Курс «Машинне навчання»

Домашнє завдання 3: «Навчання без учителя»

Як здавати роботу

Питання домашньої роботи вимагають певного обмірковування, але не вимагають довгих відповідей. Будь ласка, будьте якомога більш стислі.

1. Якщо ви маєте будь-які питання щодо цієї домашньої роботи, задавайте їх на Piazza.
2. Ви можете обговорювати домашні завдання в групах, але не показуйте іншим свої рішення і не користуйтесь готовими чужими.
3. Для теоретичних задач, можна надсилати або скановані рукописні відповіді, або підготувати електронні версії в Word чи LaTeX. Зберігайте ці звіти в форматі PDF.
4. Для задач, які вимагають написання програм, надсилайте ваш код (з коментарями) та графіки, якщо їх потрібно намалювати відповідно до умов задачі.
5. Вкажіть ваше ім'я та прізвище у звіті.
6. Потрібно здати: PDF-звіт із теоретичними завданнями (якщо такі є), код програмних завдань (якщо такі є). Ці файли мають бути здані через Moodle.
7. Будь ласка, не здавайте датасети, якщо вони не були модифіковані. Вони займають надто багато місця, надто великі файли можуть бути відхилені Moodle.

Технічні примітки

1. Для завдань з програмування використовуйте Python 3.5. Можете користуватися або офіційним дистрибутивом, або дистрибутивом Anaconda, що вже містить більшість заздалегідь встановлених пакетів.
2. Встановіть бібліотеки, вказані у `requirements.txt` (можливо, знадобляться права адміністратора):

```
pip install -r requirements.txt
```

1. Упереджені викладачі.

[30 балів]

Група студентів, що вивчає курс машинного навчання, в кінці навчання здала P курсових проектів. Курс веде T викладачів. Проекти оцінюються всіми викладачами колективно – кожен з них ставить свою оцінку $x^{(pt)}$.

Ми припускаємо, що кожен проект заслуговує на певну «істину» оцінку μ_p. Кожен викладач, читаючи фінальний звіт, намагається «вгадати» ці істину оцінку. Таким чином, $x^{(pt)}$ – «здогадка» викладача t про те, яким є справжнє значення μ_p.

Проте, викладачі – люди, і людський фактор має свій вплив на оцінку. Дехто з них вважає, що всі проекти хороші, і ставить всім високі бали. Інші можуть бути надто критичними і ставити загалом низькі бали. Також, оцінки разних викладачів можуть мати різну дисперсію, що робить одних більш надійними за інших.

Позначимо v_t упередження викладача t. Іншими словами, викладач t в середньому оцінює роботи студентва на v_t балів вище, ніж мав би.

На процес оцінювання робіт впливає величезна кількість випадкових факторів, тому ми будемо моделювати його таким чином.

$$y^{(pt)} \sim N\left(\mu_p, \sigma_p^2\right)$$

$$z^{(pt)} \sim N(v_t, \tau_t^2)$$

$$x^{(pt)} | y^{(pt)}, z^{(pt)} \sim N\left(y^{(pt)} + z^{(pt)}, \sigma^2\right)$$

Змінні $y^{(pt)}$ і $z^{(pt)}$ – незалежні. Змінні x, y, z для різних пар проект-викладач також є незалежними.

Маючи лише виставлені оцінки ($x^{(pt)}$), ми хочемо з'ясувати параметри μ_p, σ_p^2, v_t і τ_t^2. Тоді ми можемо вважати значення μ_p «істиною» кількістю балів, на яку заслуговує курсовий проект.

Ми можемо визначити ці параметри, максимізувавши інтегровану правдоподібність $\left\{x^{(pt)}; p = 1, .., P; t = 1, .., T\right\}$. Таким чином, модель має латентні змінні $y^{(pt)}$ і $z^{(pt)}$, і проблема максимізації правдоподібності не може бути вирішена у явному вигляді. Тому ми використаємо ітеративний підхід і EM алгоритм. Ваша задача – визначити кроки E і M для цієї моделі.

Ваше рішення для E і M кроків має використовувати тільки такі операції.

Над скалярами: додавання, віднімання, множення, ділення, експонента, логарифм, корінь.

Над векторами і матрицями: додавання, віднімання, множення, інвертування, детермінант.

З метою спрошення задачі, нехай невідомими будуть тільки $\{\mu_p, \sigma_p^2; p = 1, \ldots, P\}$ і $\{v_t, \tau_t^2; t = 1, \ldots, T\}$. Будемо вважати σ^2 відомою константою.

a. **[10 балів]** Крок E алгоритму.
   Спільний розподіл імовірності $p\left(y^{(pt)}, z^{(pt)}, x^{(pt)}\right)$ має форму спільного багатовимірного нормального розподілу. Виразіть середнє значення та матрицю коваріації цього розподілу через змінні $\mu_p, \sigma_p^2, v_t, \tau_t^2$ і σ^2. Зверніть увагу, що $x^{(pt)}$ може будти представлений, як $x^{(pt)} = y^{(pt)} + z^{(pt)} + \epsilon$, де $\epsilon \sim N(0, \sigma^2)$.
b. **[10 балів]** Крок E алгоритму.
   Виразіть $Q_{pt}\left(y^{(pt)}, z^{(pt)}\right) = p\left(y^{(pt)}, z^{(pt)} | x^{(pt)}\right)$, використовуючи правило залежності від підмножин спільного багатовимірного нормального розподілу.
c. **[10 балів]** Крок M алгоритму.
   Сформулюйте крок M алгоритму для оновлення змінних $\mu_p, \sigma_p^2, v_t, \tau_t^2$. Ви можете це зробити через нижню межу правдоподібності з застосуванням математичного очікування $\left(y^{(pt)}, z^{(pt)}\right)$, взятого з розподілу з густиною $Q_{pt}\left(y^{(pt)}, z^{(pt)}\right)$.

2. Кластеризація в неевклідовому просторі.

[25 балів]

Ви маєте m точок, які описують об'єкти n мультиноміальними ознаками. Тобто, за умови $n = 5$, $m^{(1)}$ та $m^{(2)}$ можуть мати вигляд:

$$m^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 6 \\ 8 \\ 2 \end{bmatrix}, m^{(2)} = \begin{bmatrix} 1 \\ 3 \\ 80 \\ 8 \\ 3 \end{bmatrix}$$

Кожне зі значень ознак кодує таку ознаку, як, наприклад, назва району або діагноз. Таким чином, порівнювати дві точки за цими ознаками ми можемо

лише на рівні «однакові» і «не однакові». Різниця цих значень не має ніякої інтерпретації в реальному світі. Значення другої ознаки з прикладу вище «2» і «3» є не однакові настільки ж, наскільки не однаковими є значення третьої ознаки «6» і «80».

Вам потрібно згрупувати точки за схожістю в k кластерів. Стандартний k-means не підходить для цієї задачі через те, що точки з такими мультиноміальними ознаками знаходяться в неевклідовому просторі.

Ваша задача – модифікувати k-means так, щоб він працював для цього простору.

Навчальна вибірка, з якою ви працюєте, є надто великою, щоб вміститися в пам'ять, тому ваш алгоритм повинен навчатися за допомогою порцій даних (mini-batches).

a. **[5 балів]** Сформулюйте правило ініціалізації центроїдів та їх оновлення для мультиноміальних ознак.
b. **[10 балів]** Для k-means в евклідовому просторі ми використовуємо функцію найменших квадратів (least squares) як функцію відхилення (cost function):

$$J(c, \mu) = \sum_{i=1}^{m} \left\| x^{(i)} - \mu_c^{(i)} \right\|^2$$

Сформулюйте аналогічну функцію відхилення для вашого алгоритму, яка працюватиме в просторі точок з мультиноміальними ознаками. Поясніть, чому вона має сенс, порівнюючи з least squares для стандартного k-means. Покажіть умови її збіжності.

c. **[10 балів]** При оновленні центроїдів порціями (mini-batch) може статися так, що найбільш часте значення (мода) певної ознаки не відповідає дійсному найбільш частому значенню у всьому кластері, що формується. Скажімо, найбільш частий діагноз в певній групі пацієнтів може бути «цукровий діабет», але в якійсь випадково вибраній з такої групи підгрупі пацієнтів, яку ви використовуєте для оновлення центроїду на поточному кроці, найбільш часто буде зустрічатися діагноз «інсульт». Аналогічно, якщо ви кинете монетку три рази, може тричі випасти «орел». Але це не значить, що нам обовязково потрібно оновлювати значення центроїду на «інсульт», або стверджувати, що ця монетка завжди падає «орлом» догори.

Маючи порцію даних (mini-batch) розміру s, сформулюйте такі умови оновлення центроїдів, щоб значення ознаки змінювалось лише тоді, коли є певність, що воно відповідає дійсній моді цієї ознаки в кластері з імовірністю щонайменше p.

Вам може знадобитися нерівність Хофдінга з лекції 11:

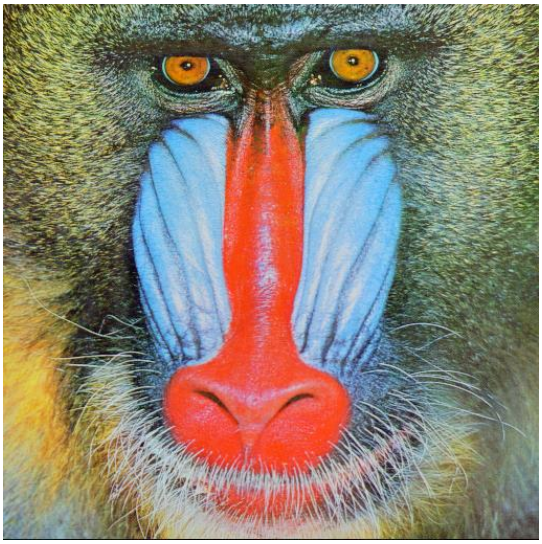
$$P\left(\left|\phi - \widehat{\phi}\right| > \gamma\right) \leq 2\exp(-2\gamma^2 m)$$

3. Стискання зображень.

[20 балів]

У цій задачі ми використаємо k-means кластеризацію для стискання зображення шляхом зменшення кількості кольорів.

Вхідне зображення складається з 512×512 пікселів, кожен із яких заданий 24-бітним кольором (по 8 бітів на кожну з RGB-компонент). Якщо це зображення зберігати попіксельно, воно займатиме $512 \times 512 \times 3 = 786\,432$ байти.



Якщо ми зможемо зменшити кількість кольорів із $\left(2^8\right)^3$ до 16, це дозволить використати лише $\log_2 16 = 4$ біти для кодування кольору замість 24, в результаті стискаючи зображення у 6 разів.

Проблема полягає лише в тому, як обрати кольори так, щоб втратити мінімум інформації. Для цього ми й застосуємо кластеризацію. Ми знайдемо 16 груп подібних кольорів, і замінимо кожен 24-бітний колір на центроїд відповідної для нього групи.

Заповніть пропущені місця в записнику **Problem Set 3 – Image compression.ipynb**:

a. **[5 балів]** Знайдіть номер найближчого центроїда для кожної точки.

b. **[5 балів]** Змістіть координати центроїдів, знаючи, до якого центроїда найближча кожна точка.

c. **[5 балів]** Обчисліть цільову функцію K-Means.

d. **[5 балів]** Імплементуйте цикл k-means, який виконується до збіжності.

Запустіть K-Means до збіжності і порівняйте зображення до стиснення і після.

4. Визначення аномального навантаження на сервери.

[25 балів]

Ви зібрали логи роботи віртуальних машин у датацентрі, які знімалися протягом певного часу з певною періодичністю. Логи складаються з записів навантаження на CPU та об'єму використаної оперативної пам'яті (RAM).

У цьому завданні ви визначите незвично малі/великі навантаження за допомогою Gaussian mixture model. Для вивчення гаусіан ви реалізуєте алгоритм Expectation-Maximization (EM).

При правильній реалізації алгоритму, функція log-likelihood буде зростати з кожною EM-ітерацією і дійде до збіжності.

Заповніть пропущені місця в записнику **Problem Set 3 – Datacenter anomaly detection.ipynb**:

a. **[5 балів]** Реалізуйте функцію густини багатовимірного нормального розподілу.

b. **[5 балів]** Реалізуйте крок E алгоритму EM.

c. **[10 балів]** Реалізуйте крок M алгоритму EM.

d. **[5 балів]** Реалізуйте розрахунок логарифму правдопобідності (log likelihood) Gaussian mixture model.

Запустіть EM до збіжності.

5. Розділення звукових доріжок.

[15 балів]

Розділення звукових доріжок – частковий випадок проблеми вечірки. До цієї ж проблеми можна звести, наприклад, розділення сигналів мозку при роботі з BCI або опрацювання радіосигналів, згенерованих різними джерелами чи відбитих від різних об'єктів. Проте, в рамках цього курсу ми будемо працювати зі звуком задля простоти оцінювання якості роботи алгоритму.

Змішані записи п'яти звукових доріжок ви можете прослухати в папці **mixed**.

У записнику **Problem Set 3 – Unmixing signals.ipynb** уже підготовлений загальний каркас коду, реалізована підготовка даних та зберігання у вигляді звукових файлів.

a. **[10 балів]** Імплементуйте вивчення матриці розділення W з даних.
b. **[5 балів]** Розділіть звукові доріжки, використовуючи вивчену матрицю розділення W.

6. Eigenfaces

[35 балів]

Eigenfaces – один з перших алгоритмів, які вивчають ознаки напряму з даних (у нашому випадку – зображень). Автори оригінальної статті застосували його до зображень обличь людей і показали, що вивчені таким чином ознаки показують високу точність на задачі розпізнавання обличь. До певної міри ми можемо казати, що eigenfaces був застосуванням філософії deep learning до задачі розпізнавання обличь задовго до ери самого deep learning (алгоритм було представлено в 1991 році).

В оригінальній статті eigenfaces було застосовано до чорно-білих зображень розміром 256×256 пікселів. В цій задачі ви застосуєте його до кольорових зображень обличь розміром 100×100 пікселів.

Зображення, які ви будете використовувати для цієї задачі, ви можете переглянути в папці **images**.

У записнику **Problem Set 3 – Eigenfaces.ipynb** уже підготовлений загальний каркас коду, реалізоване завантаження та попередня обробка даних. Вам також знадобиться ознайомитися з оригінальною статтею за посиланням вище.

a. **[5 балів]** Представте дані у вигляді матриці для SVD декомпозиції.
b. **[5 балів]** Обрахуйте eigenfaces, застовувавши SVD декомпозицію до даних.
c. **[10 балів]** Візуалізуйте 10 eigenfaces, які відповідають найбільшим власним значенням (eigenvalues) – іншими словами, є найбільш визначними ознаками, що відрізняють обличчя людей на фото від середнього обличчя. Математично, зображення eigenfaces є задачею проекції між різними просторами.
d. **[10 балів]** Обрахуйте 10 eigenfaces ітеративно, максимізуючи дисперсію. Порівняйте час їх обрахунку шляхом SVD декомпозиції та ітеративним методом.
e. **[5 балів]** Візуалізуйте 10 eigenfaces, отримані методом максимізації дисперсії, та порівняйте їх з eigenfaces, отриманих за допомогою SVD.