Princípy dátovej vedy Zima 2025
Domáca úloha č. 3
Termín: 19.12.2025
neskoré odovzdávanie (bez postihu) do 31.12.2025, 22:00

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

Dáta pre túto úlohu boli získané z kaggle.

Máte k dispozícii dáta od subjektov, ktorí vykonávali 11 rôznych aktivít s mobilom vo vrecku. Počas týchto aktivít sa odčítavali údaje z akcelerometra a gyroskopu. Ku každému odčítaniu údajov máte k dispozícii, ktorú aktivitu aktuálne subjekt vykonával, počet milisekúnd od jej začiatku a unikátne ID. Rovnaké ID znamená, že sa jedná o kontinuálne vykonávanie tej istej aktivity.

Súbor `main.py` obsahuje iba veľmi jednoduchú kostru, môžete ho ľubovoľne meniť. Odovzdávate report vo formáte pdf a upravený súbor `main.py`. Vizualizácie sú vždy vítané.

a) Deskriptívne štatistiky. Priame určovanie aktivity na základe jediného odčítania z gyroskopu a akcelerometra pravdepodobne nebude spoľahlivé. Vytvorte funkciu, ktorá spracuje okno meraní s dĺžkou T sekúnd a vypočíta z týchto dát vami navrhnuté štatistiky (ideálne nie iba priemer a varianciu), o ktorých si myslíte, že dobre zachytávajú správanie počas daných T sekúnd. Myslite na to, že počet meraní počas T sekundového okna sa môže líšiť. Parameter T by mal byť variabilný a ľahko nastaviteľný v kóde.

V reporte: Zdôvodnite váš výber štatistík a ich vhodnosť pre túto úlohu. Ak v neskorších podúlohách zistíte, že pomocou vašich štatistík klasifikátory nefungujú dobre, upravte ich. Zahrňte tento proces do reportu.

V súbore `main.py`: Implementujte spočítanie vami navrhnutých týchto deskriptívnych štatistík pre T-sekundové okno dát.

b) Trénovanie, validácia a testovanie. Za účelom ďalšej tvorby modelov potrebujeme dáta rozdeliť na trénovaciu, validačnú a testovaciu vzorku. Preveďte T-sekundové okná z dát na deskriptívne štatistiky z podúlohy a) a tieto dáta rozdeľte na tri vzorky vami zvolenou stratégiou. Budete používať prekrývajúce sa okná alebo nie? Pôvodné dáta sú časové rady, má to rovnaké implikácie, ako pri časových radoch z domácej úlohy 2?

V reporte: Detailne odargumentujte váš dizajn trénovacej, validačnej a testovacej množiny. Prečo to je vhodné pre tento typ úlohy (aj vzhľadom na použitie dát v neskorších podúlohách)?

V súbore `main.py`: Implementácia vašej train-val-test split stratégie, pričom T je variabilný argument. Dajte si pozor, aby vám nevznikli T-sekundové okná s nulovým počtom odčítaní zo senzorov.

c) Interpretovateľný model. Na trénovacej množine natrénujte jednoduchý rozhodovací strom. Pointou tejto úlohy je dostať interpretovateľný model, nie najlepší možný klasifikátor. Nemusíte klasifikovať všetkých 11 aktivít, ale môžete si vybrať nejakú ich podmnožinu z trénovacích dát. Napríklad zvoľte také aktivity, o ktorých si myslíte, že budú mať rozdielne správanie v meraniach zo senzorov, resp. v deskriptívnych štatistikách. Zdôvodnite výber tejto podmnožiny. Takisto, obmedzte hĺbku rozhodovacieho stromu a môžete pracovať aj iba s podmnožinou deskriptívnych štatistík, ak ich máte veľa. Detailne interpretujte výsledný natrénovaný model. Pokúste sa ľudskou rečou vysvetliť rozhodovacie pravidlá v každom vrchole. Túto úlohu nemusíte robiť pre viacero podmnožín aktivít a snažiť sa generalizovať správanie, stačí dobrý popis v konkrétnej situácii.

V reporte: Detailná interpretácia klasifikačných pravidiel natrénovaného rozhodovacieho stromu.

V súbore `main.py`**:** Trénovanie rozhodovacieho stromu spolu s výpisom, na základe ktorého budete výsledky interpretovať.

d) Dobrý klasifikátor. Teraz už na celej trénovacej množine natrénujte čo najlepší možný klasifikátor, samozrejme, za pomoci validácie. Metriku si môžete zvoliť, ale voľbu odôvodnite. Klasifikátor by mal mať formu hlasovacej schémy, čiže napríklad náhodné lesy alebo môžete využiť `sklearn.ensemble.VotingClassifier`. Znova, odôvodnite svoju voľbu. Nadizajnovaný a natrénovaný model otestujte pomocou validácie a zhodnoťte výsledky. Aj v tejto úlohe by mal byť T jednoducho nastaviteľný parameter.

V reporte: Zdôvodnenie voľby metriky a hlasovacej schémy. Interpretujte výsledky na testovacej množine a porovnajte so správaním na validačnej množine.

V súbore `main.py`**:** Flexibilná implementácia hlasovacej schémy, kde konkrétny dizajn je zvolený validáciou. Nezabudnite, že T je meniteľný parameter.

e) Úloha času. Váš model by mal byť schopný rozoznať aktivitu, ktorú subjekt vykonáva, v čo najkratšom možnom čase. Zistite, ako sa správa klasifikátor z podúlohy d), keď meníte šírku okna T. Koľko sekúnd meraní je potrebných na relatívne spoľahlivú klasifikáciu? Vizualizujte správanie metriky zvolenej v podúlohe d) v závislosti od T a interpretujte výsledky. Sformulujte hypotézy, prečo by sme mohli takéto správanie očakávať.

V reporte: Interpretácia správania sa klasifikátora v závislosti od T, vizualizácia a odargumentované vysvetľujúce hypotézy.

V súbore `main.py`**:** Kód na vygenerovanie potrebných vizualizácií.

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 píšte 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 vašeho konkrétneho riešenia) využite verejnú diskusiu v tomto systéme. V prípade otázok špecifických pre vaše riešenie použite e-mailovú adresu `spitalsky3@uniba.sk`.