

Princípy dátovej vedy Zima 2025

Domáca úloha č. 1

Termín: 27.10.2025, 22:00

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

Dáta pre túto úlohu boli získané z kaggle. Pôvodný zdroj: *Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the Climate Assessment. Int. J. of Climatol., 22, 1441-1453*. Dáta a metadáta dostupné na <http://www.ecad.eu>.

V tejto úlohe budete vytvárať plne automatický systém, ktorý z meteorologických dát odčítaných v deň d predikuje priemernú teplotu v deň $d + 1$. Váš program z externého dátového zdroja (funkcia `get_data`) získa $n = 10$ dátových bodov pre tréning (matica vstupných atribútov `train_x` a vektor cieľových hodnôt `train_y`), natrénuje na nich jednoduchý lineárny model a na testovacom data sete z toho istého zdroja (matica testovacích vstupných atribútov `test_x` a vektor testovacích cieľových hodnôt `test_y`) vyhodnotí úspešnosť pomocou indikátoru MAE (mean absolute error). V prípade potreby môžete vo funkcii `get_data` zafixovať konkrétny dataset nastavením parametrov `force_city` a `seed`.

Ako riešenie úlohy odovzdávate modifikovaný súbor `main.py`, ktorého kostra je prílohou zadania a do ktorého doplníte vlastnú implementáciu jednotlivých častí úlohy. Odovzdávate tiež report v PDF formáte s obrázkami a s odpoveďami na jednotlivé otázky v zadaní.

a) Chýbajúce dáta. V niektorých prípadoch nie sú v tréningových dátach z externého zdroja prítomné všetky atribúty. Po porade so šéfom ste dostali dve možnosti: zahodiť dátové body (riadky) alebo atribúty (stĺpce), ktoré obsahujú chýbajúce hodnoty. Ktorú z alternatív by ste si vybrali a prečo?

V súbore `main.py`: Implementujte vybranú stratégiu vo funkcii `clean_data`.

V reporte: Krátko zdôvodnite, prečo ste si vybrali danú stratégiu.

b) Kryptický atribút. Dodatočne sa zistilo, že dáta vytváral medzinárodný tím bez dostatočnej dokumentácie, vďaka čomu v atribúte `wind_speed` niektorí meteorológovia udávajú údaje v metrických jednotkách (m/s), kým iní v imperiálnych jednotkách (mph). Nie je možné spätne vytrasovať, kto zadával ktoré údaje. Ako by ste postupovali pri predspracovaní tohto atribútu a prečo?

V súbore `main.py`: Implementujte vašu stratégiu vo funkcii `clean_data`. V prípade, že sa rozhodnete túto podúlohu neriešiť, atribút `wind_speed` jednoducho zahod'te.

V reporte: Krátko popíšte a zdôvodnite vašu stratégiu.

c) Jednoduchá lineárna regresia. Použite tréningové dáta na natrénovanie jednoduchého lineárneho modelu. Vypíšte tri pre regresiu najdôležitejšie atribúty. Následne vyhodnoťte úspešnosť vášho modelu na testovacích dátach tak, že vypočítate testovaciu

chybu MAE a vykreslite bodový graf (scatterplot) predikovaných vs. skutočných cieľových hodnôt na testovacích dátach. (Pozor: pri každom spustení dostane váš program z externého zdroja iný dataset.)

Môžete použiť `LinearRegression` z modulu `sklearn.linear_model`.

V súbore `main.py`: Implementujte vyššieuvedené vo funkcii `basic_model`.

V reporte: Váš model na niektorých datasetoch funguje veľmi dobre a na niektorých príliš nie. Ukážte výstupy (vrátane bodových grafov) pre aspoň 2 príklady, kedy vaša regresia fungovala dobre a 2 príklady, kedy výsledky neboli príliš dobré. Pokúste sa identifikovať dôvody, prečo v niektorých prípadoch model zlyháva.

d) Vylepšený model. Navrhňte vylepšený stabilnejší model. Môžete sa napríklad pokúsiť o automatizovaný výber vhodných atribútov, generalizovanú lineárnu regresiu s nelineárnymi atribútmi a pod.

V súbore `main.py`: Implementujte váš vylepšený model vo funkcii `improved_model`.

V reporte: Popíšte a zdôvodnite váš prístup. Na konkrétnych príkladoch (výstupy vrátane bodových grafov) porovnajte váš model so základným modelom z predchádzajúcej podúlohy a demonštrujte silné a slabé stránky vášho vylepšeného modelu.

e) Regularizácia. Jednou z možností, ako zlepšiť fungovanie základného modelu, je použiť regularizáciu. Vyskúšajte regularizáciu pomocou LASSO a Ridge regresie, pričom na tom istom datasete vždy otestujete hodnoty parametra α z vami určeného intervalu a porovnajte výsledky s MAE základného modelu. Výsledky vykreslite formu jednoduchého grafu, ktorý ukazuje priebeh MAE v závislosti od parametru α .

Môžete použiť `Lasso` a `Ridge` z modulu `sklearn.linear_model`.

V súbore `main.py`: Implementujte vyššieuvedené vo funkcii `regularized_model`

V reporte: Vyberte niekoľko (aspoň 2) reprezentatívnych prípadov, ktoré ukazujú rôzne priebehy MAE regularizovaných modelov v závislosti od parametru α . Zdôvodnite, prečo tieto grafy vyzerajú tak ako vyzerajú.

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 dokumentujte 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`.