Основи машинског учења, јесен 2021. домаћи задатак №1

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] Линеарни класификатори (логистичка регресија и ГДА)**

У овом задатку, биће покривена два линеарна класификатора која су до сада обрађена на предавањима. Први, дискриминативни линеарни класификатор: логистичка регресија. Други, генеративни линеарни класификатор: Гаусова дискриминантна анализа (ГДА). Оба алгоритма проналазе линеарну границу одлуке која раздваја податке на две класе, али уз различите претпоставке. Циљ овог домаћег задатка јесте да се стекне дубље разумевање о сличностима и разликама (као и о предностима и манама) ова два алгоритма.

У склопу овог задатка, биће размотрена два скупа података, уз шаблоне изворних кодова који су дати у следећим датотекама:

- `src/linearclass/ds1_{train,valid}.csv`
- `src/linearclass/ds2_{train,valid}.csv`
- `src/linearclass/logreg.py`
- `src/linearclass/gda.py`

Свака датотека садржи n примера, један пример $(x^{(i)}, y^{(i)})$ по реду. Нарочито, i-ти ред садржи колоне $x_1^{(i)} \in \mathbb{R}$, $x_2^{(i)} \in \mathbb{R}$, и $y^{(i)} \in \{0, 1\}$. У подзадацима који следе, биће испитано коришћење логистичке регресије и Гаусове дискриминантне анализе (ГДА) како би се извршила двојна (бинарна) класификација на ова два скупа података.

(a) [20 поена]
На предавањима је приказана функција губитака за логистичку регресију:

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right),$$

где је $y^{(i)} \in \{0, 1\}$, $h_\theta(x) = g(\theta^T x)$ и $g(z) = 1/(1+e^{-z})$.
Пронаћи Хесијан H ове функције и показати да за произвољни вектор z важи

$$z^T H z \geq 0.$$

Смерница: Може се најпре показати да је $\sum_i \sum_j z_i x_i x_j z_j = (x^T z)^2 \geq 0$. Подсетити се такође да је $g'(z) = g(z)(1-g(z))$.

Напомена: Ово је један од уобичајених начина да се покаже да је матрица H позитивно семидефинитна, што се означава са "$H \succeq 0$." Ово даље имплицира да је J конвексна, односно да нема других локалних минимума изузев глобалног. Није неопходно користити горњу смерницу како би се показало да је $H \succeq 0$ већ било коју.

(b) [10 поена] **Програмерски задатак.** Пратити упутства дата у `src/linearclass/logreg.py` да се истренира класификатор заснован на логистичкој регресији користећи се Њутновом методом. Почевши од $\theta = \vec{0}$, извршавати Њутнову методу све док померај по θ не постане мали: Конкретно, тренирати до прве итерације k за коју важи $\|\theta_k - \theta_{k-1}\|_1 < \epsilon$, где је $\epsilon = 1 \times 10^{-5}$. Обавезно уписати вероватноће предвиђања на валидационом скупу у датотеку која је дата у изворном коду.

Укључити график **валидационих података** са x_1 на хоризонталној оси и x_2 на вертикалној оси. За представљање две класе користити различите маркере (симболе) за примере $x^{(i)}$ за које је $y^{(i)} = 0$ у односу на оне за које је $y^{(i)} = 1$. На истом графику исцртати границу одлуке коју проналази логистичка регресија (тј. праву која одговара $p(y|x) = 0.5$).

(c) [10 поена] Подсетити се да је у ГДА заједничка расподела (x, y) описана следећим једначинама:

$$p(y) = \begin{cases} \phi & \text{if } y = 1 \\ 1 - \phi & \text{if } y = 0 \end{cases}$$
$$p(x|y = 0) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right)$$
$$p(x|y = 1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right),$$

где су ϕ, μ_0, μ_1, и Σ параметри модела.

Претпоставимо да су ϕ, μ_0, μ_1, и Σ већ одређени и да је даље неопходно предвидети y за нову задату тачку x. Како би се доказало да ГДА као резултат даје класификатор са линеарном границом одлуке, показати да се апостериорна вероватноћа може написати као

$$p(y = 1 \mid x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-(\theta^T x + \theta_0))},$$

где су $\theta \in \mathbb{R}^d$ и $\theta_0 \in \mathbb{R}$ одговарајуће функције параметара ϕ, Σ, μ_0, и μ_1.

(d) [15 поена] За задати скуп података, тврди се да су на основу методе највеће веродостојности (МНВ) параметри дати као

$$\phi = \frac{1}{n} \sum_{i=1}^{n} 1\{y^{(i)} = 1\}$$
$$\mu_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$
$$\mu_1 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$
$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

Логаритамска функција веродостојности података је

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi). \end{aligned}$$

Максимизацијом ℓ по четири параметра, доказати да су процене ϕ, μ_0, μ_1, и Σ методом највеће веродостојности заиста онакве као у горњим једнакостима. (Може се претпоставити да постоји бар један позитиван и макар један негативан пример тако да су имениоци у дефиницијама за μ_0 и μ_1 различити од нуле.)

(e) [10 поена] **Програмерски задатак.** У датотеци `src/linearclass/gda.py` допунити изворни код тако да израчунава ϕ, μ_0, μ_1, и Σ, затим искористити ове параметре да се добије θ, и коначно употребити тако добијени ГДА модел за предвиђања на валидационом скупу података. Обавезно уписати вероватноће предвиђања на валидационом скупу у датотеку која је дата у изворном коду.

Укључити график **валидационих података** са x_1 на хоризонталној оси и x_2 на вертикалној оси. За представљање две класе користити различите маркере (симболе) за примере $x^{(i)}$ за које је $y^{(i)} = 0$ у односу на оне за које је $y^{(i)} = 1$. На истом графику исцртати границу одлуке коју проналази ГДА (тј. праву која одговара $p(y|x) = 0.5$).

(f) [5 поена] За први скуп података (`ds1_valid`) упоредити графике добијене из логистичке регресије и ГДА из претходних подзадатака и укратко у пар редова прокоментарисати запажања.

(g) [10 поена] Поновити програмерске подзадатке за други скуп података. Направити сличне графике на **валидационом скупу** и укључити их у одговор.

На ком од два скупа података ГДА ради лошије од логистичке регресије? Шта може бити узрок томе?

(h) [10 поена] За скуп података на ком ГДА ради лошије, испитати да ли је могуће пронаћи трансформацију улазних података $x^{(i)}$ такву да ГДА ради знатно боље? Која би то трансформација могла бити?