

Piotr Pokarowski
pokar@mimuw.edu.pl

Statystyczna analiza danych

30 godz. wykładów + 30 godz. laboratorium.

Wykład będzie poświęcony eksploracyjnej analizie danych i predykcji statystycznej. W pierwszej części omówione będą współczesne metody statystycznej analizy danych wielowymiarowych: streszczenia danych (miary położenia, rozrzutu i zależności między cechami), estymacja gęstości, redukcja wymiaru i klasteryzacja. Drugim tematem będzie predykcja statystyczna, czyli prognozowanie wartości cechy nieobserwowanej na podstawie cech obserwowanych. Przedstawione będą zarówno modele predykcji cechy ciągłej, nazywane tradycyjnie regresją jak również cechy dyskretnej - klasyfikacja lub alokacja. Zajęcia w laboratorium będą poświęcone wykorzystaniu pakietu "R" do analizy i wizualizacji danych oraz implementacji i porównywania algorytmów predykcji.

PROGRAM

Wstęp

Co to jest "statystyczna" analiza danych? Czy dane są próbą z populacji? Jaki jest problem obliczeniowy dla populacji? Czy dana metoda analizy danych rozwiązuje ten problem?
Główne zadania: eksploracja danych i predykcja.

Eksploracyjna analiza danych

1. Streszczenia danych.

Miary położenia: średnia, mediana, moda.

Miary rozrzutu: wariancja, odchylenie standardowe i bezwzględne, odległość międzykwartylowa, MAD, entropia, współczynnik Giniego.

Boxplot.

Miary bliskości między cechami: błąd średniokwadratowy, entropia względna = odległość Kullbacka-Leiblera, odległość chi-kwadrat.

Miary zależności między cechami: korelacja liniowa, korelacja rang, wspólna informacja, współczynnik Goodmana-Kruskala, krzywe ROC.

2. Estymacja gęstości: histogram i estymator jądrowy.

3. Redukcja wymiaru cech: analiza składowych głównych, skalowanie wielowymiarowe i analiza odpowiedniości.

4. Klasteryzacja - redukcja wymiaru danych.

Klasteryzacja oparta na modelu statystycznym czyli estymacja parametrów mieszanki rozkładów normalnych.

Metody relokacyjne: k-średnich, k-medoidów.

Metody hierarchiczne: aglomeracyjne (single-, average-, complete-linkage) i metody podziału.

5. Obserwacje odstające i brakujące.

Predykcja statystyczna

1. Wstęp do predykcji.

Regresja klasyfikacja i dyskryminacja na przykładzie metody k-najbliższych sąsiadów (knn).

$E(Y|X=x)$ - optymalna średniokwadratowa regresja; $\operatorname{argmax}_y p(y|x)$ - optymalna klasyfikacja.

Empiryczna ocena błędu predykcji: próba ucząca i testująca.

Ocena błędu predykcji za pomocą randomizacji danych: krosvalidacja, testy permutacyjne i metoda bootstrap.

2. Wielowymiarowy rozkład normalny.

Estymacja parametrów metodą największej wiarygodności.

Rozkłady pomocnicze: chi-kwadrat, t-studenta, F-Snedecora.

3. Metody parametryczne.

Modele liniowe: regresja, analiza kowariancji i analiza wariancji.

Klasyfikacja w modelu normalnym.

Liniowa analiza dyskryminacyjna.

Regresja logistyczna i logliniowa.

Sieci neuronowe.

Ocena istotności i wybór modelu, przedziały ufności dla współczynników.

Ocena i wybór modelu.

4. Metody nieparametryczne.

Regresja nieparametryczna.

Metoda knn.

Drzewa klasyfikacyjne i regresyjne.

Maszyny wektorów podpierających.

Literatura

1. W. N. Venables i B. D. Ripley, *Modern Applied Statistics with S*, Springer 2002.
2. T. J. Hastie, R. J. Tibshirani i J. Friedman, *The Elements of Statistical Learning*, Springer 2001.
3. W. Vos i L. Evers, *MSc in bioinformatics: statistical data mining*, 2004, dostępny w sieci.
4. J. Koronacki i J. Mielniczuk, *Statystyka*, WNT 2001.
5. J. Koronacki i J. Ćwik, *Statystyczne systemy uczące się*, WNT 2005.
6. J. Faraway, *Linear Models with R*, Chapman and Hall/CRS 2004. Wcześniejsza wersja pt. *Practical Regression and ANOVA using R* jest dostępna na www.r-project.org.
7. J. Faraway, *Extending the Linear Models with R*, Chapman and Hall/CRS 2005.
8. The R Development Core Team, *An Introduction to R*, www.r-project.org.
9. E. Paradis, *R for Beginners*, www.r-project.org.

Wymagania: zaliczony wykład „Statystyka I”.

Forma zaliczenia: egzamin komputerowy z programowania w R + egzamin pisemny.