Основи машинског учења, јесен 2021. домаћи задатак №2

Рок: понедељак, 8. 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 извештају који се шаље 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] Поасонова регресија**

У овом задатку биће обрађена још једна врста често коришћених уопштених линеарних модела (УЛМ) под називом Поасонова регресија. У уопштеном линеарном моделу, избор неке расподеле из експоненцијалне породице заснован је на врсти задатка који се решава. То јест, ако се решава класификациони задатак, користи се расподела из експоненцијалне породице с подршком дискретним класама (као што је рецимо Бернулијева или категоријска расподела). Слично томе, ако је излаз реална вредност, користи се Гаусова или Лапласова расподела (обе такође из експоненцијалне породице). Некада је жељени излаз предвиђање бројања, на пример, предвиђање броја писама која се очекују да ће бити примљена у току дана или броја посетилаца неке интернет странице, односно броја муштерија која се очекују да уђу у радњу у наредном сату, и слично. Расподела густине вероватноће која је природна за овакве врсте задатака, то јест подржава природне бројеве, односно бројање, јесте Поасонова расподела која такође припада експоненцијалној породици.

У подзадацима који следе, најпре ће бити показано да Поасонова расподела припада тзв. експоненцијалној породици, а затим ће бити изведен функционални облик хипотезе као и начин за тренирање модела, и коначно над датим скупом тренинг података биће практично истрениран модел који ће се користити за предвиђања на задатом тест скупу.

(a) [18 поена] Размотрити Поасонову расподелу параметризовану са λ:

$$p(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

(Овде је y природан број, а $y!$ представља факторијел y.) Показати да Поасонова расподела припада експоненцијалној породици и јасно изразити вредности за $b(y)$, η, $T(y)$, и $a(\eta)$.

(b) [11 поена] Размотрити извођење регресије користећи се уопштеним линеарним моделом са Поасоновом излазном променљивом. Која је каноничка излазна функција породице? (Може се искористити чињеница да Поасонова случајна променљива са параметром λ има средњу вредност λ.)

(c) [25 поена] За тренинг скуп $\{(x^{(i)}, y^{(i)});\ i = 1, \ldots, n\}$ нека је логаритамска веродостојност за један пример $\log p(y^{(i)}|x^{(i)}; \theta)$. Преко првог извода логаритамске веродостојности по θ_j, извести правило ажурирања стохастичког градијентном успона за учење користећи уопштени линеарни модел са Поасоновим излазима y и каноничку одзивну функцију.

(d) [36 поена] **Програмерски задатак**

Размотрити интернет страницу која жели да предвиди дневни саобраћај. Власници интернет странице су сакупили скуп података саобраћаја ка својој интернет страници из прошлости и уз то још нека својства за која су мислили да могу бити корисна у предвиђању дневног броја посетилаца. Скуп података је подељен на тренинг скуп и валидациони скуп и шаблон изворног кода је дат у следећим датотекама:

- `src/poisson/{train,valid}.csv`
- `src/poisson/poisson.py`

Биће примењена Поасонова регресија за моделовање дневног броја посетилаца. Треба напоменути да сама примена Поасонове регресије унапред претпоставља да подаци прате Поасонову расподелу чији је природни параметар линеарна комбинација улазних атрибута (то јест, $\eta = \theta^T x$). У `src/poisson/poisson.py`, имплементирати Поасонову регресију за дати скуп података и користити *потпун (нестохастички) градијентни*

успон за максимизацију логаритамске веродостојности параметра θ. Као критеријум за заустављање проверити да ли промена параметара има норму која је мања од вредности 10^{-5}.

Користећи истрениран модел предвидети очекиван број на **валидационом скупу** и направити дијаграм расејања између тачног броја и предвиђеног броја посета (на валидационом скупу). На дијаграму расејања нека на апциси буду тачне вредности за број посета, а а на ординати одговарајућа предвиђања очекиваног броја посета. Приметити да су тачне вредности целобројне, док очекиване вредности су у општем случају реалне.