Zhodnotenie zadania č.1

Jakub Brehuv

Môj návrh riešenia by som rozdelil do troch častí. Ako prvé som si dáta pripravil, následne som klasifikátor natrénoval a ako posledný krok som vyskúšal GridSearch. Prejdem postupne vysvetlím každý jeden bod môjho riešenia.

1. Príprava dát.

Ako prvé som si dáta načítal z .npy súboru a následne rozdelil na testovaciu a trénovaciu časť. Následne som zistil či sú všetky dáta v poriadku. Ukázalo sa že to tak nie je tak bolo nevyhnutné zlé dáta nejakým spôsobom nahradiť. Prvá myšlienka bola poškodené dáta nahradiť nulou a dúfať že sa s tým klasifikátor vysporiada. Po niekoľkých pokusoch som dospel k záveru že bude lepšie ak chýbajúce alebo poškodené dáta nahradím priemerom všetkých dát na rovnakej pozícii. Ako ďalší krok som použil som dáta naškáloval a tak týmto spôsobom upravil.

2. Skúšanie algoritmov

Ako další krok som začal skúšať jednotlivé algoritmy. Začal som s najzákladnejším a to SVN, tak som vyskúšal SVR, Gausian, MLP, RandomForest a DecisionTree. A každom algoritme som vyskúšal rôzne parametre algoritmov. Neskôr som si uvedomil že som nerobil nad dátami žiadnu CrossValidáciu a vo vyriešení tohto problému mi pomohol GridSearch ktorý opíšem v ďalšej časti.

3. GridSearch

Môj veľký problem s testovanim roznych algoritmov bol taký že som mal málo času a veľa kombinácii parametrov v algoritme. V tom mi pohol GridSearch ktorý vyskúšal každú kombináciu parametrov ktorú som mu zadal a vyhodil parametre s najlepšími výsledkami. GridSearch sa taktiež postaral o CrossValidáciu takže už som to nemusel riešiť osobitne cez K-Fold. Vyskúšal som taktiež regresné algoritmy a to napríklad DecisionTreeRegressor s ktorým som dostal najlepšie výsledky po tom čo som ho vylepšil ešte AdaBoostom ktorý úspešnosť na testovacích dátach dostal ešte vyššie. Ako posledný krok pred uložením dát bolo nevyhnutné dáta zaokrúhliť na celé čísla v našom prípade na 0 a 1 pretože výsledok s DecisionTreeRegressor algoritmu boli float čísla.

Nakoniec už prišlo na rad len upratovanie kódu. Rozdelil som si funkcionalitu na jednotlivé klasy. A to na klasu ktorá sa mi stará o prácu s dátami či už načítavanie, škálovanie, opravovanie a aj rozdeľovanie. Ďalšia kalsa slúži na reprezentáciu dát a vypisovanie skóre algoritmov. Ďalšie klasy pozostávajú zo samostatných algoritmov a každý algoritmus má funkciu kde trénuje so základnými parametrami no niektoré majú použitý aj GridSearch.

Moje zhodnotenie tohto zadania je také nejasné, nezaujalo ma až to takej miery ako som si predstavoval že zaujme. Jeden z dôvodov je taký že som nevedel nič o dátach, nevedel som čo znamenajú a nevedel som čo som sa vlastne snažil dosiahnuť. Všetko bola len taká pre mňa nepriama komunikácia s dátami, a zisťovanie úspešnosti technikou pokus omyl. Viem že takáto práca je veľká časť strojového učenia, no nevzbudila vo mne podobmné vzrušenia ako zadanie č. 2.