

Курс «Машинне навчання»

Домашнє завдання 2: «Навчання з учителем»

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 користуйтеся 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.

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
```

Імовірно, що XGBoost вам доведеться встановлювати [окремо](#).
3. Налаштуйте Keras таким чином, щоб він [використовував обчислення на TensorFlow](#). Переконайтеся що ви можете виконати «import keras» та побачити повідомлення «Using TensorFlow backend».

1. Створення кернелів.

[35 балів]

Під час лекцій ми говорили, що кернел у форматі $K(x, z) = \phi(x)^\top \phi(z)$ може проектувати дані в простір із більшим числом вимірів для алгоритму SVM. Один із способів створити кернел — це сформулювати проекцію в простір із більшим числом вимірів ϕ і вивести на його основі K .

Але у цій задачі ми будемо конструювати кернели напряму: ми маємо якусь функцію $K(x, z)$, яка, на нашу думку, дає хороший спосіб виміряти схожість прикладів для нашої задачі, і збираємося вставити її в SVM як кернел-функцію. Проте, щоб бути правильним кернелом, $K(x, z)$ має бути внутрішнім добутком двох векторів у якомусь просторі з більшим числом вимірів, виходячи з якогось представлення ознак ϕ . Теорема Мерсера каже нам, що $K(x, z)$ — (мерсерів) кернел тоді і тільки тоді, коли для будь-якої скінченної множини $\{x^{(1)}, \dots, x^{(m)}\}$ матриця K є симетричною та невід'ємно визначеною (positive semi-definite), де матриця $K \in \mathbb{R}^{m \times m}$ розміром $m \times m$ задана як: $K_{ij} = K(x^{(i)}, x^{(j)})$.

Нехай:

- K_1 та K_2 — кернели над векторами $\mathbb{R}_n \times \mathbb{R}_n$.
- $a, b \in \mathbb{R}^+$ — дійсні додатне значення.
- $f: \mathbb{R}^n \rightarrow \mathbb{R}$ — функція, що проектує вектор розмірністю n на дійсне число.
- $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^d$ — функція, що проектує вектор розмірністю n на вектор розмірністю d .
- K_3 — кернел над векторами $\mathbb{R}^d \times \mathbb{R}^d$.
- $p(x)$ — многочлен зі змінною x та *тільки додатними* коефіцієнтами.

Для кожної з приведених функцій вкажіть, чи є вона кернелом, чи ні. Якщо ви думаєте, що функція є кернелом, доведіть це. Якщо думаєте, що не є, наведіть один контр-приклад.

- a. $K(x, z) = K_1(x, z) + K_2(x, z)$
- b. $K(x, z) = K_1(x, z) - K_2(x, z)$
- c. $K(x, z) = a \cdot K_1(x, z)$
- d. $K(x, z) = -a \cdot K_1(x, z)$
- e. $K(x, z) = K_1(ax, bz)$
- f. $K(x, z) = K_1(x, z) \cdot K_2(x, z)$

- g. $K(x, z) = f(x)f(z)$
- h. $K(x, z) = K_3(\phi(x), \phi(z))$
- i. $K(x, z) = p(K_1(x, z))$
- j. $K(x, z) = aK_1(x, z) - bK_2(x, z)$
- k. $K(x, z) = -aK_1(x, z) - bK_2(x, z)$

2. Складні проекції ознак.

[15 балів]

В лекції ми бачили, як функція кернелу аналітично реалізує проекцію ознак x в новий простір $\phi(x)$. Ми також можемо ускладнювати цю проекцію, проектуючи точку $\phi(x)$ в простір $\phi(\phi(x))$.

- a. **[3 бали]** Якщо проекцію з x в $\phi(x)$ реалізує лінійна функція кернелу $K(x, z)$, якою буде функція кернелу, що реалізує проекцію з x в $\phi(\phi(x))$?
- b. **[5 бали]** Якщо проекцію з x в $\phi(x)$ реалізує поліноміальна функція кернелу $K(x, z)$, якою буде функція кернелу, що реалізує проекцію з x в $\phi(\phi(x))$?
- c. **[7 бали]** Якщо проекцію з x в $\phi(x)$ реалізує radial basis function кернел $K(x, z) = e^{-\epsilon(x-z)^2}$, якою буде функція кернелу, що реалізує проекцію з x в $\phi(\phi(x))$?

3. Нейронні мережі.

[15 балів]

Розглянемо нейронну мережу з активаційною функцією сигмоїди на прихованому шарі (hidden layer):

$$a_h(x) = g(x) = \frac{1}{1+e^{-W^T x}}$$

Покажіть, що існує нейронна мережа з активаційною функцією гіперболічного тангенсу, яка обраховує таку саму функцію, що і перша мережа.

$$a_h(x) = \tanh(x) = \frac{e^{W^T x} - e^{-W^T x}}{e^{W^T x} + e^{-W^T x}}$$

4. Програмування: GBM локалізації всередині приміщення.

[15 балів + 10 додаткових балів]

Локалізація людини всередині приміщення – одна з найцікавіших задач в області роздрібної торгівлі та маркетингу. Інформація про те, що людина затрималася біля певної вітрини, зайшла в певний магазин або певний відділ має можливість більш якісно пропонувати рекомендовані товари або знижки. Одним з найбільш дешевих способів такої локалізації є використання радіосигналів від якихось фіксованих джерел. Наприклад, базових станції мобільного зв'язку або wi-fi точок доступу. У цьому завданні ви працюватимете з даними про потужність сигналів від семи різних точок доступу в різних локаціях чотирьох кабінетів (джерело: [UCI Machine Learning Repository](#)).

У записнику **Problem Set 2 – Indoor localization.ipynb** уже підготовлений загальний каркас коду та реалізована підготовка даних.

- a. **[5 балів]** Побудуйте та навчіть на навчальній вибірці XGBoost regressor для визначення кімнати по потужності сигналів від семи wi-fi точок доступу. Перевірте його точність на крос-валідаційній вибірці.
- b. **[10 балів + 10 додаткових балів]**: Підберіть гіперпараметри XGBoost таким чином, щоби максимізувати точність на крос-валідаційній вибірці. Студентам, які побудують моделі, що покажуть найкращу точність на крос-валідаційній вибірці порівняно з моделями групи (окремо для бакалаврів і магістрів) буде присвоєно додаткові бали за таким принципом:
 - 1 місце: 10 додаткових балів;
 - 2 місце: 5 додаткових балів;
 - 3 місце: 3 додаткових бали.

5. Програмування: CNN для розпізнавання зображень.

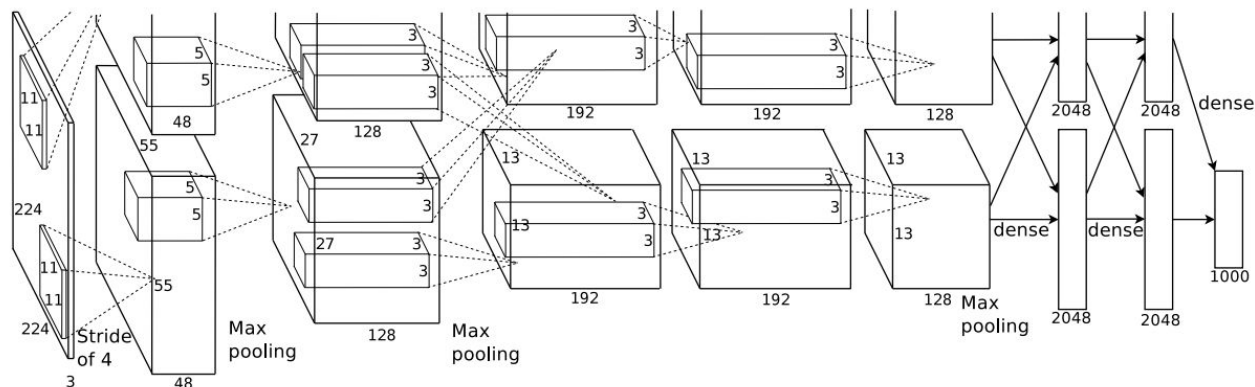
[25 балів + 15 додаткових балів]

Відродження інтересу до нейронних мереж і deep learning багато в чому почалося з перемоги Алекс Крижевський (Alex Krizhevsky) на змаганні ImageNet Competition 2012. Алекс ускладнив архітектуру [LeNet-5](#), запропоновану Яном ЛеКуном в 1998 році, і пристовував її до зображень різного розміру, з яких складається ImageNet. Ця нова мережа була названа ним [AlexNet](#).

У цьому завданні ви побудуєте і використаєте AlexNet для розпізнавання зображень.

ImageNet містить 1.2 млн. зображень і 1000 класів. Навчання мережі на такому об'ємі даних вимагає значного часу. З метою дати можливість навчати модель

швидше для цього завдання замість ImageNet ми будемо використовувати [CIFAR-10](#), який містить 60 тис. зображень і 10 клас



Вам потрібно буде ознайомитися з [документацією Keras](#) і навчитися створювати моделі шар за шаром ([Sequential](#)), компілювати їх та виконувати їх моніторинг. У записнику **Problem Set 2 – Image recognition.ipynb** уже підготовлений загальний каркас коду, реалізована підготовка даних та їх візуалізація.

а. **[10 балів]** Реалізуйте AlexNet для класифікації зображень з CIFAR-10. Зверніть увагу на такі важливі моменти.

- Розмір вхідної картини в нашому датасеті CIFAR — 32x32, що менше, ніж 224x224 на схемі вище.
- Зображення з CIFAR-10 мають 10 класів, а не 1000, як в ImageNet.
- Функцією активації у нас буде ReLU.
- Оригінальний AlexNet використовує model parallelism: модель тренується на двох GPU картках так, що одна картка «бачить» верхню половину зображення, а друга – нижню. На певних етапах ці моделі (та картки) обмінюються інформацією між собою (див. схему вище).
- Ми не будемо використовувати model parallelism для цієї задачі. Вам потрібно імплементувати лише одну мережу, яка аналізуватиме картинку повністю, без розбиття на верхню і нижню частини.

б. **[5 балів]** Натренуйте мережу на навальній вибірці і перевірте результати її роботи на крос-валідаційній вибірці.

с. **[10 балів + 15 додаткових балів]** Оригінальний AlexNet був розроблений для розпізнавання зображень набагато більшого розміру і з набагато більшою кількістю класів, ніж є в CIFAR-10. Проте мережу можна модифікувати для

кращої роботи на CIFAR. Змініть мережу так, щоб вона показувала якомога більшу точність на крос-валідаційній вибірці. Ви можете застосувати зміну кількості фільтрів чи нейронів на шарах нейронної мережі, регуляризацію, dropout та будь-які інші механізми забезпечення генералізації.

Студентам, які побудують моделі, що покажуть найкращу точність на крос-валідаційній вибірці порівняно з моделями групи (окремо для бакалаврів і магістрів) буде присвоєно додаткові бали за таким принципом:

- 1 місце: 15 додаткових балів;
- 2 місце: 10 додаткових балів;
- 3 місце: 5 додаткових балів.

6. Властивості VC розмірності

[15 балів]

У цьому завданні ми дослідимо кілька цікавих властивостей розмірності Вапніка-Червоненкіса, зокрема про те, як збільшується $VC(H)$ зі збільшенням розміру H . Для кожного пункту цієї задачі вкажіть, чи є твердженням правильним (і тоді доведіть його), або чи є воно невірним (і тоді наведіть контрприклад).

- d. **[5 балів]** Нехай для двох класів гіпотез H_1 та H_2 виконується $H_1 \subseteq H_2$.
Правильно чи ні: тоді $VC(H_1) \leq VC(H_2)$.
- e. **[5 балів]** Нехай $H_1 = H_2 \cup \{h_1, \dots, h_k\}$ - H_1 - об'єднання H_2 та k додаткових гіпотез h_i . Правильно чи ні: $VC(H_1) \leq VC(H_2) + k$
(підказка: ви можете почати з випадку, коли $k = 1$).
- f. **[5 балів]** Нехай $H_1 = H_2 \cup H_3$. Правильно чи ні:
 $VC(H_1) \leq VC(H_2) + VC(H_3)$

7. MAP та ослаблення ваг

[15 балів]

Візьмемо логістичну регресію з сигмоїдою як функцією гіпотези: $h(x) = g(\theta^T x)$ і навчальну вибірку: $\{(x^{(i)}, y^{(i)}), i = 1 \dots m\}$. Визначення ваг θ методом максимальної вірогідності (maximum likelihood) виглядає наступним чином:

$$\theta_{\text{ML}} = \arg \max_{\theta} \prod_{i=1}^m p(y^{(i)} | x^{(i)}; \theta)$$

Ми хочемо регуляризувати логістичну регресію, вводячи баєсову апіорну імовірність $p(\theta)$. Нехай ми вибрали апіорний розподіл як гаусів (нормальний) розподіл: $p(\theta) \sim N(o, \tau^2 I)$, де $\tau > 0$, а I - $n + 1 \times n + 1$ одинична матриця. Тоді MAP метод визначення вагів виглядатиме так:

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta) \prod_{i=1}^m p(y^{(i)} | x^{(i)}, \theta)$$

Доведіть, що:

$$\|\theta_{\text{MAP}}\|_2 \leq \|\theta_{\text{ML}}\|_2$$

[Підказка: ви можете використати доведення від супротивного]

Ремарка: через доведену вами у цій задачі закономірність, ця форма регуляризації також називається **ослаблення ваг (weight decay)**, оскільки вона змушує ваги θ бути меншими.