

Program szkolenia

Data mining w praktyce z wykorzystaniem R

Prowadzący: dr inż. Artur Suchwałko

1. Wprowadzenie do data miningu

- czym jest data mining (DM)?
- przegląd zastosowań DM: zastosowania w bankowości i w przemyśle, text mining, web mining
- data mining z punktu widzenia potrzeb biznesowych firmy: analiza klientów i ich zachowania, CRM, wspomaganie podejmowania decyzji, możliwe korzyści
- główne etapy w procesie odkrywania wiedzy w danych (Knowledge Discovery in Data)
- różnice pomiędzy DM, OLAP i analizą wyników zapytań bazodanowych
- DM a statystyka

2. Przegląd rodzajów zadań DM

- klasyfikacja
- analiza skupień
- odkrywanie reguł asocjacyjnych
- redukcja wymiaru

3. Przygotowanie danych do analiz metodami data mining

- analiza jakości danych
- obserwacje brakujące i nietypowe
- czyszczenie danych
- przekształcenia wstępne

4. Przegląd metod stosowanych w DM

- klasyfikacja metod/algorytmów; uczenie z nadzorem i uczenie bez nadzoru
- metody klasyfikacji (idea wielu metod, dokładniej: drzewa klasyfikacyjne, metoda najbliższego sąsiada (k-NN))
- algorytmy analizy skupień (metoda k-średnich, PAM, metody hierarchiczne)
- techniki redukcji wymiaru (metoda składowych głównych (PCA), skalowanie wielowymiarowe (MDS))
- metody odkrywania reguł asocjacyjnych (algorytm apriori) oraz wizualizacja odkrytych reguł.

5. Wybrane metody wizualizacji i eksploracyjna analiza danych (EDA)

- analiza graficzna jednowymiarowa i wielowymiarowa
- podstawowe wykresy analizy opisowej
- zaawansowane algorytmy stosowane w wizualizacji danych wielowymiarowych
- czym jest EDA?
- metody stosowane w EDA

6. Wybór najlepszych cech i optymalnego modelu/algorytmu

- wybór istotnych cech (atrybutów)
- ogólne reguły związane z wyborem modelu w praktyce
- problem przeuczenia (overfitting)
- kompromis pomiędzy złożonością struktury modelu, a jego efektywnością
- ocena efektywności zastosowanej metody/algorytmu
- podstawy oceny jakości modeli klasyfikacyjnych
- podstawy oceny jakości analizy skupień

7. Prowadzenie projektów DM

- popularne metodologie prowadzenia projektów DM:
- Virtuous Cycle of Data Mining
- CRISP-DM
- wskazówki praktyczne

8. Case studies

- przykłady typowych zadań DM na bazie wybranych problemów praktycznych

9. Dyskusja: możliwe korzyści biznesowe związane z zastosowaniem analiz data mining

- dyskusja z uczestnikami szkolenia na temat przydatności i możliwości zastosowanie metod data miningu w analizie problemów spotykanych w ich praktyce zawodowej