Princípy dátovej vedy – Domáca úloha 1
Spracovanie dát, lineárna regresia

O úlohe

Meteorologický pracovníci z rôznych miest po Európe chcú od vás, aby ste vytvorili predikčný model na priemernú teplotu v nasledujúci deň. Bohužiaľ, poskytnú vám k tomu nie úplne kvalitné dáta. Vašou úlohou bude vytvoriť zautomatizovaný framework, ktorý tieto dáta spracuje a vytvorí regresný model, ktorý z rôznych poskytnutých meteorologických údajov spoľahlivo odhaduje budúce teploty ovzdušia.

Zdroj dát, ich nespracovanú podobu a metadáta nájdete v priečinku `data/raw`. Predpripravená funkcia `get_data` v súbore `util.py`, ktorá má za cieľ simulovať meteorologického pracovníka, ktorý vám poskytne dáta, využíva predspracované `data/weather.csv`.

Pravidlá

- Úloha má predpripravenú štruktúru. Svoj kód píšte iba do súboru `main.py`. Pri hodnotení budeme používať iba tento váš skript, čiže zmeny v zvyšných súboroch sú irelevantné. Každá podúloha má predpripravenú svoju funkciu, no súbor `main.py` môžete ľubovoľne meniť.
- Kód dokumentujte.
- Debugovacie a zbytočné printy odstráňte, relevantné medzivýsledky printovať môžete.
- Dôležitou súčasťou tejto úlohy je **report**, ktorý odovzdajte spolu s úlohou. Vôbec nemusí byť dlhý. Zhrňte v ňom, čo ste robili v jednotlivých úlohách a prečo ste to robili. Taktiež v ňom odpovedzte na otázky položené v zadaní. Chceme vidieť kritický pohľad na prácu, ktorú ste spravili. Aj nekvalitný model, s dobrým zdôvodnením, prečo je nekvalitný, a prečo nebolo možné ho zlepšiť, má šancu získať plný počet bodov.
- Úlohu odovzdajte do Google Classroom zadania ako zip, ktorý obsahuje report a upravený súbor `main.py`.
- Diskusia so spolužiakmi je vítaná, ale odpisovanie je prísne zakázané. Chceme vidieť vašu individuálnu prácu. Rovnako to platí aj pre prácu s ChatGPT, Copilotom a podobnými nástrojmi. Ich použitie (ktoré nesmie byť excesívne) deklarujte v kóde.
- V prípade problémov a nejasností kontaktujte `spitalsky3@uniba.sk`. Takisto nám prosím dajte vedieť o prípadných chybách v zadaní.

Zadanie

V súbore `main.py` máte predpripravených pár riadkov kódu. V nich sa načítajú dáta potrebné pre túto úlohu. Konkrétne:

- `train_x` – matica vstupných premenných použitá na trénovanie regresného modelu
- `train_y` – vektor výstupných hodnôt, ktoré model predikuje počas trénovania
- `test_x` – matica vstupných premenných, ktorá sa nepoužíva pri trénovaní a slúži na hodnotenie modelu
- `test_y` – vektor skutočných výstupných hodnôt pre `test_x`

1. Spracovanie dát (2b)

Pred ľubovoľnou analýzou dát je potrebné ich dostať do podoby vhodnej na spracovanie. To bude vašou prvou úlohou. Dataset `train_x`, s ktorým budete pracovať, vykazuje rôzne známky nekvality, konkrétne

- chýbajúce dáta,
- číselné dáta uložené ako string,
- stĺpec irelevantný pre analýzu,
- stĺpec `wind_speed` v nekonzisentných jednotkách (m/s alebo km/h, jednotky sú uvedené pri hodnote).

Prvou úlohou je teda tieto dáta upratať, aby ste s nimi mohli jednoducho pracovať v ďalších úlohách. Dajte si však pozor, pri každom spustení `main.py` vám prichádzajú dáta z náhodného mesta (teda s rôznou množinou stĺpcov) a s náhodnými nekvalitami. Preto sa ubezpečte, že vaše funkcie sú univerzálne a nevyžadujú znalosti napr. názvu stĺpcov.

Pre túto úlohu je pripravená funkcia `clean_data`, ktorej vstupom sú nekvalitné dáta a výstupom by mali byť upratané dáta ako `pandas.DataFrame`. V reporte krátko zhrňte, ako ste k tejto úlohe pristupovali.

2. Prvé predikcie (2b)

Na upratených dátach budete robiť svoju prvú lineárnu regresiu (na tomto predmete). Pomocou dát v `train_x` a `train_y` spočítajte optimálny odhad vektoru parametrov $\hat{\beta}$, môžete použiť napríklad `sklearn.linear_model.LinearRegression`.

Na vyhodnotenie kvality modelu použijeme dáta `test_x` a `test_y`. Pomocou vyššie získaného $\hat{\beta}$ odhadnite priemerné teploty v nasledujúci deň. Vykreslite tieto odhadované hodnoty oproti skutočným priemerným teplotám v `test_y` ako scatterplot. Čo pozorujete? Keďže dostávate náhodné dáta, spustite súbor `main.py` viackrát a zhrňte vaše pozorovania (vo funkcii `get_data` môžete využiť optional argument `force_city`, ktorým si viete nastaviť mesto, z ktorého chcete dáta). Pár príkladov takýchto obrázkov (aj s mestami, z ktorých sú ich dáta) s popisom správania môžete zahrnúť do reportu.

Ako metriku nekvality modelu budeme používať mean absolute error (MAE), teda priemerná aboslútna odchylka odhadu od skutočnej hodnoty. Spočítajte ho pre tento základný model. Sú pre rôzne mestá rôzne hodnoty MAE? Ak áno, napíšte hypotézu, prečo by to tak mohlo byť.

Pre túto úlohu je pripravená funkcia `basic_model`, ktorej vstupom sú dáta `train_x`, `train_y`, `test_x` a `test_y`. Optional argument je `show_fig`. V kóde zahrňte, že pre jeho hodnotu `True` sa spomínaný obrázok vykreslí, inak nie. Výstupom tejto funkcie je `float`, konkrétne MAE natrénovaného modelu.

3. Vylepšené predikcie (2b)

Ako ste sa dozvedeli na prednáške, overtraining je fenomén, pri ktorom sa model správa príliš dobre na trénovacích dátach, ale stráca schopnosť generalizácie. Využite preto rôzne techniky (napríklad používanie iba podmnožiny stĺpcov, transformácie premenných...) na vylepšenie testovacej MAE chyby modelu. Môžete na spočítanie tejto chyby, samozrejme, použiť funkciu `basic_model`, ktorej dáte modifikované dáta. V tomto kroku nepoužívajte regularizovanú regresiu.

Pre túto úlohu je predpripravená funkcia `improve_model`, ktorej vstupom sú datasety `train_x`, `train_y`, `test_x` a `test_y` a do argumentu `MAE_basic` dajte MAE chybu z úlohy 2. Výstupom by mali byť modifikované (za účelom zníženia overtrainingu) datasety `train_x`, `train_y`, `test_x` a `test_y` a vylepšená MAE chyba. Samozrejme, relevantný výsledok je aj keď sa túto chybu zlepšiť nepodarí.

Použité techniky, prístup a pozorovania výsledkov spíšte do reportu.

4. Regularizácia (2b)

V tomto kroku sa pozrieme na to, čo sa deje pri variácii regularizačného parametra α v regularizovanej regresii. Na modifikovanom datasete z predchádzajúceho kroku natrénujte Lasso a hrebeňovú regresiu pre veľa rôznych parametrov α v hodnotách medzi 0 až 10 (odporúčam použiť `Lasso` a `Ridge` z modulu `sklearn.linear_model`). V oboch prípadoch vykreslite krivku závislosti parametra α a MAE chyby modelu.

Opíšte správanie, ktoré vidíte. Prečo je také? Je očakávané, že v oboch prípadoch bude pre vysoké hodnoty α chyba rásť. Prečo? Tiež by ste si mali v krivkách závislosti všimnúť aspoň jeden zlom pre Lasso a hladký trend chyby pre hrebeňovú regresiu. Sformulujte hypotézu, prečo to je tak. V reporte napíšte odpovede na tieto otázky, ako aj aspoň dva obrázky kriviek závislostí.

Ak pozorujete iné trendy ako sú tu uvedené, opíšte ich správanie a sformulujte hypotézu, prečo také trendy mohli vzniknúť.

Pre túto úlohu máte predpripravenú funkciu `regularized_model`, ktorej vstupom by mali byť modifikované datasety z úlohy 3. Táto funkcia nemá výstup, má vykresliť dva požadované obrázky.