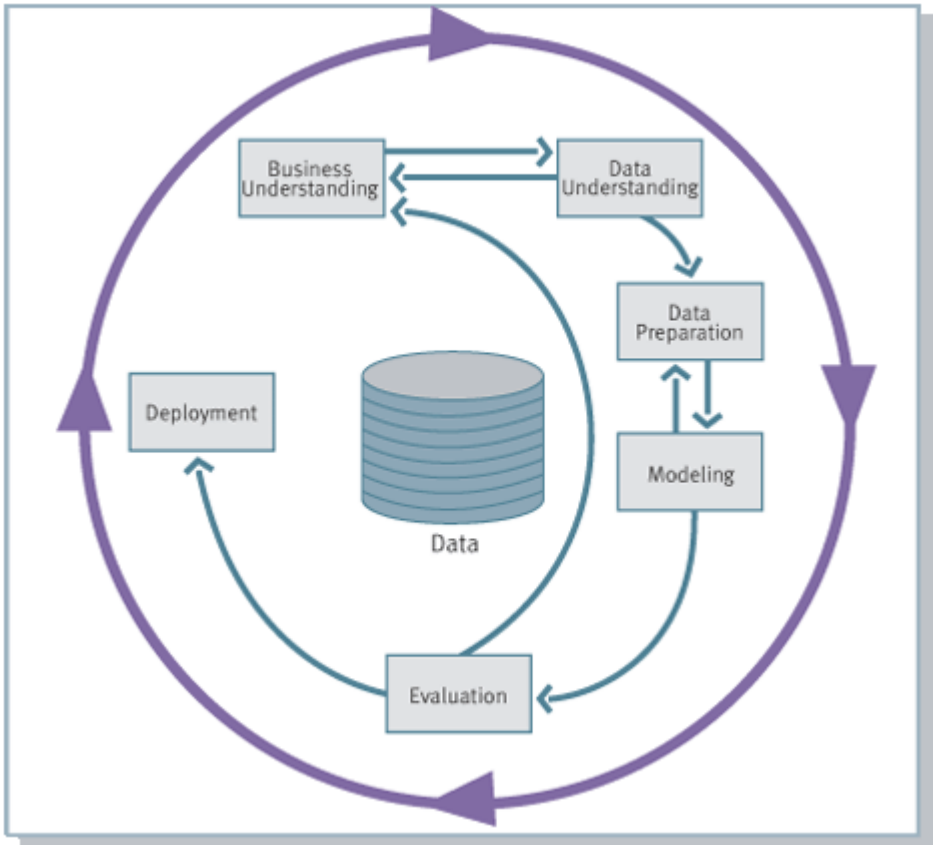


Nature Inspired
Technologies Group

Metodika CRISP-DM

(www.crisp-dm.org)





Příprava dat – Data Preparation

- příprava dat pro modelování
  - selekce příznaků – výběr relevantních příznaků
  - čištění dat
  - získávání odvozených příznaků
  - převod typů dat
  - transformace dat do jedné velké tabulky
  - formátování pro jednotlivé modelovací techniky
- nejpracnější část celého procesu
- často se provádí opakovaně

Transformace dat do jedné tabulky

- 1:1
  - prakticky pouze doplnění tabulky o nové atributy
- 1:N
  - vytvoření agregovaných hodnot
  - součet, min, max, průměr, regresní křivka
  - majoritní hodnota, počet různých hodnot, výskyt konkrétní hodnoty
  - do této skupiny patří časové řady
- M:N
  - nutná volba úlohy, zda chceme 1:N nebo 1:M
- Propozicionalizace

Datová tabulka

Filtrování a úprava instancí

Sepallength	Sepalwidth	Petallength	Petalwidth	Class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.7	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica

Filtrování a úprava příznaků

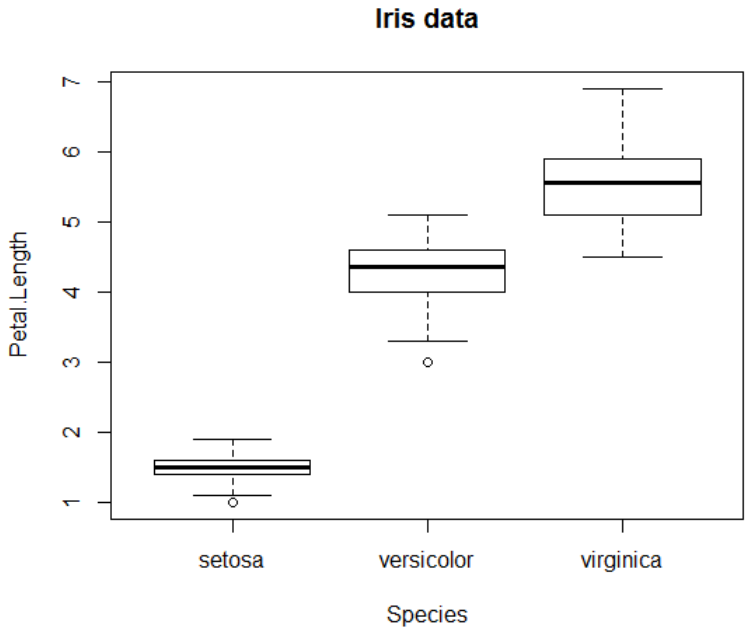
ÚPRAVA INSTANCÍ

Náhrada chybějících hodnot

- nedělat nic, náhrada hodnotou „nevím“
  - některým algoritmům chybějící hodnoty nevadí, např. rozhodovací stromy
  - Not Available NA
- ignorovat celou instanci
  - ideální pro data s minimem chybějících hodnot
  - newdata <- na.omit(mydata)
- náhrada
  - nejčetnější hodnotou
  - průměrem, mediánem replace(x, is.na(x), median(x, na.rm=T))
  - nalezení nejbližšího souseda
  - využití algoritmu pro modelování

Outliers

- Výrazně odlišné hodnoty atributu pro danou instanci
  - Outlier pro jeden atribut nemusí být outlier i pro kombinaci atributů a naopak!
  - Boxplot



Vzorkování dat

- obrovský počet instancí - pro algoritmy pracující v dávkovém režimu nutnost
  - redukce počtu dat
  - tvorba modelů na základě podmnožin a jejich následná kombinace
- rozdělení dat na trénovací a testovací část
- nevyvážená data např. třída A 95%, třída B 5%
  - každý objekt patří do majoritní třídy
  - různé ceny chybného rozhodnutí
  - výběr dat pro různé třídy s různou pravděpodobností

ÚPRAVA PŘÍZNAKŮ

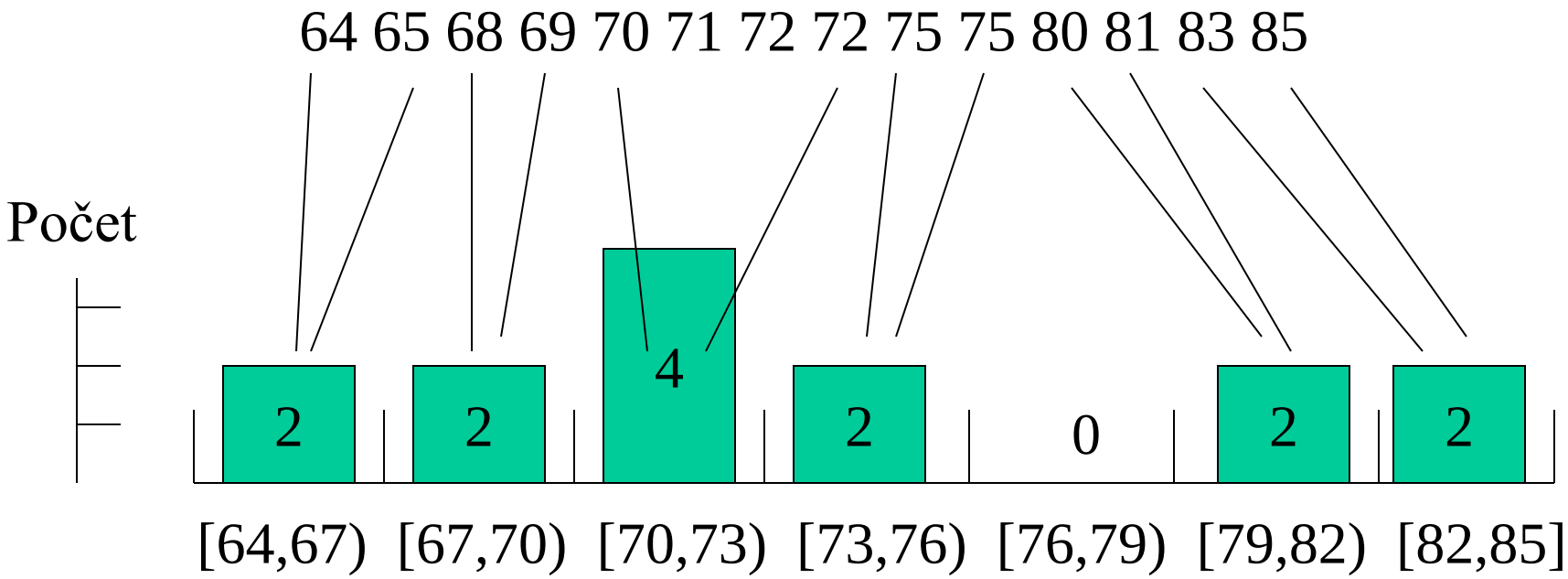
Které příznaky mají význam v DM?

- V případě prediktivních úloh musí jít především o příznaky, jejichž hodnota je známá v okamžiku, kdy chceme predikci provádět.
- Pozor na anachronické příznaky (anachronistic at.), tj. takové, že nesplňují výše uvedený požadavek.
- Příklad. Telefonní operátor a predikce těch, co přecházejí k jinému operátorovi. Mezi 500 použitými atributy se ukázal mít velkou prediktivní sílu atribut odpovídající jménu zaměstnance, který dělal s klientem poslední interview. Později se ukázalo, že jiný člověk měl na starosti klienty, kteří projevili zájem odejít.

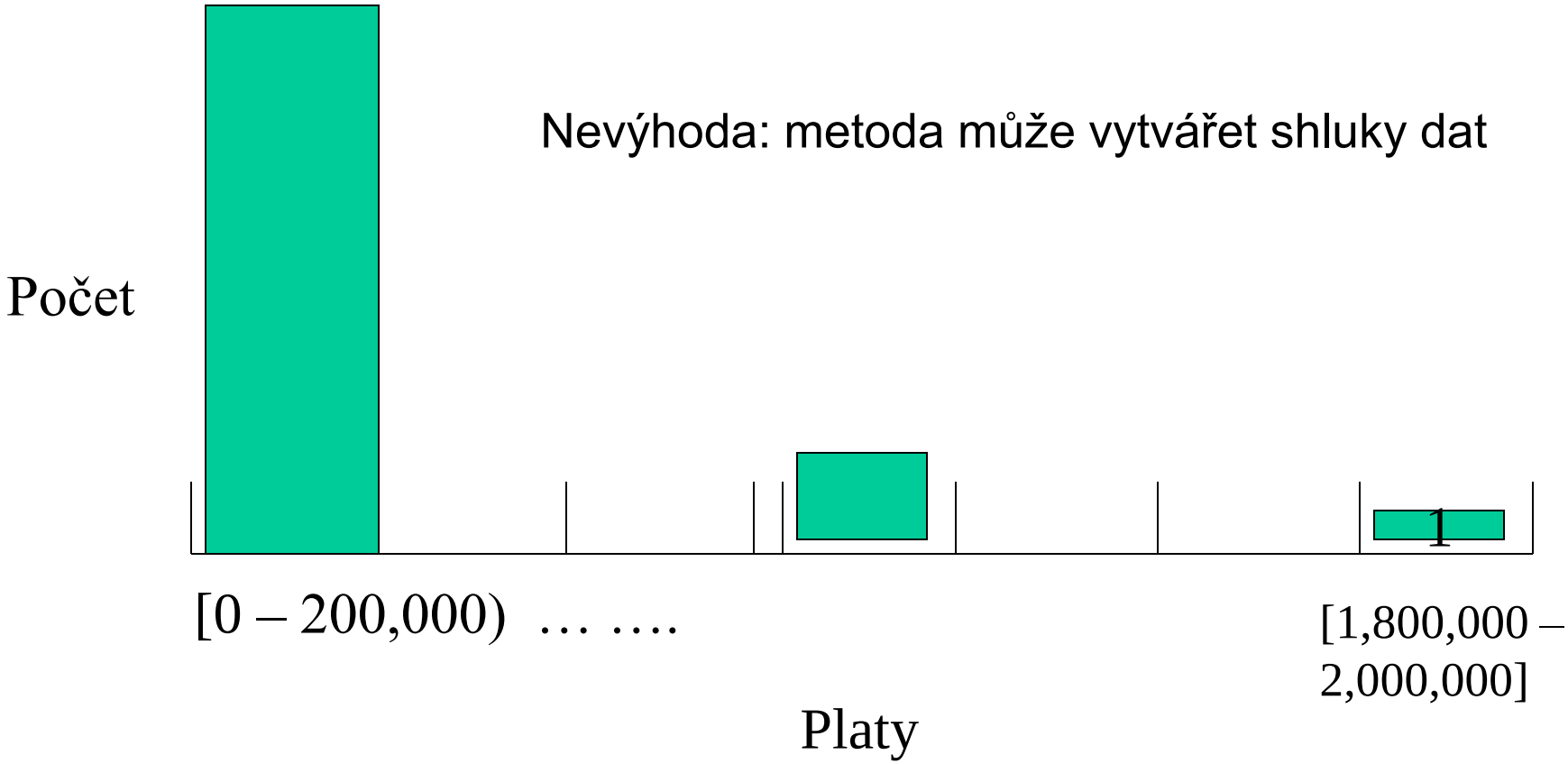
Diskretizace dat

- Neinformované metody
  - ekvidistantní intervaly
  - ekvifrekvenční intervaly
- Informované metody
  - využití znalosti o příslušnosti objekt -> třída
  - strategie rozdělování nebo spojování intervalů

Diskretizace: Ekvidistantní intervaly

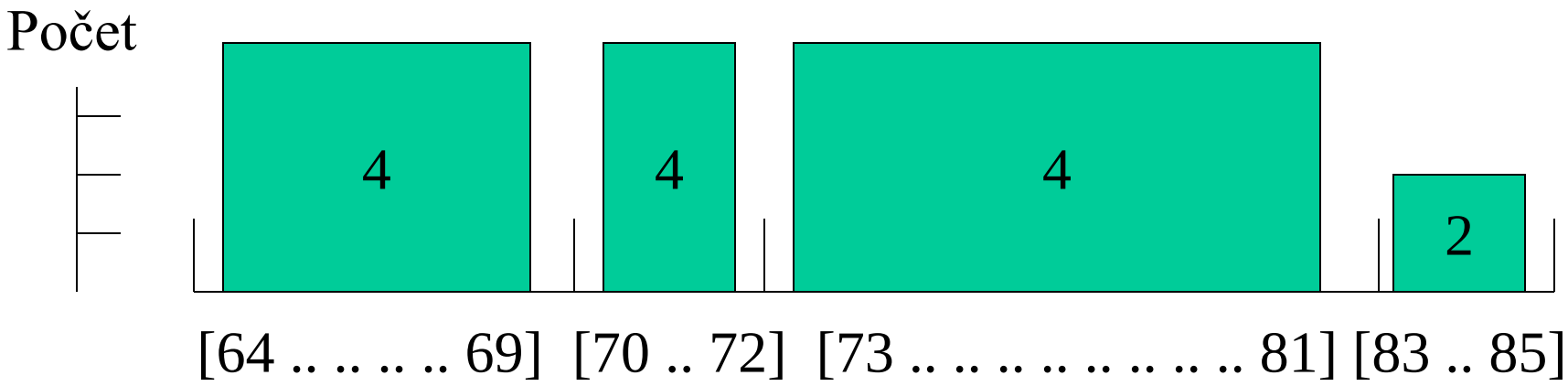


Diskretizace: Ekvidistantní intervaly



Diskretizace: Ekvifrekvenční intervaly

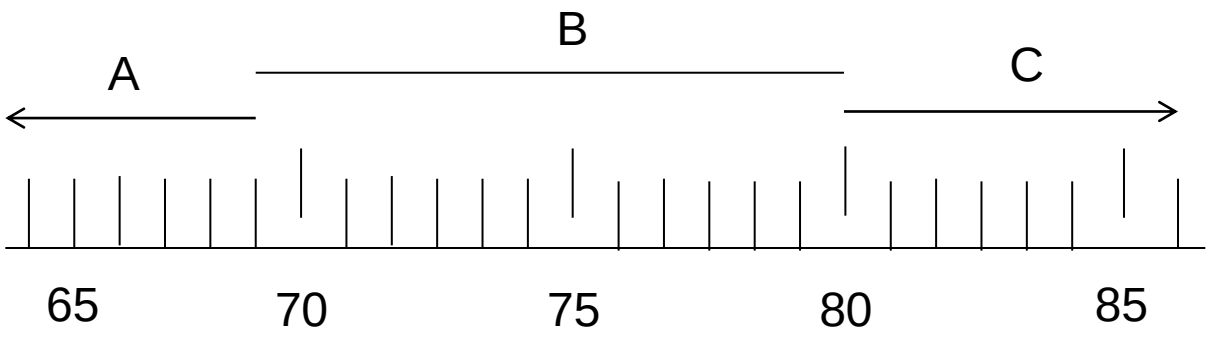
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Diskretizace: v závislosti na třídě

požadujeme minimálně 3 instance na interval

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	No	No	No	Yes	Yes	Yes	No	Yes	Yes	No



Normalizace dat

- Převod numerických hodnot do intervalu <0,1>
- Numerické atributy

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i} \quad \text{nebo} \quad a_i = \frac{v_i - Avg(v_i)}{StDev(v_i)}$$

v_i: aktuální hodnota atributu I

Odvozené atributy

- výpočet nového atributu ze stávajících
- BMI = váha(kg) / výška(m)^2
- rodné číslo => věk a pohlaví
- agregační hodnoty

Redukce počtu příznaků

- Proč je výběr příznaků důležitý?
- Extrakce příznaků
  - PCA principal component analysis
- Selekce příznaků
  - Míru pro měření kvality vybrané podmnožiny příznaků
  - Strategii prohledávání prostoru
  - Metody výběru příznaků

Proč je výběr příznaků důležitý?

Vztah mezi výběrem příznaků a strojovým učení (ML) nebo dobýváním znalostí?

- Předpokládáme-li, že informace o cílové třídě je **implicitně zahrnuta v hodnotách příznaků**, pak

- Můžeme učinit naivní závěr, že mít více příznaků
- je výhodné, neboť tím získáváme
  => víc informací
  => větší rozlišovací schopnost.

- Praktická zkušenost upozorňuje, že **často tomu tak není!**

- Další doplňkový argument:
  Optimalizace je (obvykle) výhodná. Proč se tedy nepokusit o optimalizaci kódování vstupu ?

Věta o PAC učení rozhodovacího stromu

Nechť objekty jsou charakterizovány pomocí ***n*** binárních atributů a nechť připouštíme jen hypotézy ve tvaru rozhodovacího stromu s maximální délkou větve ***k***. Dále nechť δ, ε jsou malá pevně zvolená kladná čísla blízká 0. Pokud algoritmus strojového učení vygeneruje hypotézu φ, která je konzistentní se všemi **m** příklady trénovací množiny a platí

$$m \geq m_{k\text{-DT}}(n) \geq c\,(\,n^k + \ln(1/\delta))\,/\,\varepsilon$$

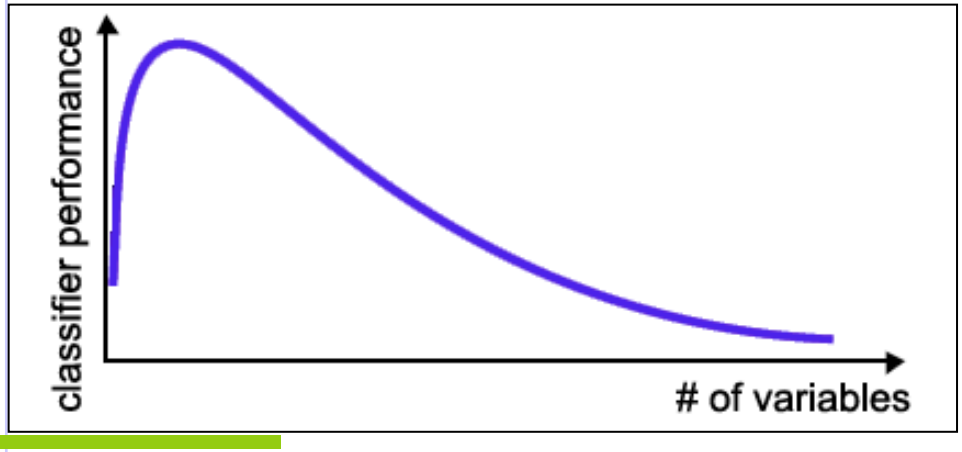
pak φ je ε-skoro správná hypotéza s pravděpodobností větší než $(1-\delta)$, t.j. chyba hypotézy φ na celém definičním oboru konceptu je menší než ε s pravděpodobností větší než $(1-\delta)$.

PAC učení = Probably Approximately Correct Learning

Prokletí dimensionality

- Počet trénovacích příkladů ***m*** potřebných proto, aby byla vytvořena hypotéza o dostatečné přesnosti roste exponenciálně s počtem atributů! PAC: ***m*** > Počet_prvků (Prostor_hypotéz)
- V praktických úlohách bývá maximální počet trénovacích příkladů pevně dán!

=> výkon klasifikátoru (classifier performance) výrazně klesá s rostoucím počtem atributů (# of variables)!



Velmi často lze docílit toho, že

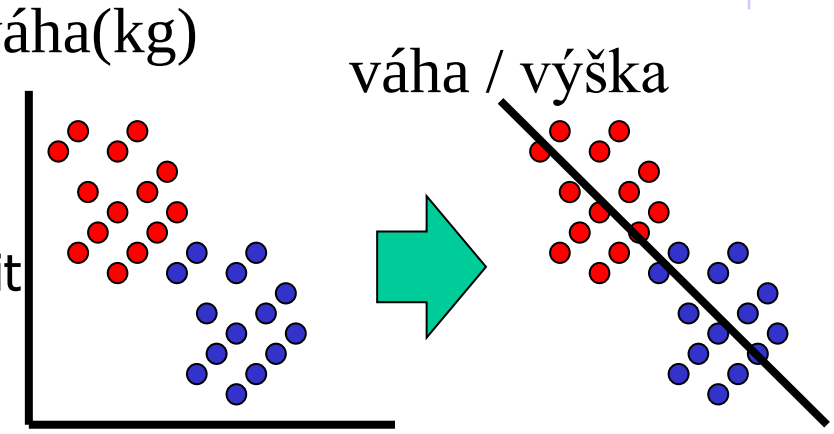
- ztráta informace vzniklá vynecháním některých atributů

je vyvážena

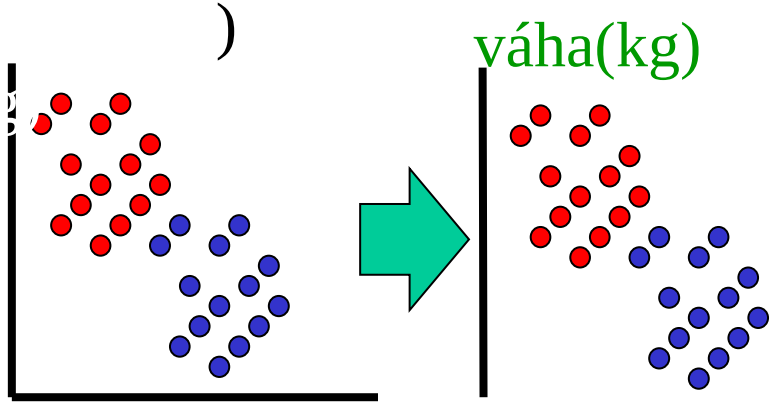
- daleko lepšími výsledky klasifikace v prostoru o nižší dimenzi !

2 cesty k redukci dimenze

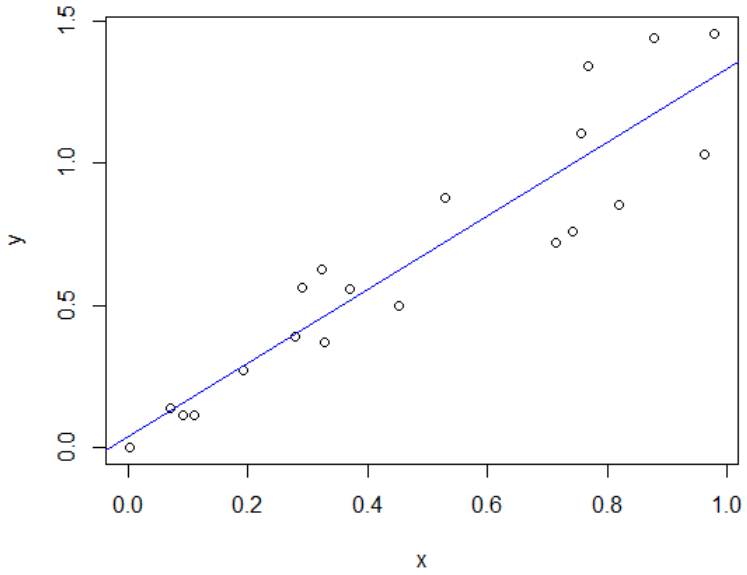
- **Extrakce příznaků** (*Feature Extraction*)
  - Vytvoří nový příznak, který může skupinu jiných nahradit
  - Např. váha/výška

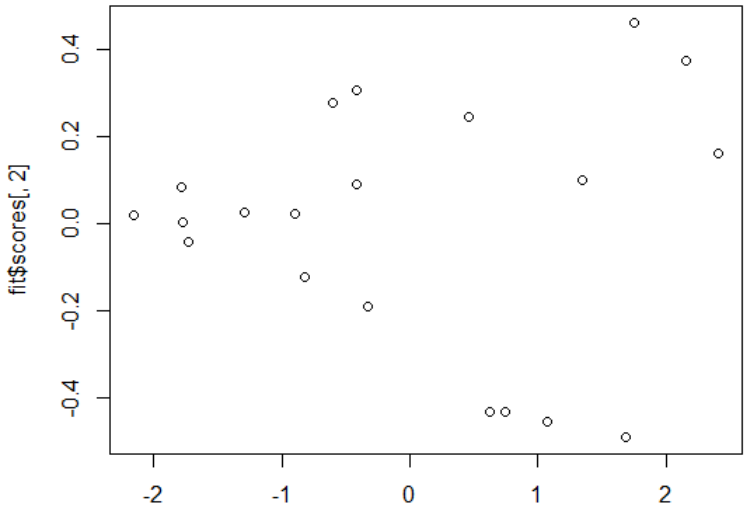


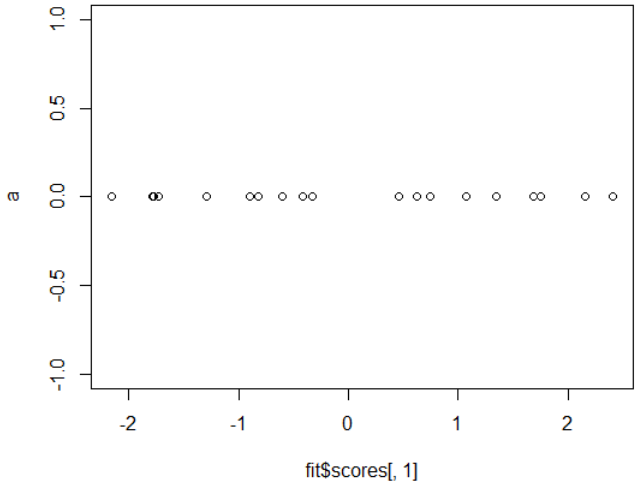
- **Výběr příznaků** (*Feature Selection*)
  - Vybere 1 nebo více příznaků, na které se sosutředí
  - např. zachová p. váha (používá příslušný průmět)
  - V tomto příkladě není klasifikace jednoznačná

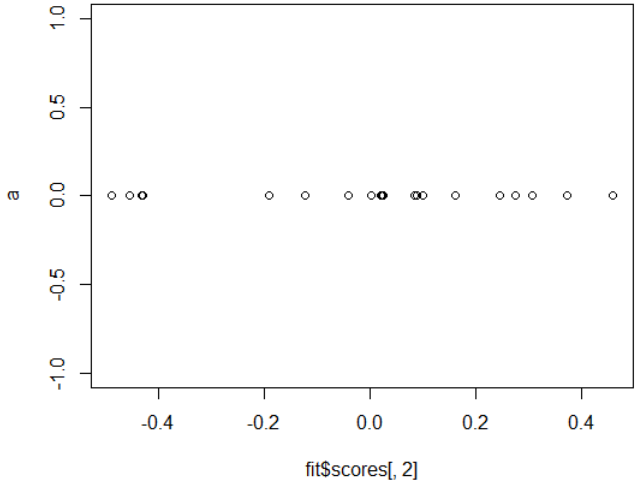


Principal Component Analysis









Výběr podmnožin příznaků

- Potřebujeme:
  - Míru pro měření kvality vybrané podmnožiny příznaků (hodnotící funkci)
  - Strategii prohledávání prostoru všech možných podmnožin

  => Good heuristics are needed!

- Používané metody:
  - Filtrační metody
  - Wrappers
  - Vnořené metody jsou zabudované do jednotlivých algoritmů strojového učení (např. uvnitř ID3)

[9] E. Amaldi, V. Kann: The approximability of minimizing nonzero variables and unsatisfied relations in linear systems. (1997)

Vhodná hodnotící kritéria

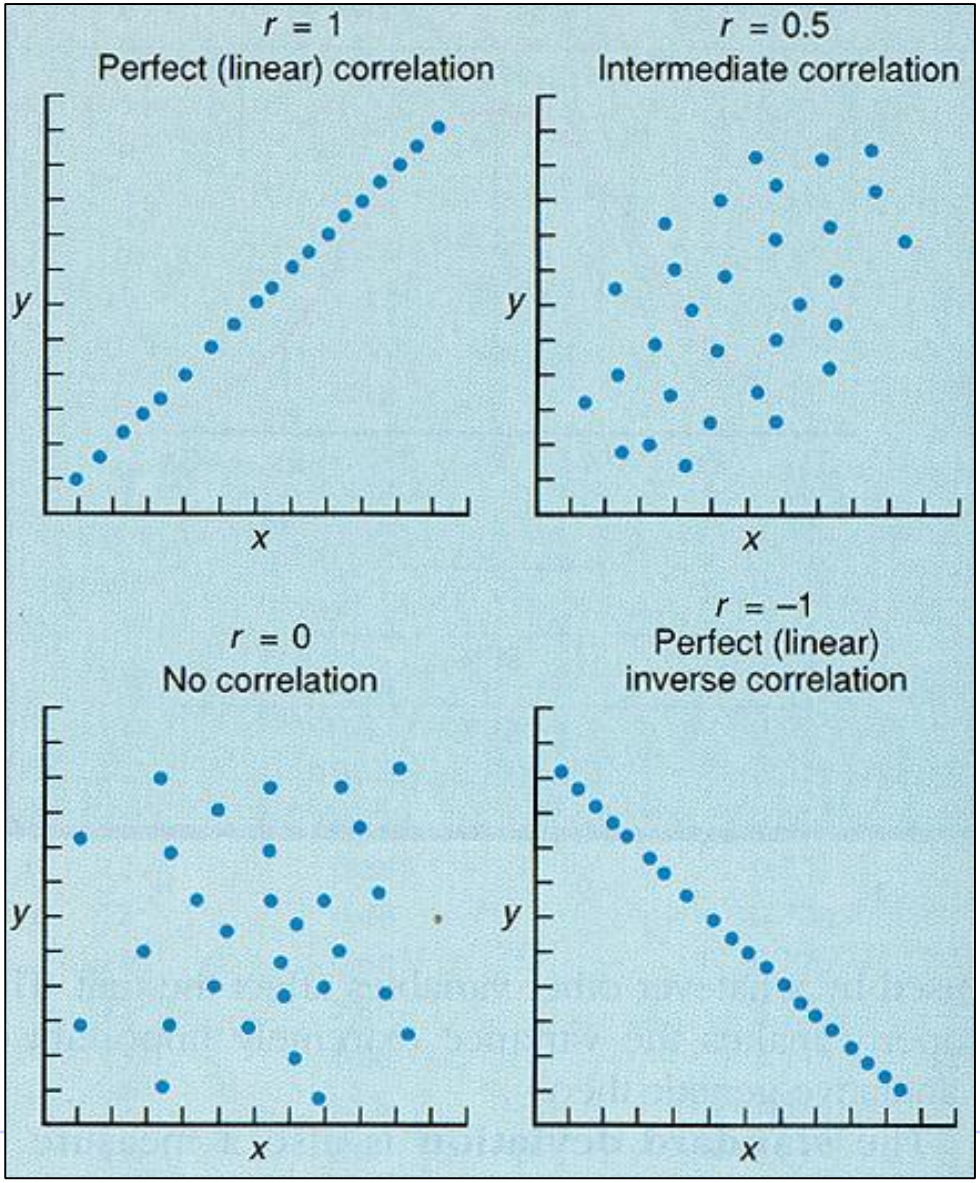
- Pearsonův korelační koeficient

$$\mathrm{R}(f_i, y) = \frac{\mathrm{cov}(f_i, y)}{\sqrt{\mathrm{var}(f_i)\,\mathrm{var}(y)}}$$

- Odhad pro m vzorků:

$$R(f_i, y) = \frac{\sum_{k=1}^{m}\left(f_{k,i} - \overline{f}_i\right)\left(y_k - \overline{y}\right)}{\sqrt{\sum_{k=1}^{m}\left(f_{k,i} - \overline{fi}\right)^2 \sum_{k=1}^{m}\left(y_k - \overline{y}\right)^2}}$$

Hodnotící kriterium – korelace

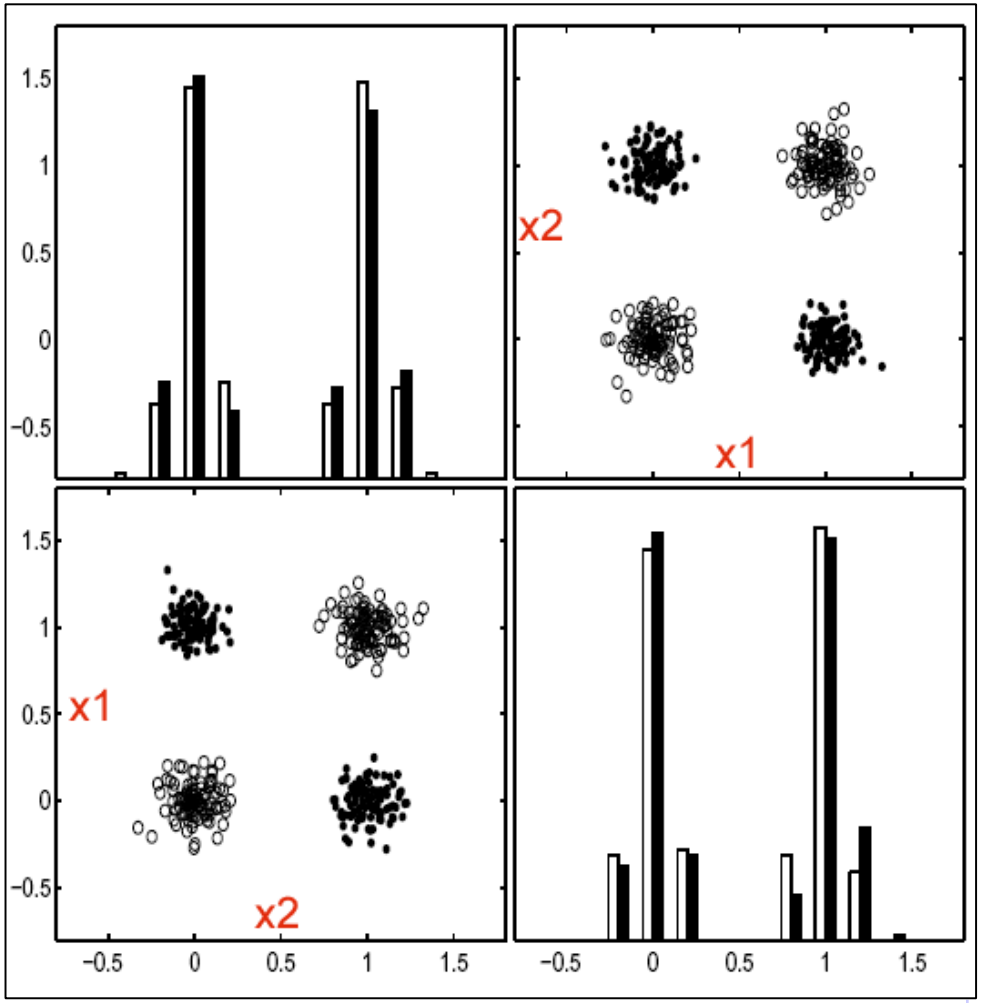


Hodnotící kriterium – korelace

- Může být skupina (alespoň 2) příznaků s nízkým hodnocením užitečná?

ANO!

Je nutné hledat další kriteria



Další hodnotící kriteria

- Korelace
- Chi-kvadrát test, entropie, informační míra závislosti
  - vychází z kontingenční tabulky
- nevýhoda: posuzujeme každý atribut samostatně – množiny atributů

Výběr podmnožin příznaků

- Potřebujeme:
  - Míru pro měření kvality vybrané podmnožiny příznaků
  - Strategii prohledávání prostoru všech možných podmnožin
  - => Good heuristics are needed!
- Používané metody:
  - Filtrační metody
    - Vybírají podmnožinu příznaků obvykle v rámci předzpracování a nezávisle na tom, jaký bude použit klasifikátor!!
  - Wrappers
    - Wrapper prohledává prostor všech možných podmnožin příznaků a každou zvažovanou podmnožinu hodnotí tak, že její kvalitu otestuje na trénovacích datech s použitím nějakého učicího algoritmu
  - Vnořené metody
    - jsou zabudované do jednotlivých algoritmů strojového učení (např. uvnitř ID3)

Závěrečné poznámky

- Vhodný výběr příznaků může významně zlepšit výkon při řešení problému strojového učení (přesnost i počítačová náročnost) – ale jedná se o náročnou úlohu!
- Je to cesta k řešení problémů s velmi mnoha atributy
- Pozor na vztah mezi relevancí a optimalitou (nelze automaticky ignorovat všechny příznaky s malým hodnocením – mohou mít význam v kombinaci!).
- Prostor pro vylepšení ?
  - Nový způsob prohledávání prostoru podmnožin příznaků
  - Odhad kvality aktuální množiny příznaků
- Malé množiny příznaků lze najít i při použití metody „boosting" (kombinace klasifikátorů) pro klasifikátory s jediným příznakem!