Извештај за предмет Проналажење скривеног знања

При реализацији алгоритама машинског учења (Линеарна регресија и К најбилижих суседа) испробао сам разне комбинације хиперпараметара као и техника у циљу побољшања крајњих резултата модела.

*Линеарна регресија*

Хиперпараметри и остали фактори који су утицали на перформанс модела су:

* Степен учења (алфа)
* Број епоха
* Подела тест-тренинг скупа
* Одабир карактеристика
* Број података на ком се тренира модел
* Регуларизација (ламбда)
* Тип градиентног спуста

За степен учења је првобитно одабрана вредност од 0.1 која се потом смањила на 0.05 како би се дошло што ближе минимуму функције коштања. Број епоха је стављен на 10000, али се у већини случајева тренирање заврши раније услед конвергирања градиентног спуста. Као најбоље решење за поделу података на скупове за тренирање и тестирање се показао сплит 95-5% у корист скупа за тренирање јер је укупан скуп података нешто већи од 40000. У разматрање сам ставио и постојање валидационог скупа података или чак и коришћење унакрсне валидације, међутим сматрао сам да то није неопходно због једноставности модела уодносу на природу проблема. Карактеристике су биране тако да од почетних 18 остане свега 7-8 најбитнијих како би се потенцијално избацили уљези који би погоршали естимацију функције лин. Регресије. Регуларизација није била потребна, односно лабда је 0, зато што модел сам по себи већ има премалу варијабилност, односно underfit-ује тако да би додатна регуларизација само погоршала перформанс модела. За градијентни спуст је коришћен класичан batchh gradient descent због не толико великог скупа података и потребног времена за тренирање када се упосле процесори + графичка картица, као и због чињенице да овај спуст сигурно иде ка конвергенцији при сваком кораку.

За евалуацију модела коришћена је R2 метрика. За конкретну имплементацију, најбољи резултат је износио 0.61.

*К најближих суседа*

Хиперпараметри и остали фактори који су утицали на перформанс модела су:

* Коришћена метрика за израчунавање дистанце
* К
* Одабир карактеристика
* Број података

При рачунању дистанце, покушао сам са више различитих метрика (Хамминг, Џакард, косинусна, еуклидска, Менхетн и Минковски). Међу овим техникама, најбоље резултате је дала Менхетн дистанца и поред своје једноставности. За параметар К је опште прихваћено да је најбоље изабрати вредност која је једнака корену укупног броја података. Поред тога је експериментисано и са другим вредностима, али се корен укупног броја података показао као најбоље решење. Одабир карактеристика је вршен као и код модела лин. Регресије. Број података је нешто већи него код лин. регресије - око 50000.

За евалуацију модела коришћена је прецизност која је за ову имплементацију износила око 65%.