I-SUNS: Zadanie č.1

Neurónové siete

Vo vybranom programovacom jazyku implementujte program, ktorý bude kategorizovať hry na platforme Steam do kategórií zadarmo a platené. V tomto zadaní budete pracovať s dátami z AIS. Pre dataset budú dostupné dva súbory - testovací a trénovací. Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 10 bodov je **18.10.2022 o 08:00/10:00/13:00 (pred vaším cvičením)**. Každý týždeň omeškania je penalizovaný stratou dvoch bodov.

- Načítajte dáta z oboch množín (trénovacej - *train.csv* a testovacej - *test.csv*) a pripravte ich na spracovanie neurónovou sieťou **spolu 5b**:
  - Prezrite si stĺpce v databáze (popis menej zrozumiteľných je aj na konci toho zadania). Stĺpce s predponou D_ sú skopírované priamo zo scrapera, ostatné sú už predspracované ako bolo prezentované na cvičení[1]. Z D_ stĺpcov získajte číselné hodnoty, min.: **2b**
    - z dátumového stringu rok,
    - počet majiteľov hry,
    - z tagov žáner alebo vizuál alebo náladu hry.
  - Odstráňte stĺpce, ktoré sa nedajú použiť pri ďalšom spracovaní a *null* hodnoty. **0.5b**
  - Analyzujte dataset cez EDA[2]. **2b**
  - Dáta správne normalizujte alebo škálujte. **0.5b**
- Nátrenujte na dátach neurónovú sieť **spolu 5b**:
  - Vytvorte validačnú množinu a trénujte sieť tak, aby ste vedeli sledovať priebeh kriteriálnej funkcie a chyby na trénovacej aj validačnej množine. Vyhodnoťte úspešnosť na testovacej množine. Nezabudnite vymazať pred trénovaním stĺpec *VYMAZAT_price*.
  - Vysporiadajte sa s problémom nevyrovnaných početností vo výstupných triedach.

[1] to ale ešte neznamená, že sú vhodné na trénovanie

[2] Odpovedajte napr. na otázky: *Aký je súvis medzi self-published a indie? Má počet jazykov vplyv na cenu? Ako sa menili trendy v hrách počas histórie Steamu?* Odpoveď na otázku znamená grafická aj slovná analýza. Min. 5 grafov, z toho 3 s viacerými premennými.

  - Zväčšujte sieť, kým nebudete pozorovať problém pretrénovania. Nájdite pre architektúru dobré hyperparametre. Pridajte do siete *early-stopping*.
  - Pridávajte/uberajte vstupné parametre z datasetu, sledujte zmenu v potrebných hyperpametroch pre dobré trénovanie a v následných úspešnostiach.

- Rozdelenie bodov:
  - Správne natrénovaná sieť - nepozorovať známky pretrénovania, úspešnosť presahuje náhodnú úspešnosť (v tomto prípade 50%), na konfúznej matici sa dá pozorovať správne vyfarbená diagonála. **1b**
  - Dokumentácia postupu riešenia - vyhodnotenie úspešnosti pre jednotlivé zmeny architektúry, vstupov a hyperparametrov (chceme vedieť, ktorá zmena viedla k akému výsledku, aj k tomu najlepšiemu z predošlého bodu), zobrazenie priebehu trénovania pre trénovaciu aj validačnú množinu, zobrazenie konfúznej matice pre trénovaciu aj testovaciu množinu.[3] **2b**
  - Vyhodnotenie vplyvu vstupných parametrov na výsledok trénovania - zvoľte si spôsob vyhodnotenia dôležitosti vstupných parametrov a zdokumentujte svoje experimenty. **2b**

Nepovinné úlohy

- Dobré parametre hľadajte pomocou Grid-searchu, výsledky graficky analyzujte (napr. *heat-mapou*). **1b**
- Nájdite (a vysporiadajte sa s) outliermi v trénovacej množine (hranice IQR, izolačné lesy ...) - analyzujte dosiahnuté výsledky, zlepšila sa vaša sieť? **1b**
- Zhodnoťte dôležitosť parametrov podľa Shapleyho hodnôt. **1b**
- Navrhnite a aplikujte iný spôsob kódovania publishera a developera než je použitý v dodanej dátovej množine. Vyhodnoťte výsledok trénovania po aplikovaní zmeny. **1b**

Upresnenie stĺpcov

- *appid* - Unikátny identifikátor na platforme Steam.

[3]Na analýzu výsledkov používame vyhodnotenie úspešnosti, konfúznu maticu, ... pre trénovacie aj testovacie dáta. Keď je v zadaní požiadavka na vyhodnotenie úspešnosti alebo konfúznej matice (prípadne inej metriky v ďalších zadaniach) vždy chceme tieto výsledky pre trénovaciu aj testovaciu množinu (nie validačnú, tú používame len na vyhodnotenie priebehu trénovania).

- *positive* - Počet pozitívnych hodnotení.
- *negative* - Počet negatívnych hodnotení.
- *score* - Pomer pozitívnych hodnotení ku celkovému počtu hodnotení.
- *reviews* - Slovné hodnotenie podľa linky.
- *ccu* - Concurrently connected users .
- *price* - Cena hry v eurocentoch v čase scrapovania.
- *english* - Je hra preložená do angličtiny.
- *languages* - Do koľkých jazykov je hra preložená.
- *tags* - Označenia pridané užívateľmi, viac informácií napr. tu.