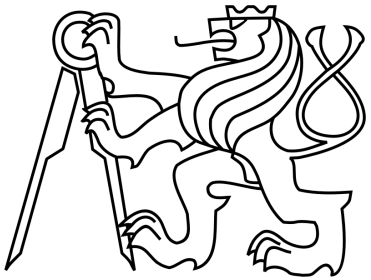
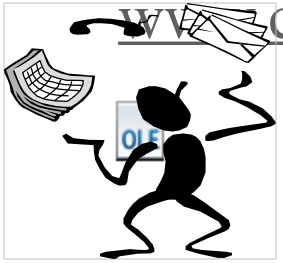
Předzpracování dat

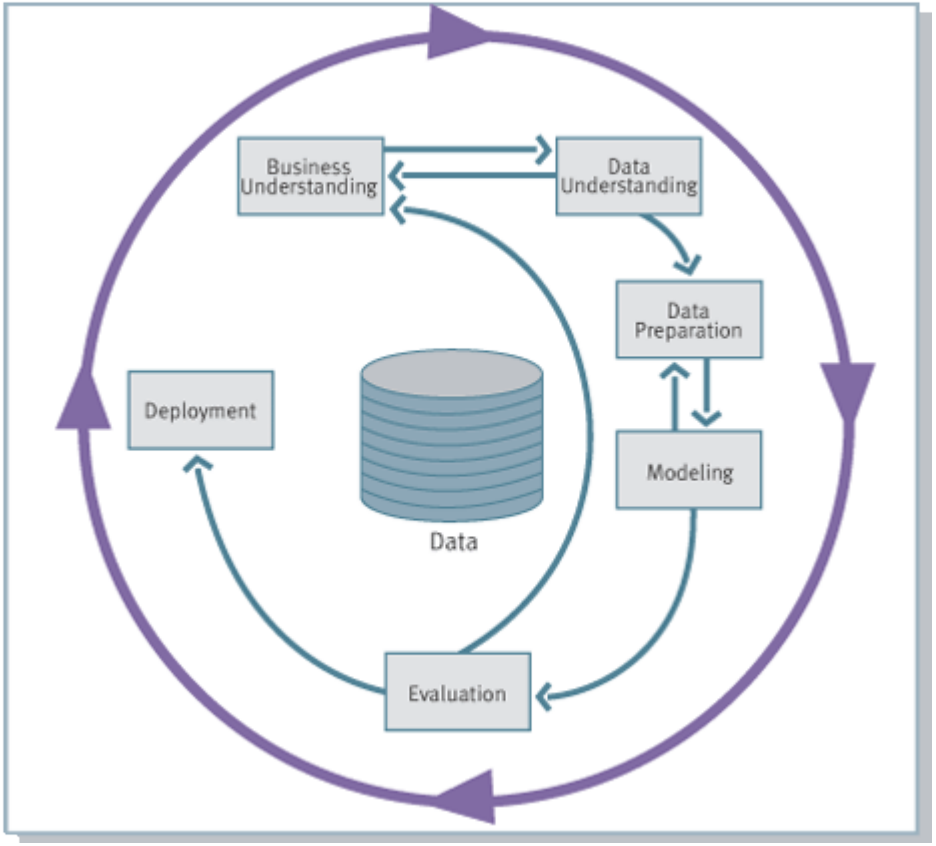
Lenka Vysloužilová



Nature Inspired
Technologies Group

Metodika CRISP-DM (www.crisp-dm.org)





Příprava dat – Data Preparation

- příprava dat pro modelování
  - selekce atributů – výběr relevantních atributů
  - čištění dat
  - získávání odvozených atributů
  - 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 atributů

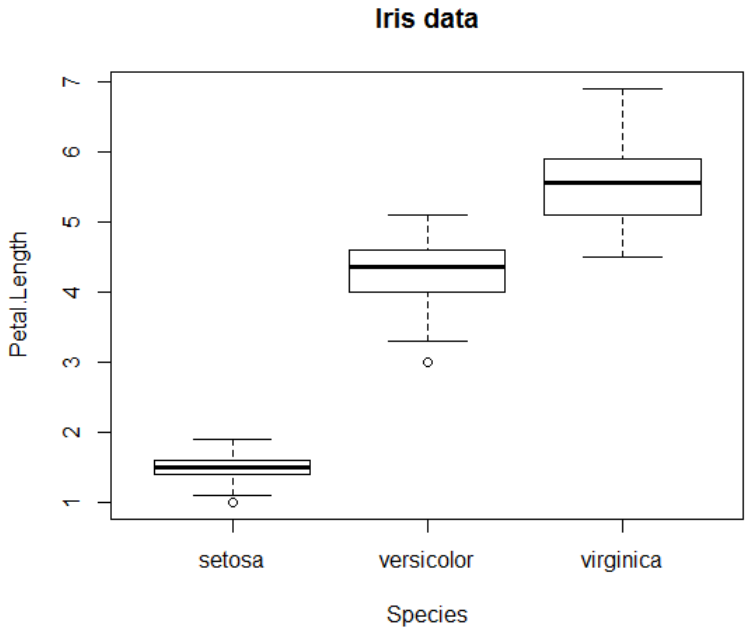
Úprava INSTANCÍ

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

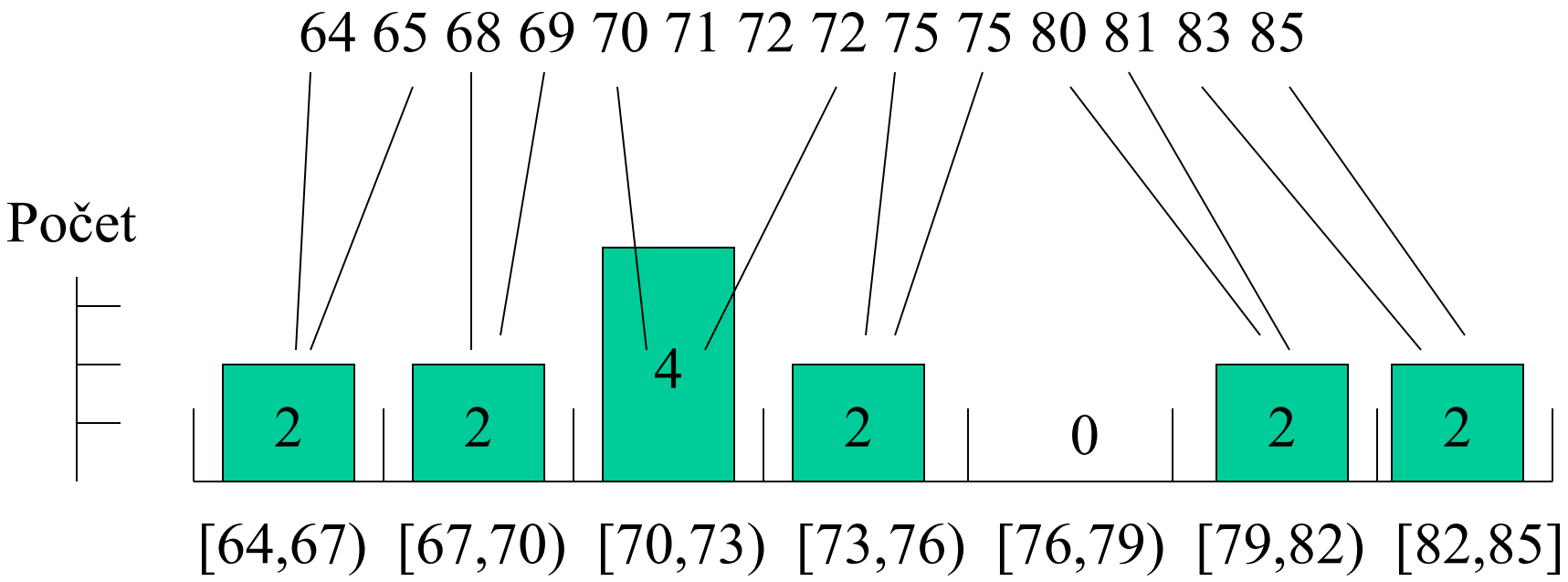
- obrovský počet instancí - pro algoritmy pracující v dávkovém režimu nutnost
  - redukce počtu dat
  - tvorba modelů na základě podmnožin a jejich následná kombinace
- rozdělení dat na trénovací a testovací část
- nevyvážená data např třída A 95%, třída B 5%
  - každý objekt patří do majoritní třídy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

Úprava atributů

Diskretizace dat

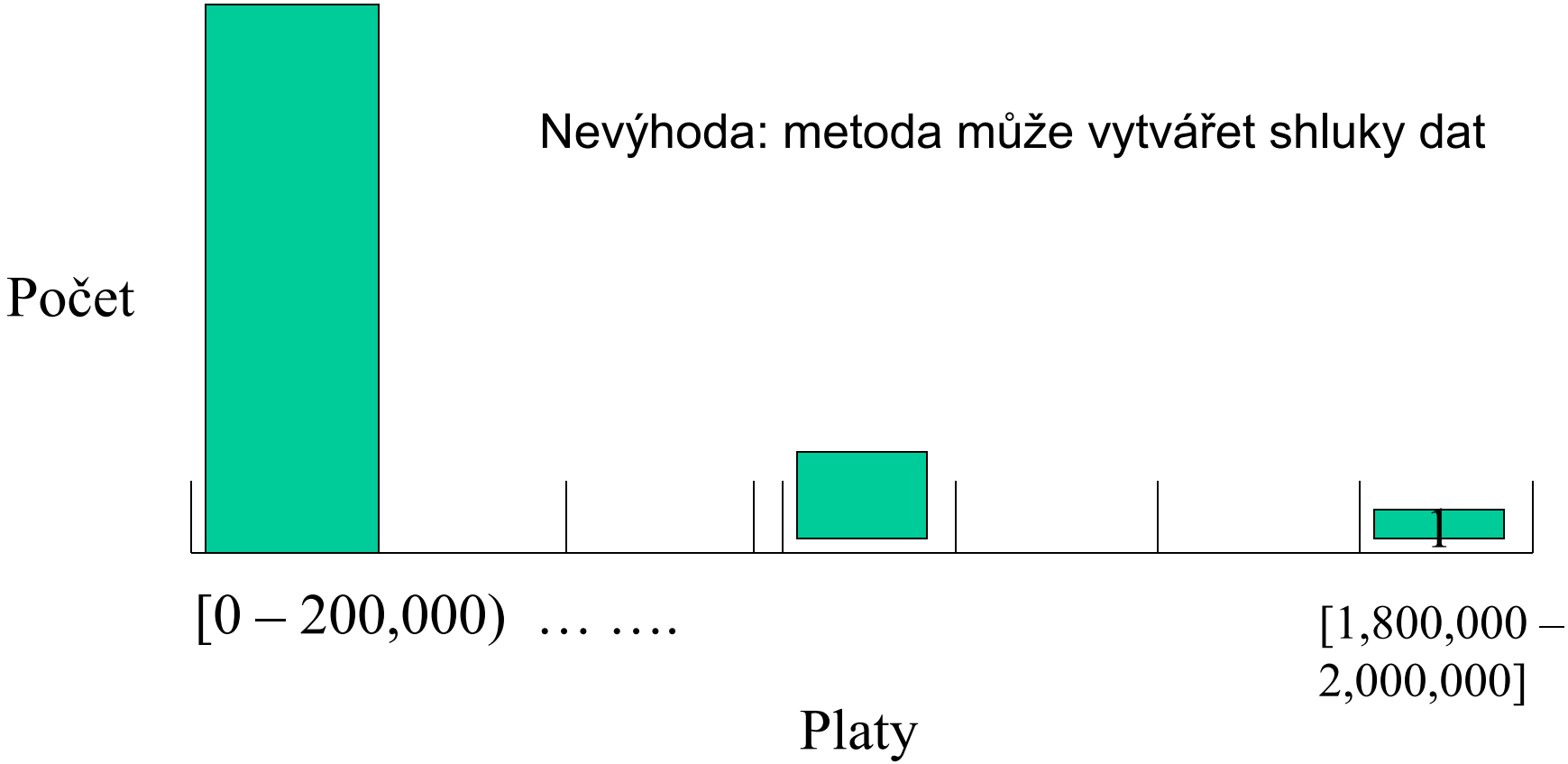
- Neinformované metody
  - ekvidistantní intervaly
  - ekvifrekvenční intervaly
- Informované metody
  - využití znalosti o příslušnosti objekt -> třída
  - strategie rozdělování nebo spojování intervalů

Diskretizace: Ekvidistantní intervaly



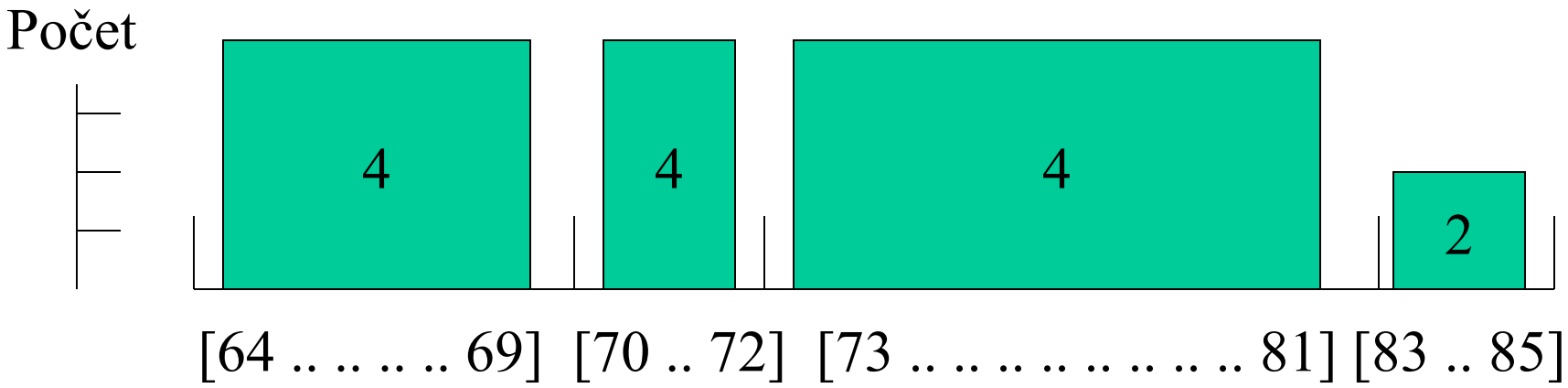
Diskretizace: Ekvidistantní intervaly

Nevýhoda: metoda může vytvářet shluky dat



Diskretizace: Ekvifrekvenční intervaly

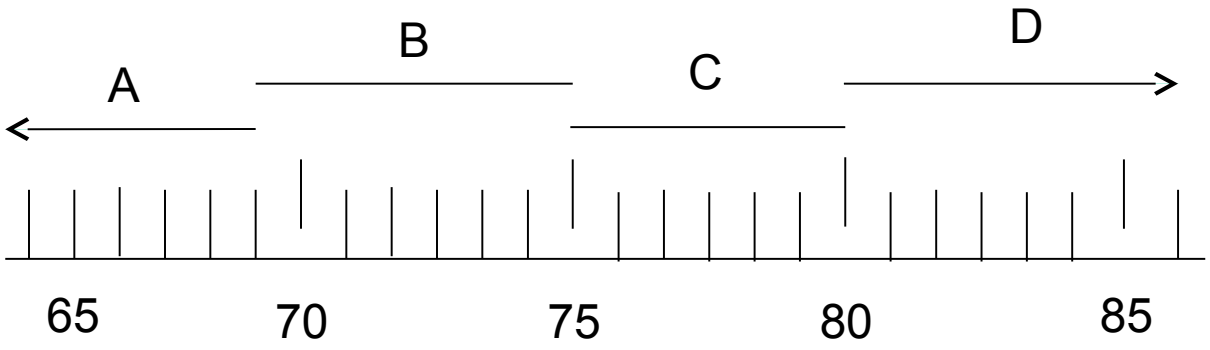
64 65 68 69 70 71 72 72 75 75 80 81 83 85



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

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

vi : aktuální hodnota atributu *I*

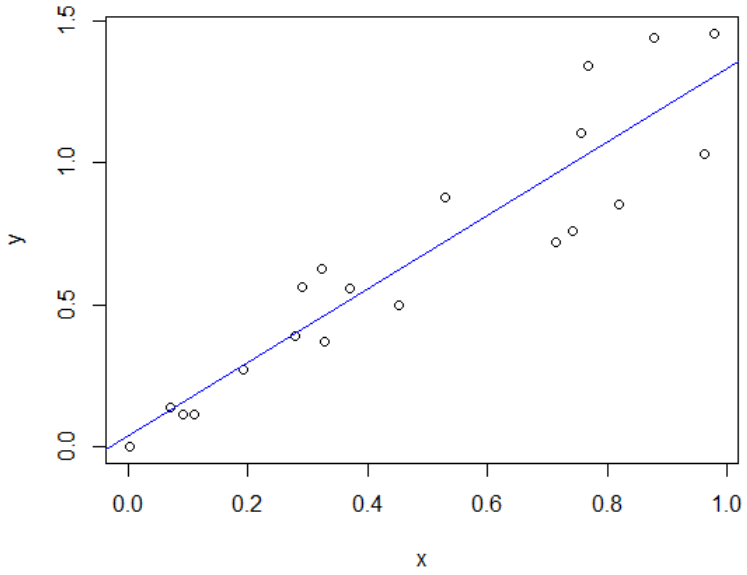
Odvozené atributy

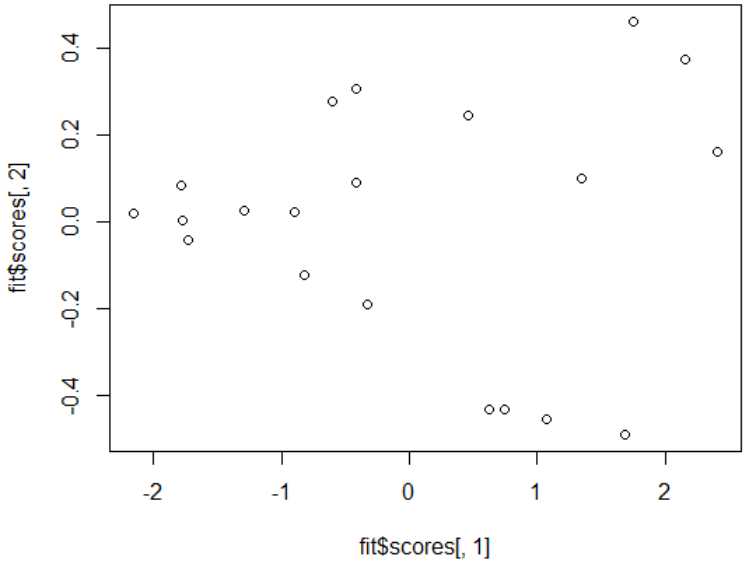
- výpočet nového atributu ze stávajících
- BMI = váha(kg) / výška(m)2
- rodné číslo => věk a pohlaví
- agregační hodnoty

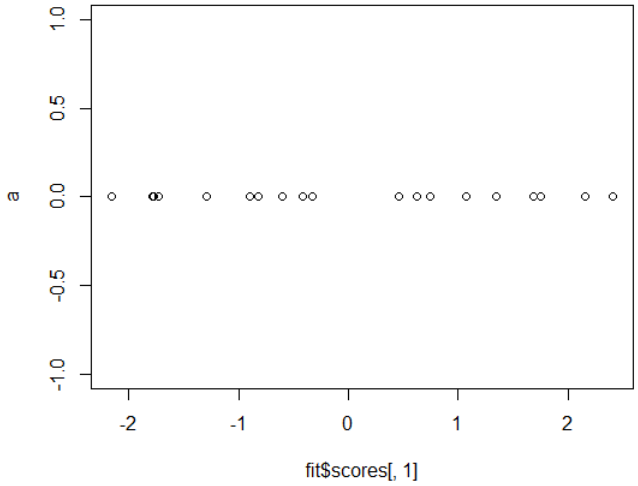
Redukce počtu atributů

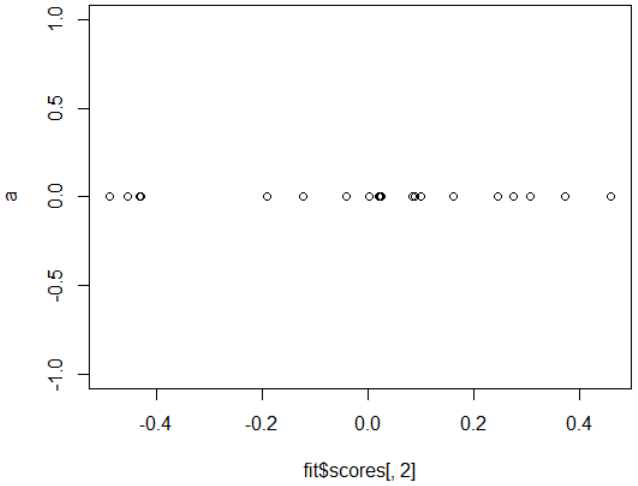
- Analýza hlavních komponent (PCA)
  - nové atributy nelze interpretovat
  - využití pro vizualizaci dat – použijeme n nejlepších komponent
  - princomp
- Selekce atributů
  - hledáme takové atributy, které nejlépe přispějí ke klasifikaci
  - metoda filtru
    - spočteme charakteristiku vyjadřující vhodnost atributu
    - korelace, chi-kvadrát, entropie, informační míra závislosti
    - vychází z kontingenční tabulky
    - nevýhoda: posuzujeme každý atribut samostatně – množiny atributů
  - metoda obálky – použití metod strojového učení

PCA









Summary

Dobrá příprava dat je klíčem k vytvoření platného a spolehlivého modelu

- Probably Approximately Correct (PAC) učení