Практически проблеми при градиентно спускане. Нормализация. Скорост на обучение

Доц. д-р Ивайло Пенев

Кат. „Компютърни науки и технологии“

Практически проблеми при градиентно спускане – мащабиране на променливи (Feature Scaling)

- Градиентното спускане може да се ускори, ако стойностите на променливите са в един или в близки диапазони
- При разлика в диапазоните на променливите параметрите θ се приближават бавно към минимума на ф-та на цената J
- В идеалният случай диапазонът е:
  $-1 \leq x_{(i)} \leq 1$ или $-0.5 \leq x_{(i)} \leq 0.5$
- Когато стойностите на променливите са в различни диапазони, те могат да бъдат преобразувани така, че стойностите им да попадат в един диапазон
- Прилагат се два метода – мащабиране (feature scaling) и нормализация на средната стойност (mean mormalization)

Мащабиране и нормализация на средната стойност

- Мащабиране (feature scaling) – стойностите на входната променлива се делят на диапазона на изменение на всички стойности на входната променлива (т.е. разликата на максималната и минималната стойност)
- Нормализация на средната стойност (mean normalization) – от всяка стойност на входната променлива се изважда средната стойност от всички стойности на променливата и получената разлика се разделя на диапазона, т.е.:

$$x_i := \frac{x_i - \mu_i}{s_i}$$

, където

μ_i - средна стойност на всички стойности на променливата x_i

s_i - диапазон на изменение на стойностите на x_i (max-min) или стандартно отклонение

Пример за нормализация

- Ако x_i е цена на m^2 за къща с диапазон от 100 до 2000 и средна стойност 1000, то нормализираната x_i има следният вид:

$$x_i \coloneqq \frac{x_i - 1000}{1900}$$

Влияние на скоростта на обучение върху градиентното спускане

Повтаряй до достигане на минимум на ф-та на цената $J(\theta)$

{

$$\theta_0 \coloneqq \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_0^{(i)}$$

$$\theta_1 \coloneqq \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_1^{(i)}$$

$$\theta_2 \coloneqq \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_2^{(i)}$$

…

}

α – скорост на обучението (learning rate) – определя стъпката на изменение на параметрите θ

Проследяване на градиентното спускане

- Как да проверим дали градиентното спускане приближава ф-та на цената $J(\theta)$ до минимална стойност?
- Чертаем графика с брой итерации на градиентното спускане по оста x и стойността на $J(\theta)$ за всяка итерация. Ако на някоя итерация $J(\theta)$ се увеличи, това означавам че трябва да намалим скоростта на обучение α
- Кога $J(\theta)$ достига минимум?
- Ако на дадена итерация $J(\theta)$ намалява със стойност по-малка от определена стойност E, то това означава, че $J(\theta)$ е достигнала минимум
- E – прагова стойност, която е достатъчно малка, напр. E = 10^{-3} (практически е трудно да се избере стойност)

Обобщение за скоростта на обучение

- Практически е доказано, че ако α е достатъчно малка, то $J(\theta)$ намалява на всяка итерация
- Проблеми
  - Ако α е прекалено малка, градиентното спускане ще се изпълнява бавно
  - Ако α е голяма, е възможно $J(\theta)$ да не достигне минимум

Подобряване на функцията на хипотезата – полиномна регресия

- Възможно е да комбинираме множество променливи в една
- Напр. можем да комбинираме променливи x1 и x2 в една променлива x3, където x3=x1.x2
- Не е задължително функцията на хипотезата да бъде права линия – може да е квадратна, кубична функция, квадратен корен и др.

Примери за нелинейни функции

- Линейна функция $h_\theta(x) = \theta_0 + \theta_1 x_1$
- Квадратна функция $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$, където е добавена нова променлива $x_2 = x_1^2$
- Кубична функция $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^3$, където $x_3 = x_1^3$
- Функция квадратен корен $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_1}$
- При такова добавяне на променливи мащабирането е много важно
- Напр., ако x_1 има диапазон от 1 до 1000, то x_1^2 има диапазон от 1 до 1000000, а x_1^3 от 1 до 1000000000