Лабораторная работа 1.
Методы линейной регрессии.

6 марта 2013 г.

Задание Есть набор данных **house_cost.txt**. Этот набор данных состоит из трех колонок:

1. Первая колонка - размер дома в футах^2
2. Вторая колонка - число спален
3. Третья колонка - стоимость дома

Необходимо найти функцию линейной регрессии стоимости дома от остальных переменных $y = b_0 + x_1 b_1 + x_2 b_2$ следующими способами:

1. С помощью встроенных функций R
2. Аналитически
3. С помощью градиентного спуска
4. С помощью наискорейшего спуска

Решение задания должно быть написано на языке R в одном файле с расширением '.R'. Для каждого способа должна быть реализована соответствующая функция:

1. lm_builtin(y,X)
2. lm_analytical(y,X)
3. lm_gradient(y,X)
4. lm_quick(y,X)

Каждая из функций должна принимать на вход в качестве параметра y вектор стоимостей домов, а в качестве параметра X - дата фрейм, содержащий в первой колонке размеры домов, а во второй - число спален, и возвращать новую функцию. Возвращаемая функция должна принимать на вход новую матрицу X и возвращать для нее соответственно вектор предсказаний $y = X\hat{b}$, где $\hat{b}$ - вектор подобранных коэффициентов модели.

Встроенные функции Необходимо использовать функцию lm().

Аналитическое решение

$$\hat{b} = (X^T X)^{-1} X^T y$$

Градиентный спуск

$$\hat{b}^{[k+1]} = \hat{b}^{[k]} - \lambda \nabla J(\hat{b}^{[k]})$$

где $J(b) = \frac{1}{2m}(Xb-y)^T(Xb-y)$ - целевая функция ошибки. Параметр λ следует выбирать не очень большим, чтобы градиентный спуск не разошелся, но и не слишком маленьким, чтобы он работал не очень медленно.

Наискорейший спуск

$$\hat{b}^{[k+1]} = \hat{b}^{[k]} - \lambda^{[k]} \nabla J(\hat{b}^{[k]})$$

$$\lambda^{[k]} = \arg\min_{\lambda} J(\hat{b}^{[k]} - \lambda \nabla J(\hat{b}^{[k]}))$$

Задачу оптимизации для поиска $\lambda^{[k]}$ можно решать аналитически.