
Data Mining
II

M9DM2 Data mining II

- S novou akreditací změna rozsahu na 2hodinový seminář a ukončení kolokviem.

- Pokračování předmětu Data mining I s důrazem na praktické použití metod.

- Prohloubení znalostí z kurzu Data mining I.

- Vybrané semináře vedené odborníky z praxe.

M9DM2 Data mining II – plán seminářů

- 16.9. a 23.9. **Diskriminační analýza** (Navrátil)
- 30.9. a 7.10. **Text mining** (Buček)
- 14.10. a 21.10. **Úvod do Pythonu** (Pokora)
- 4.11. a 11.11. **Data mining v praxi** (Kondek)
- 18.11. a 25.11. **Práce s chybějícími daty** (Zlámal)
- 2.12. a 9.12.**Grafická prezentace výsledků** (Selingerová)
- 16.12. **Prezentace projektů**

M9DM2 Data mining II – kolokvium

Podmínky pro získání kolokvia:

- Aktivní účast na seminářích
- Vypracování domácích úkolů během semestru.
- Prezentace studentského projektu na závěrečném semináři.

M9DM2 Data mining II – projekt

- Utvořte maximálně tříčlenný tým, vyberte si vhodný datový soubor a položte otázky, na které se budete snažit odpovědět.
- O této skutečnosti nás informujte e-mailem - uveďte, prosím, složení týmu, název projektu a jednu až dvě věty, co budete dělat.
- Proveďte vlastní analýzu (v libovolném softwaru).
- Připravte krátkou prezentaci (cca 15 min.).
- V prezentaci publikum seznamte s vaším problémem, jak jste jej řešili a na co jste přišli.
- Rozhodně není nutné popisovat použité metody a jiné technické záležitosti, zaměřte se hlavně na výsledky a jejich interpretaci.

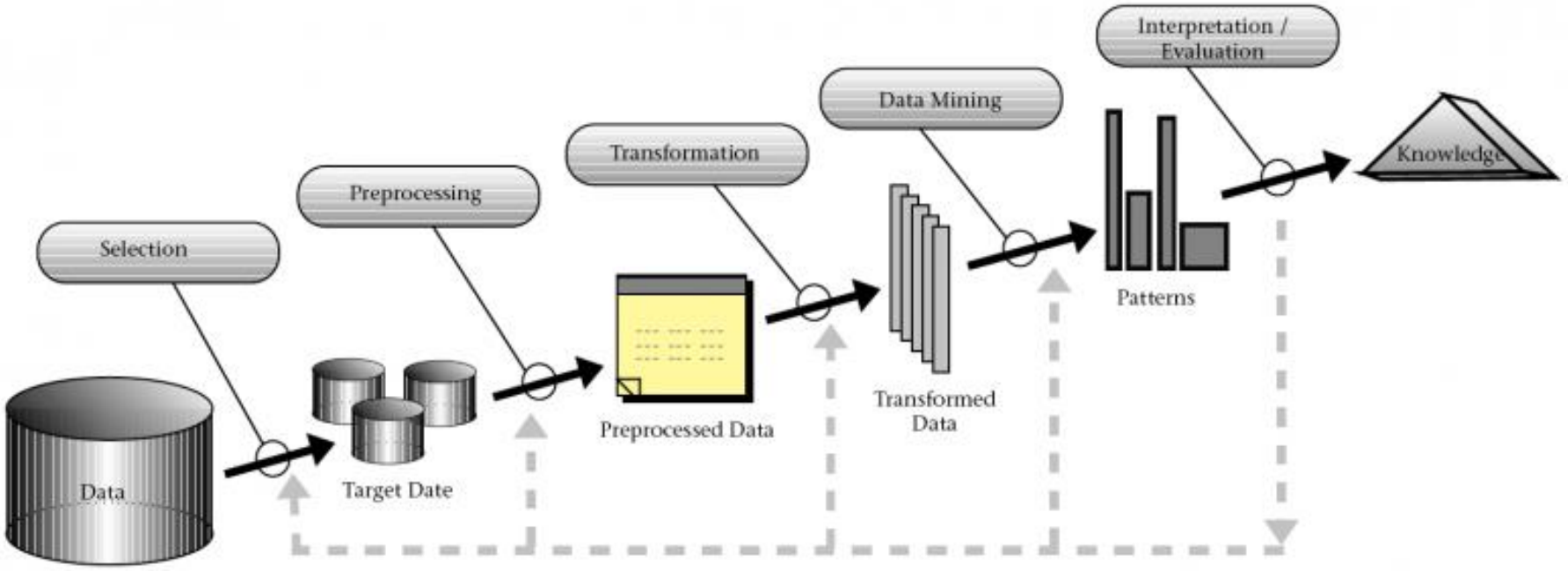
Kde jsme skončili?

Co je to data mining?

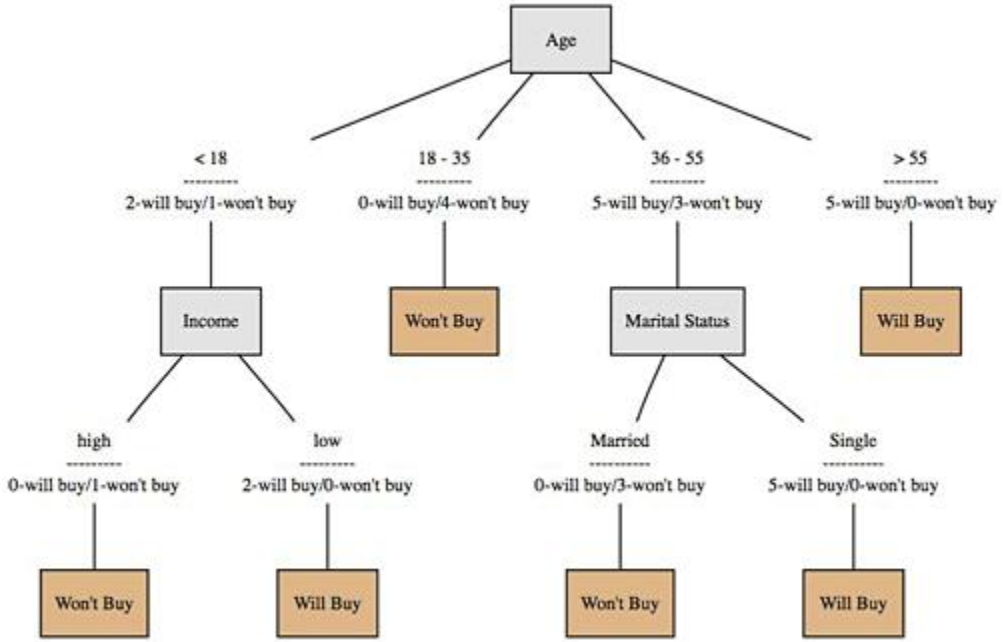
Data mining je analytická metodologie získávání netriviálních skrytých a **potenciálně užitečných informací** z dat.

[wikipedia]

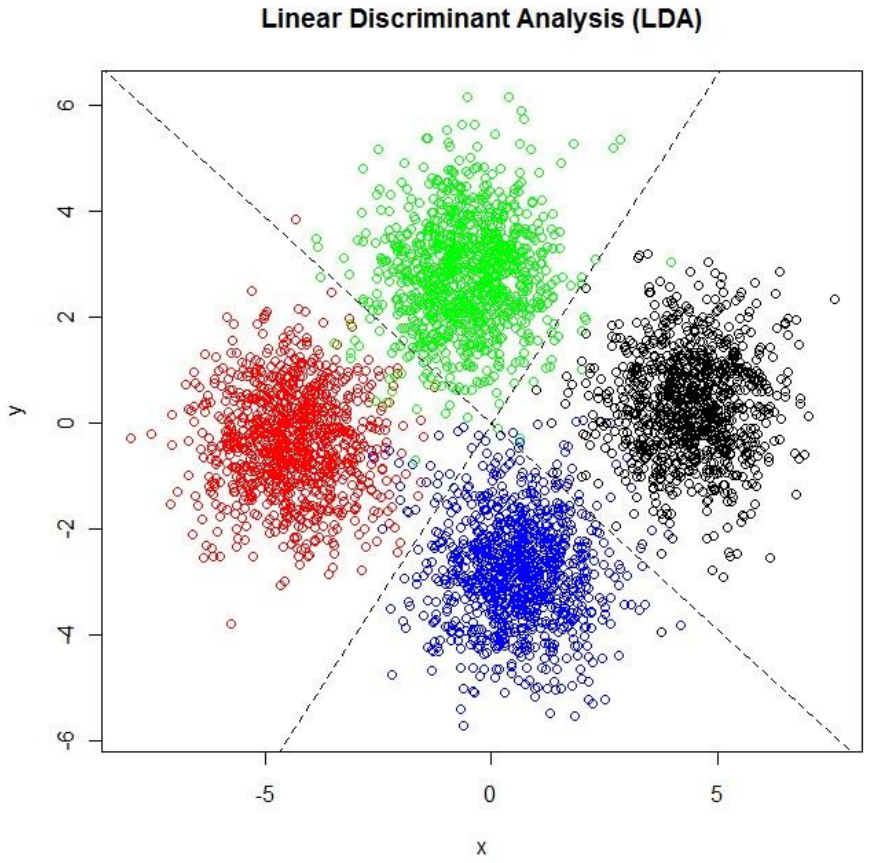
Data mining


Selection
Preprocessing
Transformation
Data Mining
Interpretation / Evaluation
Knowledge
Data
Target Date
Preprocessed Data
Transformed Data
Patterns

Rozhodovací stromy



Diskriminační analýza



Data a cíl

$(x_{i1},\ldots, x_{ip}, Y_i)$, $i=1,\ldots,n$
$\mathbf{x}_i$ je vektor **spojitých** prediktorů
Y_i udává příslušnost pozorování k dané skupině (kategoriální proměnná) – skupiny označme 1,2,…,J.

Úkol: Na základě dat zkonstruovat rozhodovací pravidlo, které bude co nejlépe klasifikovat nová pozorování $(x^*_{i1},\ldots, x^*_{ip})$ do příslušné skupiny.

Úvod a historie

- Úloha supervised learning (učení s učitelem).
- Zakladatel: R. A. Fisher (1936) – klasifikace kosatců (iris).
- Způsoby odvození klasifikačního pravidla:
  - Kanonická diskriminační analýza
  - Pravděpodobnostní modely:
    - Parametrické metody (LDA, QDA)
    - Neparametrické metody (k-nearest neighbors, metody založené na jádrových odhadech hustoty, na hloubce, apod.)

Kritéria pro hodnocení kvality modelu

- Matice záměn.
- Správnost klasifikace, chyba klasifikace.
- Správnost klasifikace do dané třídy, chyba klasifikace do dané třídy.
- Specificita a senzitivita.
- ROC – křivky.