

Lista zagadnień na egzamin magisterski na kierunku Analiza danych

Przedmioty obowiązkowe

1. Eksploracja danych i uczenie maszynowe
 - a. Algorytmy predykcji zmiennych jakościowych (klasyfikacja) – k najbliższych sąsiadów, drzewa klasyfikacyjne, lasy losowe, sieci neuronowe, maszyny wektorów podpierających.
 - b. Algorytmy predykcji zmiennych ilościowych (szacowanie) – k najbliższych sąsiadów, drzewa klasyfikacyjno-regresyjne, lasy losowe, sieci neuronowe.
 - c. M-estymatory z karą w modelach regresyjnych – regresja grzbietowa, regresja z karą „Lasso”.
 - d. Algorytmy grupowania – k średnich, hierarchiczna analiza skupień, dwustopniowa analiza skupień.
 - e. Pola losowe Markowa – omówienie, przykłady zastosowań, estymacja parametrów.
 - f. Reguły asocjacyjne – algorytm *A priori* i ocena jakości reguł.
 - g. Ocena jakości modeli – zbiory uczący i testowy, walidacja krzyżowa, miary jakości klasyfikacji, krzywa ROC, miary jakości szacowania.
2. Programowanie i algorytmika
 - a. Iteracja a rekurencja.
 - b. Podejście zachłanne a programowanie dynamiczne.
 - c. Algorytmy na tekstach.
 - d. Algorytmy geometryczne.
 - e. Złożoność obliczeniowa algorytmów.
 - f. Syntaktyka i semantyka języka programowania.
 - g. Struktury danych: listy, krotki, słowniki, zbiory.
 - h. Programowanie strukturalne a programowanie obiektowe.
3. Bazy danych
 - a. System zarządzania bazami danych, funkcje systemu, przykłady SZDB.
 - b. Modele relacyjnych baz danych, relacje, związki, klucze, wartość nieoznaczona.
 - c. Język SQL – podstawowe konstrukcje, selekcja, złączenia, agregacje.
 - d. Optymalizacja zapytań – indeksy, wyszukiwanie pełnotekstowe.
 - e. Perspektywy i wyrażenia tablicowe.
 - f. Big Data, model 4W, analiza i zastosowania.
 - g. Modele nierelacyjnych baz danych.
4. Biostatystyka
 - a. Podstawowe modele regresji – regresja liniowa, regresja logistyczna.
 - b. Analiza wariancji i porównania wielokrotne.
 - c. Analiza przeżycia – estymator Kaplana-Meiera funkcji przeżycia, estymator Coxa funkcji hazardu.
5. Text mining
 - a. Metody strukturyzacji danych tekstowych – tworzenie słownika, standardowy model boolowski, odwrócony indeks, ogólny model wektorowy, macierz tf-idf.

- b. Metody klasyfikacji tekstu – klasyfikator Rocchio, naiwny klasyfikator Bayesa, maszyny wektorów podpierających.
 - c. Analiza ukrytych własności semantycznych (LSA), modele tematyczne (pLSA, LDA).
 - d. Wektory słów (Word2vec, GloVe, FastText).
 - e. Analiza sentymentu – podstawowe pojęcia, zastosowania, analiza na poziomie dokumentu, zdania i aspektu, metody nadzorowane i nienadzorowane.
- 6. Wizualizacja danych
 - a. Estetyczne metody odwzorowywania danych. Kolor i jego znaczenie.
 - b. Wizualizacje wielkości, rozkładów, proporcji, powiązań, szeregów czasowych i niepewności. Przykłady konkretnych typów wykresów.
- 7. Aspekty informatyczne analizy danych
 - a. Przetwarzanie danych z poziomu linii poleceń systemu uniksopodobnego.
 - b. Wyrażenia regularne i ich zastosowania.
 - c. Pobieranie danych i wydobywanie informacji z serwisów internetowych i stron WWW; web crawling i web scraping.
 - d. Obliczenia równoległe w zbieraniu i analizie danych.
 - e. Zarządzanie wersjami kodu i wyników obliczeń w analizie danych i w uczeniu maszynowym. (od roku akademickiego 2023/24)
- 8. Eksploracja masywnych zbiorów danych
 - a. Algorytmy eksploracji strumieni danych (próbkiwanie, filtrowanie, zapytania, estymacja momentów częstotliwości, zbiory częste).
 - b. Algorytmy eksploracji dużych grafów (PageRank, inne warianty *Rank, SimRank, znajdowanie społeczności, klasteryzacja).
 - c. Systemy rekomendacyjne (filtrowanie społecznościowe, dekompozycja, itp).
 - d. Dane o wysokim wymiarze: znajdowanie elementów podobnych (index Jaccarda, MinHash, LSH), klasteryzacja, redukcja wymiarowości.
- 9. Statystyka bayesowska
 - a. Rozkład a priori i rozkład a posteriori. Zastosowanie twierdzenia Bayesa do obliczania rozkładu a posteriori.
 - b. Klasyfikator bayesowski, jego postać i własności optymalności.
 - c. Funkcje straty, ryzyko bayesowskie i konstrukcja reguł decyzyjnych minimalizujących ryzyko bayesowskie.
 - d. Przykłady typowych statystycznych modeli bayesowskich ze sprzężonymi rozkładami a priori.
- 10. Wprowadzenie do deep learning
 - a. Regularyzacja sieci neuronowej oraz typowe metody regularyzacji.
 - b. Sieci rekurencyjne: uczenie, zastosowania, typowe architektury, problemy pojawiające się podczas uczenia.
 - c. Budowa autoenkoderów oraz ich zastosowania.
 - d. Sieci splotowe: zasada działania, najważniejsze architektury.
- 11. Analiza szeregów czasowych
 - a. Estymacja i eliminacja trendu i współzrędnej sezonowej szeregów czasowych.
 - b. Podstawowe modele stacjonarnych i niestacjonarnych szeregów czasowych.
 - c. Prognozowanie stacjonarnych szeregów czasowych.

Przedmioty do wyboru

12. Analiza sygnałów i obrazów

- a. Transformata Fouriera i jej zastosowania.
- b. Przykłady filtrów jedno i dwuwymiarowych.
- c. Twierdzenie o próbkowaniu. Zjawisko zlewania widma.
- d. Analiza widma z wykorzystaniem krótkoczasowej transformaty Fouriera.
- e. Metody wykrywania kształtów. Transformata Hougha.
- f. Wykorzystanie histogramu do analizy obrazów cyfrowych. Wyrównanie histogramu, algorytm Otsu.

13. Analiza modeli ekonomicznych

- a. Pojęcie modelu matematycznego zjawiska ekonomicznego - modele statyczne i dynamiczne, przykłady modeli.
- b. Przykłady modeli ekonomicznych prowadzących do równań różniczkowych.
- c. Pojęcie równania różnicowego i metody rozwiązywania takich równań. Przykłady modeli ekonomicznych prowadzących do równań różnicowych.
- d. Model Leontiefa równowagi rynkowej - założenia i opis modelu ciągłego i dyskretnego.

14. Podstawy marketingu internetowego

- a. Konwersja, kanały pozyskania, atrybucja i modele atrybucji stosowane w marketingu internetowym.
- b. Struktura witryny internetowej, tagi meta, słowa kluczowe.
- c. Testy A/B witryny internetowej.
- d. Narzędzia wspomagające pracę i wdrożenia w marketingu internetowym.

15. Statystyka w medycynie

- a. Rodzaje badań statystycznych w medycynie (badania prospektywne i retrospektywne).
- b. Symulacje stochastyczne w statystyce. Testy permutacyjne.
- c. Podstawowe cele meta-analizy i odpowiednie metody statystyczne.

16. Wstęp do sieci neuronowych

- a. Perceptron prosty: architektura i jeden z algorytmów uczenia.
- b. Przeuczenie sieci neuronowej. Objawy i metody przeciwdziałania.
- c. Algorytm spadku gradientowego i sposób wykorzystania go do uczenia wielowarstwowych sieci neuronowych (MLP).