Program kursu Data Science PRO

PODSTAWY JĘZYKA PYTHON I ŚRODOWISKO PRACY

- podstawy obsługi Bash'a,
- system kontroli wersji Git,
- środowiska programistyczne PyCharm oraz Jupyter,
- wprowadzenie do języka Python: składnia języka, podstawowe typy zmiennych, wyrażenia warunkowe, pętle, funkcje,
- Python cd.: obsługa wyjątków, wyrażenie lambda, wyrażenia regularne, zapis i odczyt danych z pliku tekstowego,
- podstawy programowania obiektowego (wprowadzenie paradygmatu, implementacja prostych klas, dziedziczenie).

PRACA Z DANYMI W JĘZYKU PYTHON

- moduł numpy: wektory, macierze, tablice wielowymiarowe, operacje zwektoryzowane, agregacje,
- moduł pandas: praca z tabelami danych (tworzenie i obsługa tabel, transformacje danych, agregowanie danych),
- moduł matplotlib: wizualizacja danych,
- podstawy języka SQL - wyciąganie i agregacja danych z baz, łączenie z SQL-ową bazą danych z poziomu Pythona,
- pozyskiwanie i czyszczenie danych z różnych typów plików (csv, json, xml, xlsx)
- pobieranie danych z Internetu (web scraping oraz wykorzystanie publicznych API).

STATYSTYKA I MODELOWANIE STATYSTYCZNE DANYCH

- statystyka opisowa: miary położenia i rozrzutu, korelacja,
- podstawowe zagadnienia rachunku prawdopodobieństwa (zmienne losowe, rozkłady prawdopodobieństwa, twierdzenie bayesa),
- zagadnienia estymacji i testowania hipotez odnośnie parametrów i rozkładów, przedziały ufności,
- model regresji liniowej (problem regresji, definicja i własności modelu, metody ewaluacji modelu, regularyzacja),
- modelowanie szeregów czasowych - model ARIMA.

UCZENIE MASZYNOWE CZ. 1

- zagadnienie klasyfikacji: podstawowe metody klasyfikacji (regresja logistyczna, drzewo decyzyjne), metody oceny jakości klasyfikatorów, optymalizacja parametrów,
- problem przeuczenia, regularyzacja,
- potoki predykcyjne (pipelines) w sklearn - automatyzacja procesów uczenia maszynowego
- metody oceny istotności i selekcji cech,
- inne klasyfikatory (naiwny klasyfikator Bayesa, KNN, SVM),
- metody łączenia klasyfikatorów, komitety klasyfikatorów (bagging, lasy losowe, XGBoost),
- problem niezrównoważonych klas, problem braków w danych,
- przygotowywanie danych do modelowania.

UCZENIE MASZYNOWE CZ. 2

- metody redukcji wymiaru: dekompozycja SVD, analiza składowych głównych - algorytm PCA,
- metody analizy danych tekstowych: przygotowanie tekstu do analizy (moduł NLTK), reprezentacja danych tekstowych, ekstrakcja słów kluczowych (transformacja TfIdf), analiza tematyczna tekstów (model LDA),
- grupowanie danych (analiza skupień): algorytm K-średnich (m.in.: ocena jakości wyników, wyznaczanie liczby skupień), grupowanie hierarchiczne, algorytm DBSCAN,
- grupowanie bardzo dużych danych - algorytm MiniBatch K-means.

DEEP LEARNING, SPARK I BIG DATA

- wprowadzenie do sieci neuronowych (model neuronu, funkcja aktywacji, algorytm uczenia),
- sieci jednowarstwowe i wielowarstwowe: dobór struktury sieci, problem przeuczenia, regularyzacja klasyczna i typu “dropout”,
- klasyfikacja obrazów przy użyciu sieci konwolucyjnych sieci neuronowych,
- klasyfikacja tekstów przy użyciu rekurencyjnych sieci neuronowych,
- wektorowa reprezentacja słów - model word2vec,
- wprowadzenie do technologii Big Data - biblioteka do pracy z bardzo dużymi danymi Apache Spark.