

Princípy dátovej vedy Zima 2025

Domáca úloha č. 2

Termín: 26.11.2025, 22:00

Skôr, ako začnete riešiť domácu úlohu, si prečítajte všeobecné pokyny na konci tohto dokumentu.

Dáta pre túto úlohu boli získané z kaggle.

V tejto úlohe sa budeme zaoberať modelovaním spotreby elektriny. K dispozícii máte údaje o spotrebe elektriny pre 188 krajín v rokoch 1980 až 2021. Vašou úlohou bude preskúmať varianty jednoduchého autoregresného modelu, ktorý pre nejakú konštantu p (môžete predpokladať, že $p \leq 5$) bude používať spotrebu v rokoch $t - p, \dots, t - 1$ na predikciu spotreby v roku t . Ak chceme následne predikovať hodnotu v roku $t + 1$, použijeme ako atribúty skutočné hodnoty z rokov $t - p + 1, \dots, t - 1$ a predikciu pre rok t . Podobne môžeme pokračovať v predikcii pre ďalšie roky.

a) Trénovanie, validácia a testovanie. Ako prvý krok potrebujete navrhnuť rozdelenie dát na trénovaciu, validačnú a testovaciu množinu. Konzultačná firma Gepeto, s.r.o. vám navrhla niekoľko spôsobov, ku každému z nich napíšte prečo je alebo nie je vhodný na náš predikčný účel. Následne navrhните, zdôvodnite a implementujte vašu vlastnú metódu na rozdelenie dát.

1. Dáta z každej krajiny rozdelíme na prekrývajúce sa úseky $p + 1$ bodov. Tieto úseky následne náhodne rozdelíme na trénovaciu, validačnú a testovaciu vzorku v pomere 6:2:2.
2. Dáta z každej krajiny rozdelíme na neprekrývajúce sa úseky $p + 1$ bodov. Tieto úseky následne náhodne rozdelíme na trénovaciu, validačnú a testovaciu vzorku v pomere 6:2:2.
3. Dáta z rokov 1980-2002 použijeme ako trénovacie, z rokov 2003-2011 ako validačné a z rokov 2012-2021 ako testovacie. Dáta v každej z množín rozdelíme na prekrývajúce sa úseky $p + 1$ bodov.
4. Krajiny náhodne rozdelíme do trénovacej, validačnej a testovacej vzorky v pomere 6:2:2 (všetky dáta z konkrétnej krajiny budú buď trénovacie, testovacie alebo validačné). Dáta v každej z množín rozdelíme na prekrývajúce sa úseky $p + 1$ bodov.

V reporte: Krátke zdôvodnenie vhodnosti alebo nevhodnosti každej z navrhovaných metód a návrh vlastnej metódy s krátkym zdôvodnením.

V súbore main.py: Implementujte a v kóde jasne označte vašu metódu.

b) Predikcia spotreby krajiny. Navrhните predikčný model, ktorý bude vždy z p predchádzajúcich hodnôt spotreby predikovať m budúcich hodnôt. Aké výhody a nevýhody má natrénovanie jedného globálneho predikčného modelu z dát všetkých krajín vs.

natrénovanie predikčného modelu zvlášť pre každú krajinu? Na určenie hyperparametrov vášho modelu (vrátane hodnoty p) použite validačný data set, na konečné vyhodnotenie globálneho modelu vs. lokálnych modelov pre jednotlivé krajiny použite testovací/testovacie data sety. Hodnotiacu metriku si môžete zvoliť. Vybranú hodnotu p (z globálneho modelu) použite aj vo všetkých ďalších podúlohách.

V reporte: Popíšte a odargumentujte vhodnosť vášho prístupu k tejto úlohe (váš dizajn modelu a proces voľby hyperparametrov). Porovnajte globálny model s modelmi pre jednotlivé krajiny a vyhodnoťte, ktorý je lepší a prečo.

V súbore main.py: Kód, ktorý implementuje tréning vášho modelu, ako aj výber hyperparametrov.

c) Podobnosti medzi krajinami. Medzi profilmi spotreby jednotlivých krajín existujú podobnosti. Rozdeľte dáta zo všetkých krajín na prekrývajúce sa úseky dĺžky w (variabilný argument) a navrhните funkciu, ktorá meria vzdialenosť dvoch takýchto úsekov. Vaša funkcia by mala brať do úvahy špecifiká daného problému, napríklad, že spotreba elektriny sa často škáluje s veľkosťou krajiny (údaj o veľkosti krajiny vám netreba). Použite vašu definíciu vzdialenosti na zhľukovanie a na základe rozdelenia do zhľukov s pomocou vhodnej vizualizácie ilustrujte trendy alebo spoločné črty zhľukov v dátach. Porovnajte výsledky s použitím jednoduchšej euklidovskej vzdialenosti.

V reporte: Návrh definície vzdialenosti, jeho zdôvodnenie, výsledky, porovnanie s použitím euklidovskej vzdialenosti.

V súbore main.py: Jasne označená implementácia funkcií potrebných pre vytvorenie výsledkov v reporte.

d) k najpodobnejších úsekov. V tejto podúlohe použijeme ako referenciu všetky dáta z krajín, ktoré dostanete ako zoznam pomocou volania funkcie `get_reference_countries`. Všetky referenčné dáta rozdeľte na prekrývajúce sa úseky dĺžky $p + 1$. Pre predikovanie budúcej spotreby v každej novej krajine použijeme vašu funkciu vzdialenosti z predchádzajúcej úlohy. Konkrétne, pomocou vašej funkcie nájdeme k najbližších referenčných úsekov (porovnávame vždy s prvými p hodnotami referenčného úseku) a spriemerujeme ich cieľové hodnoty ($p + 1$ -vá hodnota v úseku).

Naimplementujte túto metódu a použite ju na predikciu spotreby elektriny v rokoch 2016-2021 pre všetky krajiny zo zoznamu `get_unaccepted_countries`. Porovnajte modely pre $k = 1$, $k = 3$, $k = 5$ a model pre danú krajinu natrénovaný v časti b) so skutočnosťou. (Nezabudnite, že od roku 2016 máte ako atribúty namiesto skutočných hodnôt používať vaše predikcie.) Ukážte príklady krajín s rôznou úspešnosťou a zamyslite sa nad dôvodmi.

V reporte: Ilustratívne grafy zobrazujúce skutočnosť do roku 2015 a hodnoty od roku 2016 predikované uvedenými modelmi spolu so skutočným vývojom pre minimálne 3 vybrané krajiny. Detailné zdôvodnenie pozorovaných trendov.

V súbore main.py: Jasne označená implementácia funkcií potrebných pre vytvorenie výsledkov v reporte.

Všeobecné pokyny

Diskusia so spolužiakmi je vítaná, ale odpisovanie je prísne zakázané. Chceme vidieť vašu individuálnu prácu, diskusiu so spolužiakmi deklarujte. Rovnako to platí aj pre prácu s ChatGPT, Copilotom a podobnými nástrojmi. Ich použitie (ktoré nesmie byť excesívne) deklarujte v kóde.

Kód píšete čitateľne a dokumentujete ho. Pred odovzdaním odstráňte zbytočné výpisy a debugovacie konštrukty.

Súbory odovzdávajte do systému `vektor.fmph.uniba.sk`. V prípade otázok všeobecného charakteru (ktoré sa netýkajú vášho konkrétneho riešenia) využite verejnú diskusiu v tomto systéme. V prípade otázok špecifických pre vaše riešenie použite e-mailovú adresu `spitalsky3@uniba.sk`.