Курс «Машинне навчання»

Домашнє завдання 1: «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 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.

Технічні примітки

1. Завдання з програмування використовують 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
3. Щоб працювати з .ipynb-файлами, використовуйте Jupyter. В командному рядку перейдіть у папку з цими файлами та виконайте: **jupyter notebook**.

1. Зважена лінійна регресія.

[15 балів]

Зважена лінійна регресія — це регресія, у якій ми по-різному оцінюємо помилку для кожного з навчальних прикладів. Для навчання зваженої лінійної регресії нам потрібно мінімізувати функцію втрат виду:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} \omega^{(i)}\left(\theta^{\top}x^{(i)} - y^{(i)}\right)^2$$

a. **[5 балів]** Покажіть, що функція втрат зваженої лінійної регресії $J(\theta)$ також може бути записана у такому вигляді:

$$J(\theta) = (X\theta - \vec{y})^{\top}W(X\theta - \vec{y})$$

де $X, \vec{y}$ визначені так само, як на лекції, а W — діагональна матриця. Поясніть, що таке матриця W та якими будуть її елементи.

b. **[10 балів]** Якщо всі $\omega^{(i)} = 1$, тоді, як ми бачили на лекції, нормальне рівняння має вигляд:

$$X^{\top}X\theta = X^{\top}\vec{y}$$

а значення θ, що мінімізує функцію втрат і дає найвищу точність передбачення:

$$\theta = \left(X^{\top}X\right)^{-1}X^{\top}\vec{y}$$

Виведіть вираз для знаходження θ у **зваженій** лінійній регресії. Знайдіть градієнт $\nabla_{\theta}J(\theta)$ і, прирівнявши його до нуля, виведіть нормальне рівняння для знаходження θ. Вираз буде залежати від X, W і $\vec{y}$.

2. Регресія Пуассона та сімейство експоненціальних моделей.

[25 балів]

Ви маєте задачу: передбачити кількість звернень у службу підтримки вашого сайту в певний день. У службу підтримки за день звертається цілочисельна кількість людей, яких зазвичай не більше 7–10, тому ви вирішили використовувати розподіл Пуассона для моделювання таких звернень.

a. **[7 балів]** Розподіл імовірностей Пуассона має вигляд:

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

Покажіть, що розподіл Пуассона належить до експоненціального сімейства і вкажіть, чому дорівнюють $b(y), \eta, T(y), a(\eta)$.

b. **[3 бали]** Якою буде канонічна функція відгуку (canonical response function) для цього розподілу?
Ви можете використати той факт, що випадкова величина з розподілом Пуассона з параметром λ має середнє значення λ.

c. **[15 балів]** Для навчальної вибірки $\{(x^{(i)}, y^{(i)}); i = 1, \dots m\}$ логарифмічна функція правдоподібності (log-likelihood) буде:

$$\ell(\theta) = \log P\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

Виведіть похідну $\frac{\partial}{\partial \theta_j}\ell(\theta)$ та сформулюйте правило оновлення ваги θ_j методом стохастичного градієнтного підйому, якщо y має розподіл Пуассона та канонічну функцію відгуку.

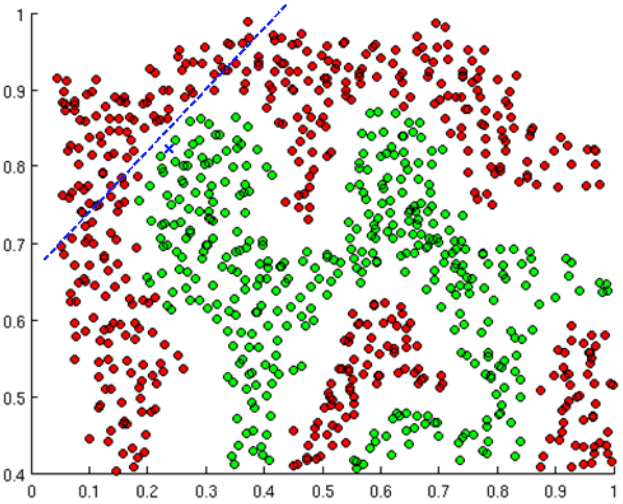
3. Зважена логістична регресія

[30 балів]

Деякі моделі навчання з вчителем, такі як нейронні мережі, наприклад, мають функцію гіпотези дуже високої варіативності. Інтуїтивно, це означає, що вони мають дуже «гнучку» межу класів (decision boundary). Проте, їх передбачення часто важко пояснити. В деяких задачах (наприклад, при автоматизованій діагностиці в медицині) пояснення рішень моделей машинного навчання так само важливо, як і їх точність. Без цього лікарі не довіряють висновкам машини.

Проте, ми можемо модифікувати логістичну регресію так, щоб вона була здатна робити якісні передбачення навіть в умовах дуже нелінійної межі класів (decision boundary). Передбачення логістичної регресії пояснювати доволі просто.

Ваша задача – модифікувати логістичну регресію таким чином, щоб вона передбачала якомога точніше найближчі точки до точки запиту (query point), аналогічно зваженій лінійній регресії.



Для обрахунку вагових коефіцієнтів ми будемо використовувати таку саму формулу, яку використовували для зваженої лінійної регресії, записану у векторній формі:

$$\omega^{(i)} = \exp\left(-\frac{\left(x^{(i)} - x\right)^T \left(x^{(i)} - x\right)}{2\tau^2}\right)$$

a. **[5 балів]** Виведіть формулу функції втрат (cost function) для зваженої логістичної регресії.
b. **[10 балів]** Виведіть правило оновлення ваг θ для зваженої логістичної регресії методом градієнтного спуску.
c. **[15 балів]** Виведіть рішення зваженої логістичної регресії методом нормальних рівнянь.

4. Масштабування ознак в лінійній регресії

[15 балів]

Ви побудували лінійну регресію для передбачення витрат палива на 100 км. шляху. Як ознаки (features) ви використали вагу автомобіля з пасажирами, об'єм двигуна та швидкість в кілометрах на годину. Після впровадження моделі у виробництво виявилось, що ознака швидкості приходить не в кілометрах на годину, а в милях на годину. Ваша модель вже реалізована на мікросхемі, змінити її чи формат даних, які поступають на вхід, дуже дорого. Проте, змінити передбачення перед тим, як видавати його користувачу, - просто і дешево. Як можна модифікувати передбачення так, щоб воно було коректним?

5. Програмування: зважена лінійна регресія.

[35 балів]

У цьому завданні ви працюватимете з реальними даними з зарплатного опитування DOU.ua за травень 2016р. Ви реалізуєте зважену лінійну регресію, яка передбачає зарплати Java-інженерів, та навчите свою модель за допомогою градієнтного спуску.

У записнику **salary-prediction-wlr.ipynb** уже реалізована підготовка даних, візуалізація та оцінка результатів моделі. Вам залишається реалізувати саму логіку зваженої лінійної регресії, заповнивши пропущені місця в коді.

Після того, як завершите, запустіть останню комірку — вона містить автоматичні тести, що перевірять правильність ваших обчислень. Ви повинні побачити повідомлення «ОК», якщо все працює вірно.

a. **[3 бали]** Реалізуйте функцію гіпотези зваженої лінійної регресії.
b. **[10 балів]** Реалізуйте функцію зважування навчальних прикладів.
c. **[7 балів]** Реалізуйте функцію втрат зваженої лінійної регресії.
d. **[12 балів]** Обчисліть градієнт функції втрат зваженої лінійної регресії.
e. **[3 балів]** Реалізуйте правило оновлення ваг при градієнтному спуску.

6. Програмування: класифікація спаму методом Баєса.

[25 балів]

Ви застосуєте наївний Баєсівський класифікатор зі згладжуванням Лапласа для навчання спам-фільтру (на основі даних SpamAssassin Public Corpus).

Пригадайте, що ми розглядали дві моделі подій для класифікації текстів — **multivariate Bernoulli event model** та **multinomial event model**. Ваше завдання — реалізувати класифікацію методом multinomial event model.

Заповніть пропущений код у записнику **spam-bayes-multinomial.ipynb**. У вас повинен вийти кращий результат, ніж при multivariate Bernoulli (> 95% точності).

a. **[2 бали]** Закодуйте лист у вигляді вектора ознак.
b. **[3 бали]** Підрахуйте кількість слів у ham- та spam- листах.
c. **[3 балів]** Обчисліть апріорні ймовірності для класів ham та spam.
d. **[10 балів]** Обчисліть імовірності появи слів в рамках кожного класу.
e. **[7 балів]** Реалізуйте функцію класифікації нового листа.