

Eksploracja danych

Syllabus

AUTOR PAWEŁ GŁADKI

1. Wprowadzenie do eksploracji danych.

Dlaczego zajmować się eksploracją danych? Czym jest eksploracja danych? Jakiego rodzaju dane są przedmiotem eksploracji? Funkcjonalność eksploracji danych. Czy wszystkie wzorce są interesujące? Klasyfikacja systemów eksploracji danych. Główne wyzwania w eksploracji danych.

2. Przegląd technik eksploracji danych.

Dane i rodzaje danych. Zbiory rozmyte. Pozyskiwanie informacji. Uczenie maszynowe. Techniki statystyczne i estymacyjne. Miary podobieństwa. Drzewa decyzyjne.

Projekt: Wprowadzenie do SciLab.

3. Klasyfikacja, cz. 1.

Problem klasyfikacji. Klasyfikacja oparta na regresji. Klasyfikacja oparta na odległości (algorytm k-NN).

Projekt: Klasyfikacja przy użyciu k-NN.

4. Klasyfikacja, cz. 2.

Drzewa decyzyjne. Klasyfikacja przy użyciu reguł. Jakość klasyfikatorów.

5. Grupowanie, cz. 1: techniki częściowe.

Czym jest grupowanie? Algorytmy partycjonujące (k-means).

Projekt: Wprowadzenie do algorytmów k-means i DBSCAN.

6. Grupowanie, cz. 2: techniki hierarchiczne.

Algorytmy hierarchiczne. Algorytmy bazujące na gęstości (DBSCAN). Techniki grupowania dużych baz danych: BIRCH oraz CURE.

Projekt: Grupowanie przy użyciu k-means vs. DBSCAN.

7. Reguły asocjacyjne, cz. 1: zestawy częstych przedmiotów.

Problem koszyka zakupów i jego zastosowania. Zestawy częstych przedmiotów. Reguły asocjacyjne. Znajdowanie reguł asocjacyjnych. Generowanie zestawów częstych przedmiotów i jego strategię. Reguła *apriori* i algorytm *apriori*.

Projekt: Eksploracja przy użyciu reguł asocjacyjnych.

8. Reguły asocjacyjne, cz. 2: szybkie algorytmy i generowanie reguł.

Bezpośrednie haszowanie i przycinanie. Generowanie reguł. Maksymalne i domknięte zestawy częstych przedmiotów.

9. Reguły asocjacyjne, cz. 3: ocena wzorców asocjacyjnych.

Ocena wzorców. Miary probabilistyczne. Miary zainteresowania.

10. Eksploracja sekwencyjnych wzorów.

Wielopoziomowe reguły asocjacyjne. Formalna definicja sekwencji. Generowanie kandydatów.

11. Eksploracja treści internetowych.

Eksploracja zawartości sieci. Ekstrakcja wzorców i relacji. Eksploracja użycia sieci. Dynamiczne zliczanie zestawów.

Projekt: Wprowadzenie do algorytmu HITS.

12. Wyszukiwarki.

Sieć i wyszukiwarki. Heurystyczny ranking. Ranking według Page Rank (Google). HITS.

Projekt: Wyszukiwanie treści internetowych.

13. Eksploracja danych i prywatność, cz. 1.

14. Eksploracja danych i prywatność, cz. 2.

Bibliografia

- [1] Margaret Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2002.
- [2] Michael Steinbach Pang-Ning Tan i Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

Zasady oceniania

Aktywność na zajęciach, praca nad projektami: 90%

Obecność na zajęciach: 10% (w semestrze można mieć jedną nieobecność nieusprawiedliwioną)