Project proposal

Machine learning Fall 2019

Jozef Kubík

V tomto dokumente sa pokúsim v krátkosti zhrnúť návrh projektu pre predmet Strojové učenie, pričom v krátkosti opíšem problém, dataset, štruktúru a navrhované metódy

1. Problém

Projekt by sa zaoberal klasifikáciou húb medzi 2 triedami: jedlé a jedovaté. Pri dodaní informácií o rôznych vlastnostiach húb (ale nie ich názvov) by som vytvoril algoritmus, ktorý by vedel s dobrou presnosťou predikovať, či huba, ktorej vlastnosti sú danými informáciami opísané, je vhodná na jedenie alebo nie.

Cieľom je teda vytvoriť algoritmus strojového určenia na klasifikáciu dát do dvoch rôznych tried na základe rôznych parametrov z datasetu.

V prípade možností (dátových/časových/schopnostných) by projekt obsahoval aj informácie o iných zaujímavostiach získaných zo spomenutých dát (ako napríklad ktoré vlastnosti huby najviac prezrádzajú o jej jedlosti).

2. Dataset

Dataset ako aj informácie o ňom pochádzajú zo stránky:

https://www.kaggle.com/uciml/mushroom-classification

Obsahuje viac než 8100 rôznych húb opísaných pomocou ich vlastností, medzi ktoré sa okrem jedlosti radí napríklad tvar, farba a povrch klobúku, vôňa, populácia, typ prsteňa, tvar hlúbika a iné. Dokopy obsahuje každá huba o sebe 23 vlastností, ktoré bude treba, samozrejme, upraviť do vhodnej vstupnej podoby pre spracovanie. V prípade možností by sa z vlastností mohli aj napríklad vybrať len určité z nich a na nich trénovať model, prípadne usudzovať závery.

Počet dát dáva veľkú nádej, že vytvorený model bude schopný vedieť s určitou pravdepodobnosťou zaradiť hubu do správnej kategórie.

3. Navrhované metódy

Metódy nie sú ešte garantované, ale s veľkou pravdepodobnosťou by sa v projekte objavili metódy Support Vector Machines, Random Forest a Convolutional Neural Network. Taktiež je možné využitie metódy Principal component analysis pre zredukovanie dimenzií pre trénovanie a vykresľovanie. V prípade možností by sa mohli pridať aj ďalšie.

4. Štruktúra

Projekt by okrem samotnej implementácie metód obsahoval aj iné dôležité faktory spojené so strojovým učením. Pred samotným začatím hocijakých úprav by sme sa najskôr pozreli na samotné dáta – čo o nich vieme už teraz povedať, ako približne vypadajú, či sú všetky dáta korektné. Následne by nasledovala ich úprava pre použitie. Dáta by sa rozdelili do tréningovej a testovacej množiny a pomocou rôznych metód by sme vyskúšali, akú dokážu dosiahnuť pravdepodobnosť správneho zaradenia huby. Výsledky by sme vykreslili a medzi sebou porovnali.