Zápočtový program pro předmět Programování 2: Lineární regrese po částech

Jolana Štraitová

Srpen 2024

Zadání práce

Program provádí v jazyce Python lineární regresi po částech (*anglicky piece-wise linear regression*) na zadaných datech přečtených ze souboru a proložení těchto dat lineárními funkcemi vykresluje v grafu. Následně porovnává tuto vlastní implementaci lineární regrese po částech s již existují funkcí pro lineární regresi po částech dostupnou v knihovně *numpy*. Obě implementace vykreslí proti sobě v jednom grafu a na závěr pak pro porovnání provádí pomocí několika metrik statistickou validaci obou implementací.

Uživatelský návod pro práci s programem

- **Příprava dat** – program načítá data ze souboru ve formátu *csv*, popřípadě *xls*, se dvěma sloupci s hodnotami x a y
- **Vizuální expertýza dat** – po načtení dat se zobrazí v grafu pro vizualizaci a určení breakpointů: bodů, kde se hodnoty dat lámou a kde na sebe budou navazovat lineární funkce, které je prokládají
- **Určení breakpointů** – zadání libovolného množství x-ových hodnot bodů zlomu do příkazové řádky, resp. konzole: čísla psát vzestupně za sebe a oddělovat mezerou. Breakpointy lze zapisovat až po zavření okna s grafem.
- **Vizualizace výsledků** – výsledné proložení dat se vykreslí v grafu společně s výsledkem proložení dat provedeným knihovnou *numpy*. Program pokračuje po zavření všech oken s grafy.
- **Porozumění výsledkům statistické validace** – po provedení regrese jsou výsledky validovány pomocí několika statistických metrik:
  - **Root Mean Squared Error (RMSE)**: jeho hodnota ukazuje průměrnou velikost chyby mezi skutečnými a modelovanými hodnotami podle vzorce $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, kde n je počet dat (datových hodnot), y_i jsou zadaná data a $\hat{y}_i$ jsou modelované hodnoty. Tedy nižší hodnota RMSE indikuje lepší model.
  - **Mean Absolute Error (MAE)**: ukazuje průměrnou absolutní chybu mezi skutečnými a modelovanými hodnotami podle vzorce $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, kde n je počet dat (datových hodnot), y_i jsou zadaná data a $\hat{y}_i$ jsou modelované hodnoty. Opět je lepší nižší hodnota MAE.
  - **R–squared**: měří, jak dobře data odpovídají regresnímu modelu. R–squared nabývá hodnot mezi 0 a 1, přičemž vyšší hodnota znamená, že model lépe vysvětluje variabilitu v datech, tedy jde o lepší model.
  - **Akaike Information Criterion (AIC)**: vhodný parametr pro porovnávání různých modelů, hodnotí kvalitu modelu i počet parametrů. Používá vzorec $AIC = 2k + n\ln(\frac{RSS}{n})$, kde k je počet parametrů modelu, n je počet dat (datových hodnot) a RSS je reziduální suma čtverců (Residual Sum of Squares), tzn. součet druhých mocnin rozdílů mezi hodnotami zadaných dat a hodnotami, které vypočítal model. Nižší hodnota AIC je lepší, značí lepší model s ohledem na jeho přesnost a jednoduchost z hlediska počtu parametrů.

  - **Bayesian Information Criterion (BIC)**: funguje velmi podobně jako AIC, složitost, reso. počet parametrů modelu však penalizuje ještě přísněji, konkrétně podle vzorce $BIC = k\ln(n) + n\ln(\frac{RSS}{n})$. Opět nižší hodnota BIC znamená lepší model.

- **Interpretace výsledků** – výsledky statistické validace pro oba modely se zobrazí v terminálu, resp. konzoli, a lze za pomoci vysvětlení výše porovnat, která z metod lineární regrese po částech poskytuje přesnější výsledky.

Technický popis programu

Použité knihovny

Pro načtení dat ze souboru a manipulaci s daty v tabulce používá program knihovnu ***pandas***. Dále pro základní lineárně-algebraické operace používá knihovnu ***numpy***. Vizualizace dat a vykreslování grafů je prováděna pomocí knihovny ***matplotlib***. Data jsou proložena lineární regresí po částech pomocí knihovny ***scipy*** funkcí ***curve_fit***. Porovnání je provedeno s modelem vytvořeným knihovnou ***numpy***, konkrétně ***np.piecewise***. Pro závěrečnou statistickou validaci je použita knihovna ***sklearn.metrics***.

Vytvořené funkce

- **linear_regression(*xs*, *ys*)**: implementace základní lineární regrese pomocí metody nejmenších čtverců. Vstupy jsou vektory *xs* a *ys*,
- **piecewise_linear_fit(*x*, *y*, *breakpoints*)**: funkce provějící lineární regresi po částech na zadaných segmentech určených breakpointy,
- **plot_my_piecewise_regression(*x*, *y*, *breakpoints*)**: vizualizace výsledků vlastní metody regrese po částech,
- **compare_regressions(*x*, *y*, *breakpoints*)**: porovnání vlastní implementace lineární regrese po částech s implementací z knihovny *numpy* – zobrazení obou modelů na jednom grafu,
- **validate_regression(*y*, *y_fit*, *k*)**: výpočet statistických metrik pro statistickou validaci kvality modelu, kde y jsou skutečné hodnoty, *y_fit* modelované hodnoty a k počet parametrů modelu.