Домашнє завдання 1

«Машинне навчання»

Lviv IT School

Зверніть увагу

- питання цієї домашньої роботи вимагають певного обмірковування, але не вимагають довгих відповідей. Будь ласка, будьте якомога більш стислі.
- якщо ви маєте будь-які питання щодо цієї домашньої роботи, будь ласка напишіть ваше питання на форумі курсу: http://my.qa-school.lviv.ua/mod/forum/view.php?f=75
- ви можете обговорювати домашні завдання в групах, але не показуйте іншим свої рішення і не користуйтесь готовими чужими.
- для задач, які вимагають написання програм, будь ласка, включіть у звіт ваш код (з коментарями) та ті графіки, які (і якщо) потрібно намалювати відповідно до умов задачі.
- будь ласка, вкажіть ваше ім'я та прізвище у звіті.

1. Лінійна регресія [25 балів]

У вас є дані вартості оренди квартир у Львові, зібрані з www.ria.com. Дані збережені в файлі prices.csv. Погляньте на нього, але не змінюйте, тому що інакше автоматичне тестування ваших відповідей може не працювати.

В архіві разом з цим описом домашнього завдання також є файли:

- prices.csv
- getError.m
- getGradient.m
- getGradientDescentStep.m
- getLinear.m
- getNormalEquations.m
- getWeightedLRPrediction.m
- runTests.m

Кожен з файлів містить коментарі з описом структури каркасу та функціональності, яку вам потрібно дописати самостійно.

(a) **[2 бали]** У файлі getLinear.m імплементуйте функцію гіпотези

$$h(x) = \sum_{i=1}^{n} x_i \theta_i$$

у матричній формі (підказка: зверніть увагу на розмірність вхідних матриць і векторів і вихідного вектору).

(b) **[3 бали]** У файлі getError.m імплементуйте функцію втрат (cost function) найменших квадратів (least squares) $J(\theta)$ у матричному вигляді.

(c) **[3 бали]** У файлі getGradient.m імплементуйте градієнт функції втрат $J(\theta)$.

(d) **[5 балів]** У файлі getGradientDescentStep.m імплементуйте один крок групового градієнтного спуску (batch gradient descent) – оновлення θ.

(e) **[3 бали]** У файлі getNormalEquations.m імплементуйте знаходження θ методом нормального рівняння.

(f) **[9 балів]** У файлі getWeightedLRPrediction.m імплементуйте передбачення вартості оренди квартири за допомогою зваженої лінійної регресії (weighted linear regression).

runTests.m – код, що допоможе протестувати вашу реалізацію. Окрім автоматичного тестування вашої реалізації, він також виводить додаткову інформацію, яка допоможе вам зрозуміти, що ваш код працює правильно:

- зі збільшенням кількості пройдених градієнтним спуском кроків загальна помилка має зменшуватись;
- ваги, вивчені за допомогою градієнтного спуску та нормальних рівнянь, мають мати близькі одне до одного значення;
- передбачення ціни квартири за допомогою зваженої лінійної регресії має бути більш точним, ніж зі звичайною лінійною регресією.

2. Логістична регресія (logistic regression) [10 балів]

(a) **[10 балів]** Логарифмічна функція імовірності для логістичної регресії:

$$l(\theta)=\sum_{i=1}^{m} y^{(i)} \log h\left(x^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-h\left(x^{(i)}\right)\right)$$

Знайдіть матрицю Гессе цієї функції H і покажіть, що для будь-якого вектору z виконується така нерівність:

$$z^T Hz \leq 0$$

Підказка: ви можете почати з доведення, що $\sum_i \sum_j z_i x_i x_j z_j = \left(x^T z\right) \geq 0$

3. Зважена лінійна регресія (weighted linear regression) [10 балів]

Зважена лінійна регресія – регресія, у якій ми по-різному оцінюємо помилку для кожного з навчальних прикладів (training example). Для навчання зваженої лінійної регресії нам потрібно мінімізувати функцію втрат (cost function) виду:

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

На лекції ми вивели, що станеться, якщо всі ваги $\omega^{(i)}$ є однаковими. Ц цій задачі ми приведемо зважену лінійну регресію до узагальненого вигляду, а також реалізуємо її в коді.

(a) **[3 бали]** Покажіть, що функція втрат $J(\theta)$ також може бути записана як:

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

для діагональної матриці W, а X та $\vec{y}$ визначені так само, як на лекції. Поясніть, що таке матриця W та чим будуть її елементи.

(b) **[7 балів]** Якщо всі $\omega^{(i)}=1$, тоді, як ми бачили на лекції, нормальне рівняння є таким:

$$X^T X\theta=X^T\vec{y},$$

а значення θ, що мінімізує функцію втрат (і дає найвищу точність передбачення) є $\left(X^T X\right)^{-1}X^T\vec{y}$.

Знайдіть похідну $\nabla_\theta J(\theta)$ і, прирівнявши її до нуля, виведіть нормальне рівняння для знаходження θ, що мінімізує функцію втрат для зваженої лінійної регресії, де кожна вага $\omega^{(i)}$ має своє значення. Рівняння буде залежати від X, W і $\vec{y}$.

4. Регресія Пуассона і сімейство експоненціальних моделей [18 балів]

(a) **[5 балів]** Розподіл імовірності Пуассона:

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

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

(b) **[3 бали]** Ви маєте задачу: передбачити кількість звернень в службу підтримки вашого сайту в певний день. В службу підтримки за день звертається зазвичай не більше 7-10 чоловік, тому ви вирішили використовувати розподіл Пуассона для моделювання таких звернень.
Якою буде канонічна функція відгуку (canonical response function) для цього сімейства? (Ви можете використати те, що випадкова величина за розподілом Пуассона з параметром λ має середнє значення λ).

(c) **[10 балів]** Для навчальної вибірки (training set) $\{(x^{(i)}, y^{(i)}); i = 1,..,m\}$, логарифмічна функція імовірності (log-likelihood) буде:

$\log p\left(y^{(i)} \mid x^{(i)}; \theta\right)$. Виведіть градієнт логарифмічної функції імовірності від θ_j та сформулюйте правило оновлення θ_j методом стохастичного градієнтного підйому (для максимізації імовірності) з відгуком y, що має розподіл Пуассона, та канонічну функцію відгуку (canonical response function).