Základy dátovej analýzy v Pythone

Martin Vozár

Cieľ kurzu

Očakávaným výsledkom tohoto kurzu je učastníka prakticky previesť a doviesť k vlastnej realizácií pri základnej práci spracovania a vizualizácie dát.

Kurz bude pozostávať zo série praktických cvičení spojených s výkladom teórie podľa rozsahu kurzu.

Počas cvičenia účastník produkuje vlastný kód v jazyku Python, prípadne v prostredí Jupyter Notebook. Načítanie a spracovanie bude realizované za pomoci niektorých zaužívaných knižníc - numpy, pandas, scipy, matplotlib, v prípade pokročilejších metód scikit-learn, tensorflow.

Medzi aplikované metódy budú v základnej variante kurzu patriť lineárna a parametrická regresia. Vo variante pre pokročilejších účastníkov bude navyše aplikácie klasifikačného algoritmu rozhodovací strom. Možnosťou pri vyššej vstupnej znalosti účastníkov je implemetovať neurónové siete, či iné pokročilejšie metódy.

Očakávané vstupné znalosti

V základnej variante kurzu účastník nepotrebuje žiadne predchádzajúce znalosti, či skúsenosti. V tejto variante je úvodná časť kurzu obohatená o prechod základmi jazyka Python vrátane inštalácie prostredia postačujúceho (nie len) pre absolvovanie kurzu.

Vo všeobecnosti je výhodou, vo variante pre pokročilých prerekvizitou znalosť základou jazyka Python, práca s knižnicami, schopnosť čítania dokumentácie knižníc pri samostanom adresovaní problémov.

Disclaimer

Tento dokument je pracovnou verziou návrhu kurzu. Jeho finalizácia je predmetom ďalšej činnosti.

Obsah

Úvod

1 Inštalácia a zoznámenie sa s prostredím

Cieľom tejto časti je nainštalovať u účastníkov spoľahlivo funkčné prostredie pre ďalšiu prácu. Súčasťou tohoto procesu je nainštalovanie príslušného prostredia, pričom účastníci su prevedení jednou z viacerých variánt tejto inštalácie. V procese tejto inštalácie môžu byť oboznámení s alternatívami k jednotlivým prvkom.

Účastníci kurzu sú prevedený inštaláciou interpretera Python. Následne je predstavený package manager pip, spomenutý package manager Anaconda. Ďalej práca s package managerom pip na inštaláciu knižníc. Rýchly priebeh inštaláciou VSCode, Jupyter Lab, ako vhodných prostredí.

Začína sa úvodným programom "Hello World!". Nasleduje import knižnice numpy a výpis základných typov poľa ako np.array, np.arrange, np.linspace. Pokračuje zavádzanie natívnej funkcie v jazyku Python a práca s knižnicou numpy v tejto súvislosti.

Nasleduje import knižnice pandas a základná práca s ňou. Primárne pôjde o zoznámeni sa s obejaktami pd.DataFrame, pd.Series. Prevedenie prechodu z pd.Series do np.array.

Import knižnice matplotlib, resp. matplotlib.pyplot a vizualizácia individuálnych sérií, vizualizácia viacerých sérií. Zoznámenie sa s niektorými možnosťami grafickej knižnice v zámere produkovať čitateľné a prehľadné grafy.

2 Práca s datasetom Iris

2.1 Načítanie dát zo súboru

Prevedenie načítania dát zo súboru do dátových typov známych z predchádzajúceho celku. Následuje zobrazenie dát v dvojrozmernom priestore s kategóriou znázornenou farbou datapointu na grafe. Vysvetlenie rozdielu medzi spojitou a kategorickou veličinou. Nahrádzanie kategorickej veličiny celočíselným indexom.

2.2 Regresia

Import knižnice scipy, resp. funkcie scipy.optimize.curve_fit na realizáciu optimalizácie. Na závislosti jednotlivých veličín je aplikovaná lineárna regresia. Vysvetlenie princípu metódy najmenších štvorcov. Aplikácia parametrickej regresie. Vysvetlenie princípu optimalizácie funkcie viacerých voľných paramtrov.