Основи машинског учења, јесен 2021. домаћи задатак №6 решења

ИМЕ И ПРЕЗИМЕ (БРОЈ ИНДЕКСА)

Рок: среда, 29. децембар 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] Анализа независних компоненти**

На предавања током којих је обрађивана анализа независних компоненти (АНК), такође позната и као метода независних компоненти (МНК), дато је неформално објашњење зашто није могуће применити представљене технике на изворе који прате Гаусову, односно нормалну расподелу. Такође је речено да за било коју другу расподела (изузев нормалне) поменуте технике раде па је искоришћена логистичка расподела. У овом домаћем задатку биће дубље истражено зашто извори који прате Гаусову расподелу представљају проблем. Такође ће бити изведена анализа независних компоненти са Лапласовом расподелом и примењена на такозвани проблем коктел журке.

У кратком подсетнику, нека су $s \in \mathbb{R}^d$ изворни подаци који су произведени од стране d независних извора. Нека су $x \in \mathbb{R}^d$ примљени подаци, такви да $x = As$ где се $A \in \mathbb{R}^{d\times d}$ назива *матрица мешања*. Под претпоставком да је A инвертибилна, односно несингуларна матрица, тада се $W = A^{-1}$ назива *матрица размешавања*. Стога важи $s = Wx$. Циљ анализе независних компоненти је да пронађе W. Нека се са w_j^T означи j-та врста матрице W. Приметити да ово имплицира да се j-ти извор може реконструисати помоћу w_j и x, пошто је $s_j = w_j^T x$. Дат је тренинг скуп $\{x^{(1)}, \ldots, x^{(n)}\}$ који ће бити коришћен у наредним подзадацима. Означимо целокупан тренинг скуп дизајн матрицом $X \in \mathbb{R}^{n\times d}$ у којој сваки пример одговара једној врсти матрице.

(a) [20 поена] **Гаусови извори**

За овај подзадатак претпоставимо да извори прате стандардну нормалну расподелу, то јест $s_j \sim \mathcal{N}(0,1), j = \{1,\ldots,d\}$. Логаритамска веродостојност матрице размешавања је дата са:

$$\ell(W) = \sum_{i=1}^{n}\left(\log|W| + \sum_{j=1}^{d}\log g'(w_j^T x^{(i)})\right),$$

где је g кумулативна функција расподеле, а g' функција расподеле вероватноће (у овом подзадатку стандардна нормална расподела). Док је на предавањима изведено правило ажурирања како би се до W дошло итеративним путем, за Гаусове изворе могуће је аналитички доћи до резултујуће матрице W.

Извести израз у затвореној аналитичкој форми за W у зависности од X када је g стандардна нормална кумулативна функција расподеле. Закључити у ком су односу W и X у најједноставнијим цртама и јасно истаћи нејасноће (у погледу ротационе симетрије) при израчунавању W матрице.

Одговор: Знамо да је: $p_x(x^i) = p_s(Wx^i)|W|$

Максимизацијом функције по W добијамо:

$l(W) = \sum_{i=1}^{n} log(p_x(x^i)) =$

$= \sum_{i=1}^{n} log(p_s(Wx^i)|W|) =$

$= nlog|W| + \sum_{i=1}^{n} log(\frac{1}{(2\pi)^{(\frac{d}{2})}} exp(-(\frac{1}{2})(Wx^i)^T(Wx^i))) =$

$= nlog|W| + \sum_{i=1}^{n} log(\frac{1}{(2\pi)^{(\frac{d}{2})}} exp(-\frac{1}{2}(x^i)^T W^T W x^i))$

$\triangledown l(W) =$

$= \nabla(nlog|W|) + \nabla(\sum_{i=1}^{n}(-\frac{1}{2}(x^i)^T W^T W x^i))) =$

$= nW^{-T} - \frac{1}{2}\sum_{i=1}^{n}(\nabla((x^i)^T W^T W x^i))) =$

$= nW^{-T} - W(\sum_{i=1}^{n}(x^i)^T x^i) =$

$= nW^{-T} - WX^TX$

потом резултат градијента по W изједначимо са 0: $nW^{-T} - WX^TX = 0$
$\rightarrow W^{-T}W = (\frac{1}{n}X^TX)^{-1}$
Уочавамо да је $(\frac{1}{n}X^TX)^{-1}$ позитивна семи-дефинитивна матрица.

(b) [40 поена] **Лапласови извори.**

У овом подзадатку претпоставка је да извори прате стандардну Лапласову расподелу, то јест $s_i \sim \mathcal{L}(0,1)$. Лапласова расподела $\mathcal{L}(0,1)$ описана је функцијом расподеле вероватноће $f_{\mathcal{L}}(s) = \frac{1}{2}\exp(-|s|)$. Уз ову претпоставку, извести правило ажурирања за један пример у облику

$$W := W + \alpha(\ldots).$$

Одговор: Знамо да је: $l(W) = \sum_{n}^{i=1}(log|W| + \sum_{d}^{j=1} log(g'(w_j^T x^i)))$
односно: $l_i(W) = (log|W| + \sum_{d}^{j=1} log(p_s(w_j^T x^i)))$
Такође је из поставке познато да је: $p_s(w_j^T x^i) = \frac{1}{2}exp(-|s|)$ то замењујемо у горе наведену функцију.
$l_i(W) = log|W| + \sum_{d}^{j=1} log(\frac{1}{2}exp(-|s|)))$
$= log|W| + \sum_{d}^{j=1} log(\frac{1}{2}exp(-|w_j^T x^i|))) =$
$= log|W| - dlog(2) - \sum_{d}^{j=1}(-|w_j^T x^i|) =$
$= log|W| - \sum_{d}^{j=1}(-|w_j^T x^i|)$
За следећи градијент потребно је знати $\frac{d|x|}{dx} = sign(x)$

$\rightarrow \nabla l(W) = W^{-T} - \sum_{j=1}^{d} \nabla(-|w_j^T x^i|)$
$= W^{-T} - \begin{bmatrix} sign(w_1^T x^i|) \\ sign(w_2^T x^i|) \\ \ldots \\ sign(w_d^T x^i|) \end{bmatrix} x^{(i)T}$
Ово унисмо у правило ажурирања:

$$W := W + \alpha(W^{-T} - \begin{bmatrix} sign(w_1^T x^i|) \\ sign(w_2^T x^i|) \\ \ldots \\ sign(w_d^T x^i|) \end{bmatrix} x^{(i)T}).$$

(c) [30 поена] **[Програмерски подзадатак] Проблем коктел журке**

У овом подзадатку биће имплементиран АНК алгоритам претпостављајући Лапласове изворе (као што је изведено у претходном подзадатку) уместо Логистичких извора који су обрађени на предавањима. Датотека `src/ica/mix.dat` садржи улазне податке који се састоје из матрице са пет колона, где свака колона одговара једном измешаном сигналу x_i. Шаблон изворног кода за овај подзадатак налази се у `src/ica/ica.py` где треба допунити `update_W` и `unmix` функције.

Након тога може се покренути `ica.py` да се измешани аудио запис раздвоји на компоненте. Измешани аудио записи ће бити снимљени у `mixed_i.wav` у излазном директоријуму. Раздвојени аудио записи ће бити снимљени у `split_i.wav` у излазном директоријуму.

Како би се уверили да је решење исправно, преслушати раздвојене аудио записе. (Неко преклапање или шум у изворима може бити присутан, али различити извори би требало да буду јасно раздвојени.)

Послати матрицу размешавања W (5×5) која је добијена тако што ће и датотека `W.txt` са овим вредностима бити укључена поред изворног кода.

У исправној имплементацији, излаз `split_0.wav` би требало да звучи јако слично запису у `correct_split_0.wav` који је укључен ради провере.

Напомена: Стопа учења α може се постепено смањивати како би се убрзало тренирање, односно учење. Поред променљиве стопе учења која може убрзати конвергенцију, могуће је такође изабрати случајну пермутацију тренинг података и покренути стохастички градијентни успон по том редоследу над њима.

Одговор: $W =$

$$= \begin{bmatrix} 52.83492974 & 16.79598806 & 19.9411949 & -10.19841036 & -20.8977174 \\ -9.9368057 & -0.97879563 & -4.68186342 & 8.0430365 & 1.79099473 \\ 8.31143332 & -7.47699382 & 19.31554724 & 15.17460858 & -14.32640472 \\ -14.66729873 & -26.64481368 & 2.44071692 & 21.38223128 & -8.42094492 \\ -0.26917605 & 18.37373974 & 9.31200636 & 9.10275731 & 30.59390495 \end{bmatrix}$$