Основи машинског учења, јесен 2021. домаћи задатак №3

Рок: понедељак, 15. 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] Линеарна регресија: линеарна по чему?**

На предавањима је показано како се помоћу регресије линеарном функцијом могу апроксимирати подаци. У овом задатку биће приказано како се линеарна регресија може искористити за апроксимацију нелинеарних функција података користећи мапирање, односно пресликавање својстава. Такође ће бити испитана и нека ограничења која ће у будућим лекцијама бити превазиђена напреднијим техникама.

(a) [18 поена] **Учење полинома трећег степена улазних података**

Нека је дат скуп података $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ где $x^{(i)}, y^{(i)} \in \mathbb{R}$. Задатак је да се полином трећег степена $h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x^1 + \theta_0$ апроксимира скуп задати скуп података. Кључно запажање на овом месту јесте да је функција $h_\theta(x)$ и даље линеарна по непознатом параметру θ који се тражи, иако није линеарна по улазу x. Ово омогућава у наставку посматрање задатка као проблем линеарне регресије.

Нека је $\phi : \mathbb{R} \to \mathbb{R}^4$ функција која претвара изворне улазе x у четвородимензионални вектор дефинисан као

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \in \mathbb{R}^4 \tag{1}$$

Нека је $\hat{x} \in \mathbb{R}^4$ скраћени запис за $\phi(x)$, и нека је $\hat{x}^{(i)} \triangleq \phi(x^{(i)})$ трансформисани улаз из скупа тренинг података. Може се направити нови скуп података $\{(\phi(x^{(i)}), y^{(i)})\}_{i=1}^{n} = \{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ заменом изворних улаза $x^{(i)}$ са $\hat{x}^{(i)}$. Може се приметити да је апроксимација кривом $h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x^1 + \theta_0$ почетног скупа података еквивалентна апроксимацији линеарном функцијом $h_\theta(\hat{x}) = \theta_3 \hat{x}_3 + \theta_2 \hat{x}_2 + \theta_1 \hat{x}_1 + \theta_0$ новодобијеног скупа података због тога што важи

$$h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x^1 + \theta_0 = \theta_3 \phi(x)_3 + \theta_2 \phi(x)_2 + \theta_1 \phi(x)_1 + \theta_0 = \theta^T \hat{x} \tag{2}$$

Другим речима, може се искористити линеарна регресија над новим скупом података да се одреде параметри $\theta_0, \ldots, \theta_3$.

Написати 1) функцију губитака $J(\theta)$ за линеарни регресиони проблем над новим скупом података $\{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ и 2) правило за ажурирање у алгоритму потпуног (нестохастичког) градијентног спуста за линеарну регресију над скупом података $\{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$.

Терминологија: У машинском учењу, ϕ се често назива и пресликавање својстава које пресликава изворне улазе x у нов скуп променљивих. Како би се направила разлика између ова два скупа променљивих, обично се x називају улазни **атрибути**, док се $\phi(x)$ називају **особине** или **својства**. (Нажалост, различити аутори користе различите термине да опишу ове две ствари.)

(b) [18 поена] **Програмерски задатак: регресија полиномом трећег степена**

За овај подзадатак биће коришћен скуп података дат у следећим датотекама:

`src/featuremaps/{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 ознака.

Користећи формулацију из претходног подзадатка имплементирати линеарну регресију са **нормалном једначином** користећи преслигавање својстава полиномом трећег

степена. Користити шаблон који је обезбеђен у `src/featuremaps/featuremap.py` да би се имплементирао алгоритам.

Направити дијаграм расејања тренинг података и нацртати научену хипотезу као глатку криву над истим. Укључити дијаграм у извештај као решење овог задатка.

Напомена: Претпоставити да је $\widehat{X}$ матрица података трансформисаног скупа података. Понекад се може срести нерегуларна, то јест сингуларна, матрица $\widehat{X}^T\widehat{X}$. Да би се добило нумерички стабилно решење увек треба користити `np.linalg.solve` како би се параметри добили директно, уместо да се експлицитно израчунава инверзија, а затим множи са $\widehat{X}^T y$.

(c) [18 поена] **Програмерски задатак: регресија полинома k-тог степена**

Сада се горња идеја проширује на полиноме k-тог степена разматрањем $\phi : \mathbb{R} \to \mathbb{R}^{k+1}$ које је

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \in \mathbb{R}^{k+1} \tag{3}$$

Пратити исти поступак као у претходном подзадатку и имплементирати алгоритам за $k = 3, 5, 10, 20$. Направити сличне дијаграме као у претходном подзадатку и исцртати криве хипотезе за сваку вредност k користећи различиту боју. Укључити и натписе у дијаграму да се укаже која боја одговара којој вредности k

Укључити дијаграм у извештај као решење овог подзадатка. Посматрати како се апроксимација тренинг скупа мења када се k повећава. Укратко прокоментарисати запажања.

(d) [18 поена] **Програмерски задатак: друга пресликавања својстава**

Уочено је да је неопходан релативно висок степен k како би се дати тренинг скуп података апроксимирао, а разлог томе је што се овај скуп података не може врло добро објаснити (то јест апроксимирати) полиномима ниског степена. Визуализацијом података, може се закључити да се y може добро апроксимирати простопериодичном (синусоидалном) функцијом. Заправо, подаци су генерисани одабирањем функције $y = \sin(x) + \xi$, где је ξ шум са Гаусовом расподелом. Допунити пресликавање својстава ϕ тако да укључује и синусну трансформацију као у наставку:

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \\ \sin(x) \end{bmatrix} \in \mathbb{R}^{k+2} \tag{4}$$

Са овако допуњеним пресликавањем својстава истренирати различите моделе за вредности $k = 0, 1, 2, 3, 5, 10, 20$ и исцртати резултујуће криве хипотезе преко података као и малочас.

Укључити дијаграм као решење овог подзадатка. Упоредити апроксимационе моделе са претходним подзадатком и укратко прокоментарисати приметне разлике у апроксимацији која користи ово пресликавање својстава.

(е) [18 поена] **Преучење изражајних модела над малобројним подацима**

За преостали део овог задатка биће размотрен мали скуп података (случајан подскуп скупа података који је до сада коришћен) са далеко мањим бројем примера који је дат у следећој датотеци:

`src/featuremaps/small.csv`

Биће истражено шта се дешава када број својстава буде већи од броја примера у тренинг скупу. Покренути алгоритам на овом малобројном скупу података користећи следеће пресликавање

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \in \mathbb{R}^{k+1} \tag{5}$$

са $k = 1, 2, 5, 10, 20$.

Направити дијаграм у коме је свака од хипотеза представљена различитом кривом, као и у претходним подзадацима. Запазити како се апроксимација над тренинг скупом података мења с порастом k. Укључити дијаграме у извештај и укратко прокоментарисати запажања.

Напомена: Феномен који се запажа где модел најпре добро апроксимира тренинг скуп података, а затим одједном "подивља" је услед ефекта који се назива и *преучење*. Интуиција коју треба развити јесте да када је количина података за тренирање релативно мала у односу на изражајну моћ породице могућих модела (то јест, класе хипотеза која је у овом конкретном случају случај породица свих полинома степена k) долази до преучења.

Грубо говорећи, скуп хипотеза је "врло флексибилан" и може се лако натерати да прође кроз све примере, односно тачке на прилично неприродан начин. Другим речима, модел покушава да предвиди чак и шум у тренинг скупу података који не би уопште требало да се предвиђа. Ово по правилу шкоди предвиђању модела на тест примерима. Феномен преучења ће бити детаљније разматран на предавањима када се буде обрађивала теорија учења и компромис између помераја и варијансе.