Основи машинског учења, јесен 2021. домаћи задатак №4

Рок: понедељак, 22. новембар у 23:59 на Moodle-у.

Упутства: (1) Ова питања захтевају размишљање, не и дуге одговоре. Будите што сажетији. (2) Уколико има било каквих нејасноћа, питајте предметног наставника или сарадника. (3) Студенти могу радити и послати решења самостално или у паровима. У случају заједничког рада, имена и презимена оба студента морају бити назначена у Gradescope-у и није дозвољено радити са истим колегом више од једном. (4) За програмерске задатке, коришћење напредних библиотека за машинско учење попут scikit-learn није дозвољено. (5) Кашњење приликом слања односно свака пошиљка након рока носи негативне поене.

Сви студенти морају послати електронску PDF верзију својих решења. Препоручено је куцање одговора у LaTeX-у које са собом носи 10 додатних поена. Сви студенти такође морају на Moodle-у послати и zip датотеку која садржи изворни код, а коју би требало направити користећи `make_zip.py` скрипту. Обавезно (1) користити само стандардне библиотеке или оне које су већ учитане у шаблонима и (2) осигурати да се програми извршавају без грешки. Послати изворни код може бити покретан од стране аутоматског оцењивача над унапред недоступним скупом података за тестирање, али и коришћен за верификацију излаза који су дати у извештају.

Кодекс академске честитости: Иако студенти могу радити у паровима, није дозвољена сарадња на изради домаћих задатака у ширим групама. Изричито је забрањено било какво дељење одговора. Такође, копирање решења са интернета није дозвољено. Свако супротно поступање сматра се тешком повредом академске честитости и биће најстроже кажњено.

1. **[90 поена] Локално-тежинска линеарна регресија**

Размотрити проблем линеарне регресије у којој постоји жеља да се "нагласе" одређени тренинг примери. Конкретно, претпоставити да треба минимизирати функцију

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)} \left(\theta^T x^{(i)} - y^{(i)}\right)^2 \quad .$$

На предавањима је објашњено шта се дешава у случају када су сви тежински фактори $w^{(i)}$ међусобно једнаки. У овом задатку, биће уопштене неке идеје за такозвану тежинску поставку.

(a) [8 поена] **Тежинска матрица**
Показати да се $J(\theta)$ такође може написати и као

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

за одговарајућу матрицу W, где су матрица X и вектор y дефинисани на уобичајен начин. Јасно означити вредност сваког члана матрице W.

(b) [14 поена] **Нормална једначина с тежинском матрицом**
Под условом да су сви тежински фактори $w^{(i)}$ јединични, показано је на предавањима да се нормална једначина може написати у облику

$$X^T X\theta = X^T y \quad , \tag{1}$$

као и да је вредност θ за коју функција губитака $J(\theta)$ има свој глобални минимум дата изразом $(X^T X)^{-1} X^T y$. Израчунавањем извода $\nabla_\theta J(\theta)$ и изједначавањем истог са нулом, уопштити нормалну једначину за овакву тежинску поставку и дати нову вредност θ која минимизира $J(\theta)$ у затвореном облику као функцију X, W и y.

(c) [14 поена] **Нормална расподела са различитим варијансама**
Претпоставити да је дат скуп података $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ са m независних примера у којима $y^{(i)}$ прати условну расподелу вероватноће са различитим нивоима варијансе $(\sigma^{(i)})^2$. Конкретно, претпоставка се може изразити и расподелом вероватноће

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2}\right) \quad ,$$

то јест, свако $y^{(i)}$ се извлачи из нормалне, односно Гаусове, расподеле са средњом вредношћу $\theta^T x^{(i)}$ и варијансом $(\sigma^{(i)})^2$ где су $\sigma^{(i)}$ унапред познате константе. Показати да се проналажење θ методом највеће веродостојности своди на решавање задатка локално-тежинске линеарне регресије. Јасно изразити тежинске коефицијенте $w^{(i)}$ у функцији стандардне дефијације $\sigma^{(i)}$.

(d) [36 поена] **Програмерски задатак: локално-тежинска регресија**
Размотрити следећи скуп података задат у датотекама

`src/locweightreg/{train,valid,test}.csv`

Свака датотека садржи две колоне: x и y. У терминологији описаној у уводу, x је атрибут (у овом случају једнодимензионални), а y је излазна вредност. У шаблону

`src/lwr.py` имплементирати локално-тежинску линеарну регресију користећи претходно изведену уопштену нормалну једначину уз тежинске факторе дате

$$w^{(i)} = \exp\left(-\frac{||x^{(i)} - x||_2^2}{2\tau^2}\right)$$

Истренирати модел на тренинг подскупу података користећи вредност $\tau = 0.5$, а затим проверити модел на валидационом подскупу и дати средњу квадратну грешку. Приказати предвиђања модела на валидационом подскупу (означити тренинг скуп плавим 'x' маркерима, а валидациони скуп црвеним 'o' маркерима)? Да ли је модел подучен или преучен?

(e) [18 поена] **Програмерски задатак: хиперпараметар τ**

У овом подзадатку биће мењан хиперпараметар модела τ. У шаблону `src/tau.py` израчунати средњу квадратну грешку модела на валидационом подскупу за све вредности τ из скупа $\{0.03, 0.05, 0.1, 0.5, 1, 10\}$. За свако τ исцртати предвиђања модела на валидационом скупу у формату истом као у претходном подзадатку. Дати вредност параметра τ за које је средња квадратна грешка на валидационом скупу најмања и коначно дати средњу квадратну грешку на тестном подскупу користећи ову вредност τ параметра.