I-SUNS: Zadanie č.1

Neurónové siete

Vo vybranom programovacom jazyku implementujte program, ktorý bude kategorizovať počasie do 4 kategórií(Rainy, Cloudy, Sunny a Snowy). V tomto zadaní budete pracovať s dátami z AIS. Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 15 bodov je **18.10.2024 o 9:00/11:00(pred vaším cvičením)**. Každý týždeň omeškania je penalizovaný stratou dvoch bodov. Zadanie je rozdelené na 3 časti:

- Načítajte dáta, predspracujte ich, natrénujte jednoduchý model a vyhodnoťte ho **(spolu 5b)**:
  - Prezrite si stĺpce v databáze (popis stĺpcov na konci tohto zadania). Podľa popisu zistite, či sa v jednotlivých stĺpcoch nachádzajú outliery (neobvyklé hodnoty) a odstráňte ich. **0.5b**
  - Odstráňte stĺpce, ktoré sa nedajú použiť pri ďalšom spracovaní a *null* hodnoty (pozor na odstraňovanie *null* hodnôt - aby ste neprišli o príliš veľa dát, ak v stĺpci chýba príliš veľa hodnôt, zvážte, či nie je rozumnejšie odstrániť celý stĺpec). **0.5b**
  - Nečíselné stĺpce vhodne zakódujte. **0.5b**
  - Vytvorte vstupné (X) a výstupné (y) dátové množiny. Vo vhodnom pomere rozdeľte dáta na trénovaciu, validačnú a testovaciu množinu. **0.5b**
  - Dáta správne normalizujte alebo škálujte. **0.5b**
  - Natrénujte jednoduchý klasifikačný model (tu Vám stačí Sklearn).**1b**
  - Experimentujte s rozdelením dátových množín. Natrénujte model a vyhodnoťte na trénovacej **A** testovacej množine viackrát. Vytvorne **1** tabuľku s aspoň 4 experimentami s dosahnutými úspešnosťami. Zamyslite sa nad jednotlivými výsledkami a zhodnoťte ako rozdelenie vplýva na celkovú úspešnosť.**1b**
  - Pre najlepšie natrénovaný model vykreslite konfúznu maticu pre trénovaciu aj testovaciu množinu.**0.5b**
- Analyzujte dataset cez EDA. Pracujte s upraveným aj pôvodným datasetom, aby ste mali k dispozícii všetky údaje (pracujte s dátami po odstránení outlierov a chýbajúcich hodnôt, ale použite hodnoty pred kódovaním, aby ste vedeli použiť slovné hodnoty). Niektoré vzťahy sa Vám budú hľadať lepšie predtým, než dáta

upravíte. Nájdite aspoň 5 zaujímavých vzťahov v dátach (každý za 1b), vizualizujte ich pomocou grafov (nie 5x rovnaký), inšpirujte sa ukážkami z iných projektov **(spolu 5b)**:

  - grafy bez slovného popisu nebudú hodnotené plným počtom bodov,
  - histogramy, minimálne a maximálne hodnoty nebudú hodnotené plným počtom bodov,
  - ak budú všetky nájdené vzťahy len voči cieľovej premennej, nebudú hodnotené plným počtom bodov.

- Natrénujte neurónovú sieť. Pre splnenie tejto časti zadania odporúčame použiť sofistikovanejšiu knižnicu (Keras, Pytorch, ...), prípadne doplniť funcionalitu knižnice Sklearn tak, aby ste boli schopní splniť všetky nasledovné body **(spolu 5b)**:

  - Zvoľte architektúru a nastavenia hyperparametrov tak, aby ste dosiahli pretrénovanie. Demonštrujte pomocou grafov priebehu trénovania, vyhodnotenia úspešností a konfúznej matice pre trénovaciu aj testovaciu množinu. **1b**
  - Odstráňte pretrénovanie tak, že do trénovania zavediete EarlyStopping pre skoré zastavenie trénovania. Demonštrujte pomocou grafov priebehu trénovania, vyhodnotenia úspešností a konfúznej matice pre trénovaciu aj testovaciu množinu. **1b**
  - V tomto bode by ste už mali mať správne natrénovanú sieť[1]. Skúste zmeniť niektoré hyperparametre, prípadne architektúru siete (aspoň 2) tak, aby ste natrénovali sieť aspoň 5x (napr. v prvých 3 experimentoch zmeníte hodnotu parametra rýchlosti učenia, v ďalších 2 počet neurónov v skrytej vrstve siete). Body budú udelené nasledovne:
    * Jednotlivé konfigurácie sú prehľadne zapísané v JEDNEJ tabuľke (stačí zapísať rozdiely v experimentoch, netreba pre každý experiment uvádzať všetky parametre, ak ostali zachované). **1b**
    * Pre všetky trénovania sú v tabuľke vyhodnotené dosiahnuté úspešnosti pre trénovaciu aj testovaciu množinu. **1b**
    * Pre najlepšie a najhoršie trénovanie je navyše zobrazený priebeh trénovania a konfúzne matice pre trénovaciu aj testovaciu množinu. **1b**

[1]Správne natrénovaná sieť - nepozorovať známky pretrénovania, úspešnosť presahuje náhodnú úspešnosť (v tomto prípade 25%), na konfúznej matici sa dá pozorovať správne vyfarbená diagonála.

Na čo si dať pozor!

- Nezabudnite zo vstupnej množiny odstrániť sledovanú výstupnú hodnotu, vo vašom prípade stĺpec *Weather Type*.
- Pri normalizácii/škálovaní dbajte na to, aby ste pri nastavení scaler-a (prípadne manuálnom výpočte použitých minimálnych/maximálnych hodnôt, priemeru a smerodajnej odchýlky) použili trénovacie dáta a následne už nastavený scaler (príp. vypočítané hodnoty) použili pre validačnú a testovaciu množinu.
- 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).
- Každý bod zadania musí byť zkodumentovaný, t.j. ak máte napr. demonštrovať pretrénovanie a následne ho odstrániť pridaním EarlyStopping-u, je potrebné dať do dokumentácie aj priebeh trénovania, dosiahnuté výsledky a nastavenia pred jeho použitím, inak nemáme možnosť vyhodnotiť, či ste splnili daný bod zadania.

Nepovinné úlohy

- Dobré parametre hľadajte pomocou Grid-searchu, prehľadávajte aspoň 10 rôznych kombinácií parametrov - dosiahnuté výsledky analyzujte (najlepší/najhorší výsledok, pretrénovanie, ...). **1b**
- Naštudujte si regularizačnú techniku Dropout a použite ju vo svojej štruktúre siete. **0.5b**
- Porovnajte rôzne modely založené na rôznych metódach z knižnice Sklearn (aj vysvetliť). **1-2b**

Popis stĺpcov

- Temperature: Obsahuje rôzne hodnoty teplôt od -25°C do 109°C.
- Humidity: Relatívna vlhkosť sa pohybuje v rozmedzí od 20% do 109%.
- Wind Speed: Rýchlosť vetra sa pohybuje od 0 km/h do 48,5 km/h.
- Precipitation (%): Pravdepodobnosť zrážok sa pohybuje od 0% do 109%.

- Cloud Cover: Možné hodnoty sú partly cloudy (čiastočne oblačno), clear (jasno), overcast (zatiahnuté) a cloudy (oblačno).
- Atmospheric Pressure: Atmosférický tlak sa pohybuje od približne 984 hPa do 1067 hPa.
- UV Index: Hodnoty UV indexu sú od 0 do 14.
- Season: Možné sezóny sú Winter, Spring, Summer a Autumn.
- Visibility (km): Viditeľnosť sa pohybuje od 0 km do 20 km.
- Location: Možné hodnoty sú inland, mountain a coastal.
- Irradiance: obsahuje hodnoty slnečného žiarenia od 200 do 800 W/m2.
- Weather Type: Možné typy počasia sú Rainy, Cloudy, Sunny a Snowy.