Report - leto

Osadský: Exploratory data analysis

V rámci ročníkového projektu sme v letnom semestri spravili nasledovné:

Sumarizácia:

- Cieľom narábania s datasetom bolo trénovanie a pochopenie rozhodovacích stromov a náhodných lesov, na čo bolo samozreme nutné pochopiť linárnu regresiu. Pôvodný plán bol dostať sa v práci s datasetom až po Support vector machines (SVM), projekt sa však dostal len po ľahké trénovanie náhodných lesov. Pravdepodobne sa nám totiž podarilo natrénovať biasnutý model, alebo sme zvolili nesprávne atribúty pre trénovanie.

Postup:

1. **Pokračovanie s dátami z minulého semestra:**
   - Predpripravené dáta z minulého semestra
2. **Hľadanie vhodných atribútov na trénovanie datasetu:**
   - Pravdepodobne najtažšia časť projektu. (A zároveň najnáchylnejšia na chyby.) Naším cieľom bolo vybrať vhodné atribúty, ktoré by sme mohli v datasete analyzovať. Kedže ich bolo rádovo v stovkách (keďže sa jedná o anotovaný ľudský genóm, ktorý obsahuje mnoho dát) bolo potrebné vybrať subset atribútov, ktoré by ukázali vhodné výsledky. Po skúšaní rôznych kombinácií sme sa napokon rozhodli pre množinu atribútov takú, kde je korelácia medzi danými atribútmi aspoň 0,5 a nie 1. (1 > corr >= 0,5)
3. **Očistenie subsetu:**
   - Tento vybraný subset sme postupne iterovali a snažili sa očistiť jednotlivé výskyty od outlierov (štatistických odchýliek a chýb). Častokrát sme však vymazali príliš veľa dát zo subsetu a preto sme boli pri tolerancii odchýlky benevolentní.
4. **Študovanie lin. regresie a rozhodovacích stromov:**
   - Na správne implementovanie subsetu bolo potrebné naštudovať si lineárnu regresiu a rozhodovacie stromy, pričom sme postupovali hlavne podľa cvičení predmetu exploratívnej analýzy. Samozrejme na štúdium sme využili aj iné zdroje, ale primárne sme sa riadili podľa toho, lebo pracuje s knižnicou Scikit.
5. **Implementácia do Jupyter notebooku:**
   - Následne sme na subsete spravili lineárnu regresiu na korelujúcich atribútoch a začali implementovať rozhodujúce stromy. Problém nastal pri výsledkoch, lebo sme častokrát mali málo hodnôt jedného atribútu a tak sme pravdepodobne trénovali biasnutý model.
6. **Výsledky dátovej analýzy na subsete:**
   - Výsledky trénovania sme uložili do excel súboru, no kvôli krátkosti času sme sa nedostali k trénovaniu s odmenami ani hlbšej analýze výsledkov. Ďalší postup na projekte je využiť natrénovaný model ku SVM, no hlavne poriadne overiť, či sú dáta správne, alebo treba subset alternovať.