Inteligentná analýza údajov, zimný semester 2018/19

Hodnotenie a podmienky absolvovania predmetu

Priebežné hodnotenie

- projekt: **max. 40 bodov**
    - práca na projekte a úlohách v rámci cvičení: 5 bodov
    - 1. fáza - prieskumná analýza (v 6. týždni): 10 bodov
    - 2. fáza - predspracovanie údajov (v 10. týždni): 13 bodov
    - 3. fáza - strojové učenie (v 13. týždni): 12 bodov
- Projekt bude riešený vo dvojiciach.

Záverečné hodnotenie

- body získané počas semestra: max. 40 bodov
- skúška: max. 60 bodov

Podmienky získania zápočtu

1. vypracovanie všetkých častí projektu v akceptovateľnej kvalite, jeho odovzdanie a prezentovanie podľa harmonogramu
2. aktívna účasť na cvičeniach
3. získanie aspoň 25 bodov počas semestra

Neskoré odovzdanie

V prípade nedodržania termínu na odovzdanie jednotlivých fáz projektu do systému AIS bude možné danú fázu odovzdať do siedmich dní s 50% penalizáciou. Neskoršie odovzdanie nebude možné. Neodovzdanie niektorej z fáz projektu bude mať za následok neudelenie zápočtu.

Podmienky absolvovania predmetu

Pre absolvovanie predmetu platia všeobecné podmienky absolvovania a hodnotenia predmetov. Za samozrejmosť sa považuje dodržanie pravidiel akademickej korektnosti. Akékoľvek vydávanie cudzej práce za vlastnú je neakceptovateľné, a to v akomkoľvek rozsahu. Ak sa ho dopustíte, automaticky to znamená neúspešné ukončenie predmetu (FX) a podnet na disciplinárnu komisiu.

Projekt

Cieľom projektu je osvojiť si základné koncepty a techniky analýzy dát, pochopiť, ako fungujú a získať intuíciu pre ich vhodnú aplikáciu za účelom objavovania znalostí v dátach. Mali by ste tiež získať predstavu, aké otázky vieme pomocou analýzy dát zodpovedať a byť schopní aplikovať a vyhodnotiť základné prístupy strojového učenia.

Projekt sa vypracúva vo dvojiciach. Pri riešení sa očakáva využitie jazyka `Python` a dostupných knižníc na analýzu dát (`pandas`, `matplotlib` a pod.). V každej fáze sa odovzdáva vykonateľný `Jupyter Notebook`, ktorý by mal zachytávať a vhodne dokumentovať všetky vykonané transformácie nad dátami. Odovzdaný notebook musí obsahovať nielen kód, ale aj jeho výsledky (vypočítané hodnoty, výpisy, vizualizácie a pod.) spolu s komentárom intrepretujúcim získané výsledky a z toho plynúce rozhodnutia pre ďalšie kroky dátovej analýzy. Schopnosť dobre odkomunikovať a vybrať relevantné výsledky analýzy bude predstavovať významnú zložku hodnotenia.

Dáta

Každá dvojica bude pracovať s im náhodne pridelenou dátovou sadou. Dáta predstavujú záznamy o pacientoch s chorobou štítnej žľazy. Vašou úlohou je vedieť predikovať hodnotu **Y** (môže sa líšiť v závislosti od pridelenej dátovej sady). Budete sa musieť pritom vysporiadať s viacerými problémami, ktoré sa v dátach nachádzajú (formáty dát, chýbajúce, nezmyselné alebo vychýlené hodnoty a pod.).

Prieskumná analýza (max. 10b)

Prieskumná analýza je kľúčovou časťou analýzy dát. Bez nej nie sme schopní dáta spracúvať, pretože nevieme, čo sa v nich nachádza. Využíva sa pritom predovšetkým deskriptívna štatistika a rôzne podporné vizualizácie.

V tejto fáze sa od vás očakáva:

- **Opis dát spolu s ich charakteristikami** (počet záznamov, počet atribútov, ich typy, distribúcie, základné deskriptívne štatistiky a pod.) vrátane preskúmania vzťahov medzi dvojicami atribútov.
- **Formulácia a overenie hypotéz o dátach.** Mali by ste sformulovať aspoň dve hypotézy o dátach, ktoré budú relevantné v kontexte zadanej predikčnej úlohy. Príkladom hypotézy môže byť, že *pacienti s chorobou štítnej žľazy majú v priemere inú (vyššiu/nižšiu) hodnotu nejakej látky alebo hormónu ako pacienti bez danej choroby*. Vami sformulované hypotézy overte vhodne zvoleným štatistickým testom.
- **Identifikácia problémov v dátach spolu s predpokladaným scenárom riešenia v ďalšej fáze**, t. j., čo budete musieť v rámci predspracovania vyriešiť (aj s naznačením možností, ako tieto problémy plánujete v ďalšej fáze riešiť). Medzi problémy, na ktoré môžete v rámci analýzy naraziť, patria napr.:

    - nevhodná štruktúra dát (dáta nie sú v tabuľkovej podobe alebo jedna entita je opísaná viacerými riadkami tabuľky)
    - duplicitné záznamy, resp. nejednoznačné mapovanie medzi záznamami
    - nejednotné formáty dát
    - chýbajúce hodnoty
    - vychýlené (odľahlé) hodnoty
    - a ďalšie, t. j. v dátach sa môžu nachádzať aj iné, tu nevymenované problémy, ktoré tiež treba identifikovať a vo vašej analýze adresovať.

V odovzdanej správe (`Jupyter Notebooku`) by ste tak mali vedieť zodpovedať na otázky:

- Majú dáta vhodný formát pre ďalšie spracovanie? Ak nie, aké problémy sa v nich vyskytujú?
- Sú niektoré atribúty medzi sebou závislé? Od ktorých (jednotlivých) atribútov závisí predikovaná premenná?
- Sú v dátach chýbajúce hodnoty? Ako sú reprezentované? Ako plánujete riešiť problém chýbajúcich hodnôt pre jednotlivé atribúty, resp. pozorovania? (Pre rôzne atribúty môže byť vhodné použiť rôzne stratégie.)
- Nadobúdajú niektoré atribúty nezmyselné (nekonzistentné) či inak výrazne odchýlené hodnoty? Ktoré?
- Ako plánujete v ďalšej fáze tieto identifikované problémy adresovať / riešiť?

Správa sa odovzdáva v 6. týždni semestra na cvičení (dvojica svojmu cvičiacemu odprezentuje vykonanú prieskumnú analýzu v `Jupyter Notebooku`). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do **nedele 28.10.2018** do **23:59**.

Celkové hodnotenie za 1. fázu (10b) je rozdelené nasledovne:

- Základný opis dát - 2b
- Párová analýza - 2b
- Formulácia a štatistické overenie hypotéz - 2b
- Identifikácia problémov v dátach - 2b
- Celkový dojem - 2b

Predspracovanie (max. 13b)

Na základe identifikovaných problémov v dátach a návrhu ich riešenia v predchádzajúcej fáze treba zrealizovať predspracovanie. Výsledkom by mala byť upravená dátová sada (vo formáte `csv`) vo vhodnom tvare pre zvolený algoritmus strojového učenia (v našom prípade *rozhodovacie stromy*). Zároveň, keďže predspracovaním sa mohol zmeniť tvar a charakteristiky dát (počet atribútov, distribúcie hodnôt a pod.), treba znovu zrealizovať podstatné časti prieskumnej analýzy. Významnú časť hodnotenia bude predstavovať znovupoužiteľnosť (replikovateľnosť) predspracovania.

V tejto fáze sa od vás očakáva:

- **Integrácia dát a prípadná deduplikácia záznamov.** Výsledkom by mala byť jednotná tabuľková reprezentácia dát, ktorá bude predstavovať vstup pre ďalšie spracovanie a (v 3. fáze) strojové učenie.

- **Realizácia krokov predspracovania dát a ich zdokumentovanie.**
    - Pri riešení chýbajúcich hodnôt vyskúšajte rôzne stratégie (očakáva sa vyskúšanie a porovnanie *minimálne dvoch* stratégií, pričom aspoň jedna z nich musí byť zvolená z posledných troch menovaných):
        - nahradenie chýbajúcej hodnoty mediánom
        - nahradenie chýbajúcej hodnoty priemerom
        - nahradenie chýbajúcej hodnoty pomerom ku korelovanému atribútu
        - nahradenie chýbajúcej hodnoty priemerom segmentu
        - nahradenie chýbajúcej hodnoty pomocou lineárnej regresie
        - nahradenie chýbajúcej hodnoty pomocou algoritmu k-najbližších susedov
    - Podobne postupujte aj pri riešení vychýlených (odľahlých) hodnôt, pričom porovnajte odstránenie vychýlených pozorovaní s aspoň jednou zvolenou stratégiou ich nahradenia, resp. transformácie:
        - odstránenie vychýlených (odľahlých) pozorovaní
        - nahradenie vychýlenej hodnoty hraničnými hodnotami rozdelenia (5 percentilom, resp. 95 percentilom)
        - transformácia atribútu s vychýlenými hodnotami pomocou zvolenej funkcie (logaritmus, odmocnina a pod.)
- **Opätovná realizácia podstatných častí prieskumnej analýzy.** V rámci nej by ste mali vedieť zodpovedať na otázku, ako sa zmenili distribúcie hodnôt po realizácii krokov predspracovania.

Správa sa odovzdáva v 10. týždni semestra na cvičení (dvojica svojmu cvičiacemu odprezentuje vykonané predspracovanie v `Jupyter Notebooku`). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do **nedele 25.11.2018** do **23:59**.

Celkové hodnotenie za 2. fázu (13b) je rozdelené nasledovne:

- Integrácia dát a deduplikácia záznamov - 3b
- Oprava dát - 3b
- Znovupoužiteľnosť predspracovania - 3b
- Opätovná prieskumná analýza - 2b
- Celkový dojem - 2b

Strojové učenie (max. 12b)

Pri dátovej analýze nemusí byť naším cieľom získať len znalosti obsiahnuté v aktuálnych dátach, ale aj natrénovať model, ktorý bude schopný robiť rozumné predikcie pre nové pozorovania (v našom prípade pre nových pacientov). Na to sa využívajú techniky strojového učenia. V tomto projekte sa zameriame na *rozhodovacie stromy* vzhľadom na ich jednoduchú interpretovateľnosť.

V tejto fáze dostanete nový dataset, na ktorom oddemonštrujete znovupoužiteľnosť vami realizovaného predspracovania. Vami natrénované klasifikátory budú porovnané medzi sebou; uvidíte tak, ako dobre ste sa umiestnili v rámci vášho cvičenia, resp. celého predmetu.

V poslednej fáze sa od vás očakáva:

- **Predspracovanie nového datasetu vami realizovaným postupom predspracovania.** Spustite postup predspracovania realizovaný v predchádzajúcom kroku nad novým datasetom. Nový dataset bude mať rovnakú štruktúru ako váš pôvodný, nebudú sa v ňom však možno nachádzať niektoré problémy (nové vám nepribudnú). Ak si spustenie predspracovania vyžiada zmeny v kóde, opíšte ich.
- **Manuálne vytvorenie a vyhodnotenie rozhodovacích pravidiel pre klasifikáciu.** Vyskúšajte jednoduché pravidlá zahŕňajúce jeden atribút, ale aj komplikovanejšie zahŕňajúce viacero atribútov (ich kombinácie). Pravidlá by v tomto kroku mali byť vytvorené manuálne na základe pozorovaných závislostí v dátach. Pravidlá (manuálne vytvorené klasifikátory) vyhodnoťte pomocou metrík *správnosť* (angl. *accuracy*), *presnosť* (angl. *precision*) a *úplnosť* (angl. *recall*).
- **Natrénovanie klasifikátora s využitím rozhodovacích stromov.** Využite algoritmus dostupný v knižnici `scikit-learn` (`CART`). Preskúmajte hyperparametre tohto algoritmu a vyskúšajte ich rôzne nastavenie tak, aby ste minimalizovali preučenie. Vysvetlite, čo jednotlivé hyperparametre robia. Pri nastavovaní hyperparametrov algoritmu využite 10-násobnú krížovú validáciu na trénovacej množine.
- **Vyhodnotenie natrénovaných klasifikátorov.** Vizualizujte natrénované pravidlá. Porovnajte algoritmy navzájom, ako aj s vašimi manuálne vytvorenými pravidlami z druhého kroku. Vyhodnoťte ich pomocou metrík *správnosť* (angl. *accuracy*), *presnosť* (angl. *precision*) a *úplnosť* (angl. *recall*).
- **Vyhodnotenie vplyvu zvolenej stratégie riešenia chýbajúcich hodnôt na správnosť klasifikácie.** Zistite, či použitie zvolených stratégií riešenia chýbajúcich hodnôt vplýva na správnosť (angl. *accuracy*) klasifikácie. Ktorá stratégia sa ukázala ako vhodnejšia pre daný problém?

Správa sa odovzdáva v 13. týždni semestra na cvičení (dvojica svojmu cvičiacemu odprezentuje vykonanú prieskumnú analýzu v `Jupyter Notebooku`). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do **nedele 16.12.2018** do **23:59**.

Celkové hodnotenie za 3. fázu (12b) je rozdelené nasledovne:

- Realizácia predspracovania nad novým datasetom a opis prípadných zmien - 2b
- Návrh a overenie manuálne vytvorených pravidiel - 2b
- Porovnanie stratégie riešenia chýbajúcich hodnôt - 2b
- Natrénovanie a otestovanie klasifikátora (klasifikátorov) - 2b
- Optimalizácia hyperparametrov - 2b
- Celkový dojem - 2b