МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОННОЙ ТЕХНИКИ
Институт системной и программной инженерии
и информационных технологий (Институт СПИНТех)

Лабораторный практикум по курсу "Интеллектуальные системы" (09/22 – 01/23)

Лабораторная работа 1
Линейная регрессия как задача контролируемого (индуктивного) обучения [1].

На этом занятии компьютерного практикума вы изучите *линейную регрессию* и получите представление, о том, как данная процедура применяется для обработки данных. Во многих приложениях нейронные сети реализуют регрессионные вычисления (статистический метод исследования влияния одной или нескольких независимых переменных x_1, x_2, ... , x_p на зависимую переменную y) или решение задач классификации. При этом, в частности, линейная регрессия (англ. *Linear regression*) представляет собой регрессионную модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

Прежде чем приступить, собственно к программированию, настоятельно рекомендуется ознакомиться с материалом лекций, а также с дополнительными материалами, имеющими отношение к задаче градиентного спуска и к области минимизации функционалов.

Файлы, включенные в задание:

ex1.m - скрипт, реализующий пошаговое выполнение задания 1 части задания по разделу «Линейная регрессия с одной переменной» (простая линейная регрессия);
ex1_multi.m - скрипт, реализующий пошаговое выполнение 2-й части задания по разделу «Многомерная линейная регрессия»;
ex1data1.txt - набор данных для простой линейной регрессии;
ex1data2.txt - набор данных для многомерной линейной регрессии;
warmUpExcercise.m - простая тестовая функция в MATLAB;
plotData.m - функция, отображающая набор данных;
computeCost.m - программа, с помощью которой производится вычисление функции стоимости (целевой функции);
gradientDescent.m - программа, реализующая метод градиентного спуска;
computeCostMulti.m - программа, с помощью которой производится вычисление функции стоимости (целевой функции) для многомерной линейной регрессии;
gradientDescentMulti.m - программа, реализующая метод градиентного спуска для случая нескольких переменных;

[1] Материал лабораторной работы составлен на основании аналогичного задания по курсу «Машинное обучение» на портале онлайн обучения Coursera.org (профессор Эндрю Ын, Стэнфордский университет - https://ru.wikipedia.org/wiki/Ын,_Эндрю)

f*eatureNormalize.m* - функция, с помощью которой производится нормализация признаков;
normalEqn.m - программа, реализующая вычисления по методу наименьших квадратов.

В этой лабораторной работе Вы будете использовать скрипты *ex1.m* и *ex1_multi.m*, не изменяя их. В скриптах (коротких программах) подготовлены обращения к исходным данным. Далее производится вызов функций, написанных Вами, и отображаются результаты вычислений. Необходимо дописать функции в файлах по инструкциям упражнения.

1 Пример простой функции в Matlab

Первый скрипт *ex1.m* поможет Вам еще раз попрактиковаться с MATLAB.

Измените файл *warmUpExercise.m,* чтобы функция *A=warmUpExercise()* выводила единичную матрицу 5 х 5. Затем, запустите *ex1.m* (если ваша текущая директория совпадает с директорией задания, введите в командной строке MATLAB ≪*ex1*≫). Теперь *ex1.m* приостановит выполнение программы до нажатия любой кнопки. После нажатия продолжается выполнение очередного задания. Если Вы хотите прервать выполнение сценария (скрипта), нажмите ≪*ctrl+c*≫.

2. Линейна регрессия с одной переменной

В этой части упражнения Вы реализуете линейную регрессию с одной переменной для прогнозирования прибыли сервисного центра по обслуживанию, например, изделий бытовой электроники. Допустим, рассматриваются кандидатуры нескольких городов для такого сервисного центра. Сеть подобных сервисных центров уже функционирует и у Вас есть данные по прибыли в зависимости от числа жителей города (или от количества изделий, находящихся на обслуживании в городе). Вы бы хотели использовать эти данные с тем, чтобы понять, в каком городе открыть дополнительный сервисный центр или спрогнозировать прибыль в зависимости от числа жителей.

Файл *ex1data1.txt* содержит набор данных для задачи с линейной регрессией. Первый столбец – число жителей города (количество проданных изделий бытовой электроники), второй – прибыль центра в этом городе. Отрицательные значения прибыли означают убыток.

Файл *ex1.m* уже настроен для загрузки этих данных в программу.

2.1 Построение данных

До начала выполнения любого задания было бы полезным представить данные в графическом виде. Для визуализации набора данных (в дальнейшем, при решении аналогичной задачи с помощью нейронной сети, этот набор данных представляет собой обучающие данные) можно использовать точечную диаграмму, т.к. имеется только два свойства – прибыль и население города (количество аппаратов). Многие задачи из реальной жизни имеют большее количество свойств и параметров, так что построить их на двумерном графике, конечно же, не удастся.

В *ex1.m* данные загружаются из файла данных в переменные X и y.

```
data = load('ex1data1.txt');     % Чтение данных, разделённых запятой
X = data(:, 1); y = data(:, 2);
m = length(y);                   % Число обучающих примеров
```

Далее программа вызывает функцию *plotData* для построения точечного графика. Ваша задача – построить график, завершив код в файле *plotData.m*. Напечатайте следующий код после открытия файла plotData.m в любом текстовом редакторе:

```
plot(x, y, 'rx', 'MarkerSize', 10);              % Построение данных
```

```
ylabel('Прибыль (10 тыс руб)');                    % Подпись y-оси
xlabel('Число жителей города (десятки тысяч');    % Название x-оси
```

Для продолжения запустите *ex1.m*. Результат должен выглядеть так, как показано на Рис. 3 с теми же красными крестиками ≪x≫ и подписями к осям.

Указание: Узнать больше о команде *plot* можно напечатав *help plot* в командной строке MATLAB или в интернете. Для того чтобы поменять знак маркера на красный ≪x≫ в команде *plot* была использована опция ≪rx≫: *plot (..,[перечисление опций],.., 'rx'); ...*)

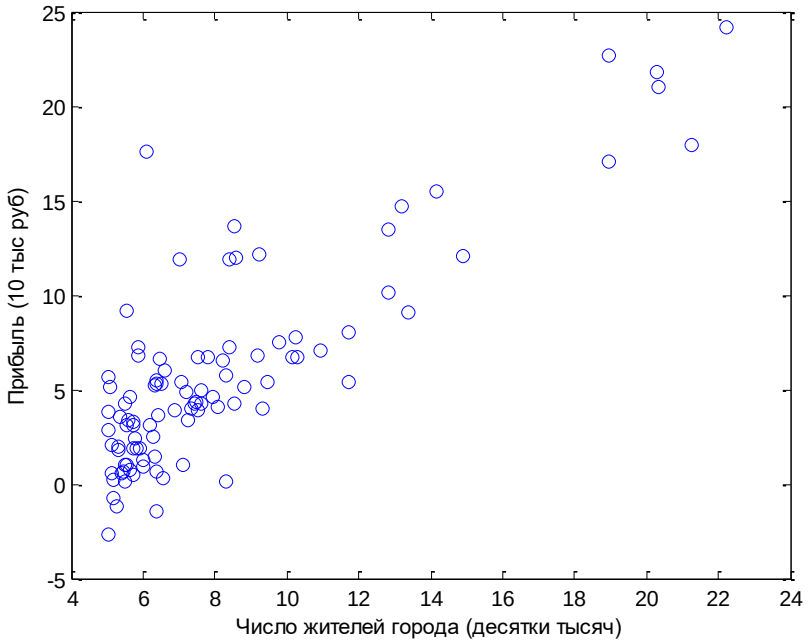


Рис. 3. Точечный график обучающих данных

2.2 Метод градиентного спуска для одной переменной

В этой части производится настройка параметров линейной регрессии θ на основе метода градиентного спуска.

2.2.1 Вычисления

Задача линейной регрессии – минимизировать функцию стоимости (затрат)

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

в котором уравнение гипотезы h_θ (x) представляет собой линейную модель. Напомним, что параметры модели, это значения θ_j. Именно их необходимо подобрать с целью минимизации стоимости $J(\theta)$. Один из способов это сделать - использовать алгоритм наискорейшего спуска, где каждая итерация обновляет

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

(одновременно обновляет θj для всех j).

С каждым шагом градиентного спуска параметры θ_j становятся ближе к оптимальным значениям, при которых достигается наименьшая величина стоимости *J(θ)*.

Указание: Здесь каждый пример хранится в строке матрицы X. С целью учёта дополнительного коэффициента θ_0, к X добавлен единичный столбец, что позволяет рассматривать θ_0 как дополнительный параметр.

2.2.2 Выполнение

В *ex1.m* данные для линейной регрессии уже определены. Добавим дополнительную размерность с целью учёта θ_0. Кроме того, установим начальные параметры, равные нулю, а также скорость обучения равную 0.01.

```
X = [ones(m, 1), data(:,1)];   % Добавить единичный столбец к X
theta = zeros(2, 1);           % Установка начальных значений
iterations = 1500;             % Количество итераций
alpha = 0.01;                  % Скорость обучения
```

2.2.3 Вычисление стоимости J(θ)

Осуществляя вычисления на основе метода градиентного спуска в процессе минимизации функции стоимости *J* (*θ*), полезно отслеживать сходимость вычислительного процесса.

В этой части задания предстоит написать функцию стоимости *J(θ)*, при помощи которой можно отслеживать сходимость уже реализованного процесса спуска.

Следующее задание – завершить код в файле computeCost.m, содержащий функцию *J(θ)*.

Указание: Переменные X и y - не скалярные величины, а матрицы, чьими строками являются примеры из обучающей выборки (набора данных).

Закончив программировать функцию стоимости *J(θ),* запустите скрипт *ex1.m*, который выполнит функцию *computeCost* используя *θ* с начальными значениями, равными нулю. В результате вычислений на экране появится стоимость примерно равная 32.07.

2.2.4 Метод градиентного спуска

Далее завершите код градиентного спуска в файле *gradientDescent.m*. Необходимый цикл был заранее подготовлен и Вам остается лишь дополнить корректировку *θ* для каждой итерации.

В процессе программирования учтите, что стоимость *J(θ)* зависит от параметра *θ*, а не от *X* и *y*. Т.е. необходимо оптимизировать *J(θ)* изменяя значения вектора *θ*, а не *X* и *y*. Хороший способ проверить правильность работы градиентного спуска – *посмотреть* на значения *J(θ)* и проверить их уменьшение с каждой итерацией. Начальный код *gradientDescent.m* вызывает функцию *computeCost* на каждой итерации и выводит на экран значение стоимости. При безошибочном функционировании кода и функции *computeCost*, значения функции стоимости *J(θ)* не возврастают и вычислительный процесс должен сходиться к устойчивой величине в конце вычислений.

После завершения, *ex1.m* использует ваши конечные параметры для построения *линейного* тренда. Результат должен быть схож с Рис. 4.

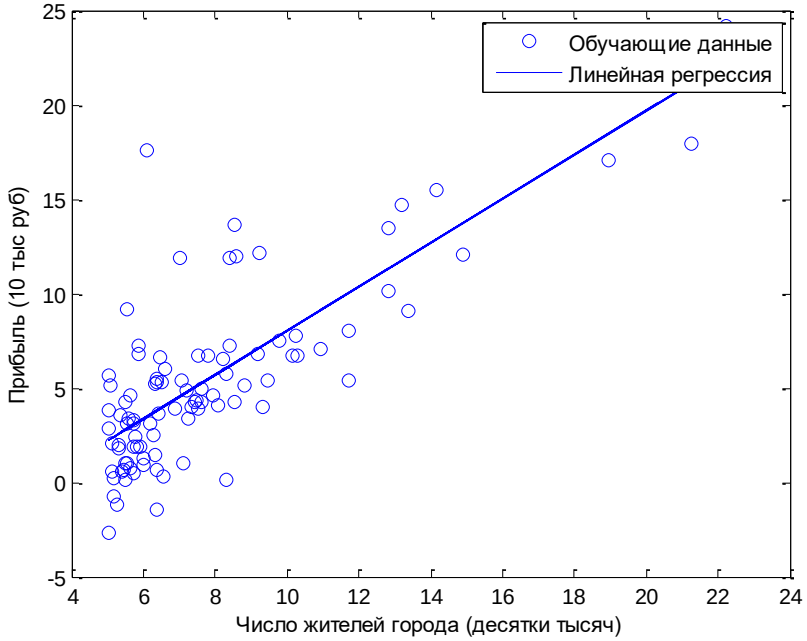


Рис.4. Представление данных в виде тренда линейной регрессии

Оптимальные величины θ могут быть использованы для прогнозирования прибыли в городе с числом жителей (проданных изделий бытовой электроники) 35,000 и 70,000. Обратите внимание на способ, которым следующие строки кода в *ex1.m* используют перемножение матриц, вместо явного сложения или использования цикла для вычисления вашего прогноза. Это пример векторизации (преобразования данных в векторную форму) в MATLAB.

```
predict1 = [1, 3.5] * theta;
predict2 = [1, 7] * theta.
```

2.3 Отладка

Указание: несколько полезных советов для осуществления градиентного спуска:

- Индексы в MATLAB начинаются с 1, а не с 0. Сохраняя θ_0 и θ_1 в векторе theta их значения будут theta(1) theta(2) соответственно;
- Некоторые ошибки возникают из-за не соответствия размеров матриц для операций сложения или умножения.

По умолчанию MATLAB интерпретирует математические операнды как матричные. Это распространенный источник ошибок несовместимости. Если Вы хотите исключить матричные умножения, то следует ставить точку перед операндом умножения. Например, A*B перемножает матрицы, а A.*B перемножает матрицы поэлементно.

2.4 Наглядное представление стоимости J(θ)

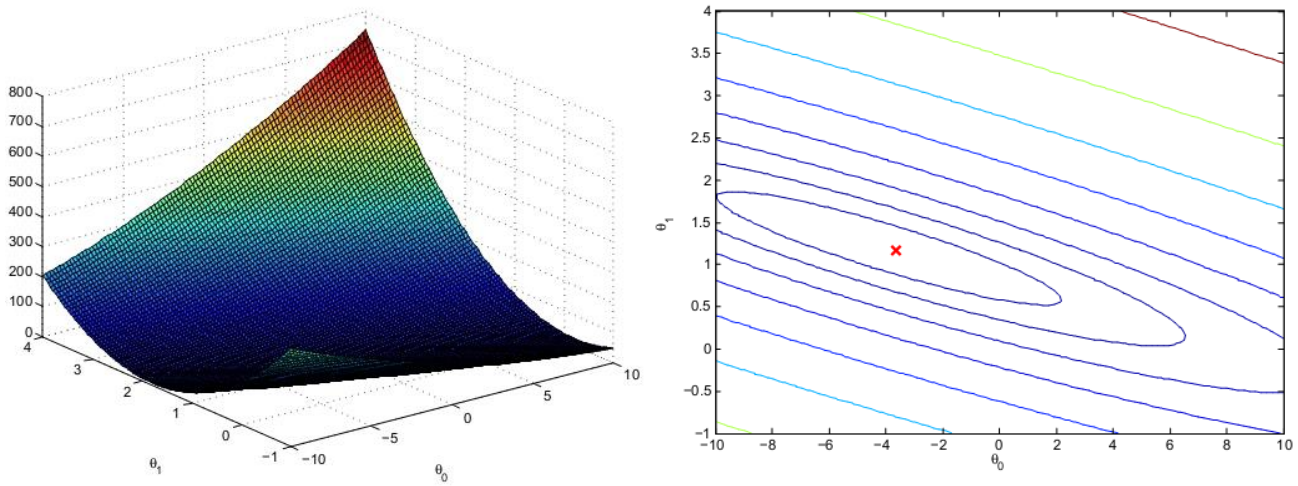
Для лучшего понимания поведения целевой функции (функции стоимости) *J(θ)* постройте двумерный график в зависимости от значений θ_0 и θ_1. Вам не потребуется для этого писать новый код, но предстоит разобраться, как функционирует уже написанная программа.

Следующий шаг в *ex1.m* вычислить набор значений *J(θ)* с использованием написанной вами ранее функции *computeCost*.

```
% задать J_vals нулевой матрицей
J_vals = zeros(length(theta0_vals), length(theta1_vals));
```

```
% вычислить J_vals
for i = 1:length(theta0_vals)
for j = 1:length(theta1_vals)
t = [theta0_vals(i); theta1_vals(j)];
J_vals(i,j) = computeCost(X, y, t);
end
end
```

После выполнения этих строк кода Вы будете располагать двумерным массивом значений *J(θ)*. Программа *ex1.m*, затем, использует эти значения для построения поверхности и контурного графика *J(θ)* с помощью функций *surf* и *contour*. Графики будут выглядеть примерно так, как показано на Рис. 5:



а) поверхность

б) контурный график с обозначенным минимумом

Рис. 5. Функция стоимости *J* (*θ*)

Цель данного графического представления - показать влияние θ_0 и θ_1 на *J(θ)*. Целевая функция (функция стоимости) *J(θ)* приняла форму чаши или цилиндра и имеет глобальный минимум (что будет отчетливее видно на контурном графике, нежели на поверхности). Этот минимум и есть оптимальная точка θ_0 и θ_1 к которой градиентный спуск с каждым шагом всё ближе и ближе.

3. Линейная регрессия для нескольких переменных *

После успешного выполнения предыдущих заданий, Вы лучше владеете методом линейной регрессии и можете им пользоваться для решения практических задач. Это, в свою очередь, позволит в дальнейшем понять процедуру предсказания или интерполяции данных, которая реализуется на основе нейронно-сетевого подхода.

Здесь Вы осуществите метод линейной регрессии для нескольких переменных, например, для оценки стоимости, приобретаемой квартиры, состоящей из нескольких комнат. Один из способов сделать это – накопить информацию о недавних подобных сделках и сконструировать ценовую модель.

Файл *ex1data3.tx*t содержит набор недавних цен на квартиры. Первый столбец – площадь в м^2, второй – количество комнат, третий - цены.

Рекомендуется использовать программу *ex1_multi.m* для выполнения упражнения.

3.1 Нормализация свойств

Файл *ex1_multi.m* сначала загрузит и отобразит выборочные данные. Обратите внимание на то, что численные значения площади квартиры, выраженной в м^2, во много раз больше численных значений, соответствующих количеству комнат в квартире. В таких случаях, когда свойства отличаются значительно (в том числе, на несколько порядков), сначала проводят их нормализацию, что существенно ускоряет процесс схождения градиентного спуска.

Ваша задача завершить программу *featureNormalize.m* с тем, чтобы:

- вычесть среднее значение каждого свойства из набора данных;
- затем поделить полученные значения на соответствующие стандартные отклонения.

≪Стандартное отклонение≫ это измерение (*max* - *min*) диапазона значений для каждого конкретного свойства (большинство отклонений располагается в пределах ± 2 единиц от среднего). В MATLAB стандартное отклонение вычисляет функция *std*. Например, внутри *featureNormalize.m* величина X(:,1) содержит все значения обучающих данных - x1 (площадей квартир), так что команда std(X(:,1)) вычисляет стандартное отклонение данного признака. Во время вызова *featureNormalize.m* первый столбец с единичными значениями x0 = 1, упомянутый выше, пока не был добавлен к X (подробности см. в *ex1_multi.m*). Предстоит сделать эти приготовления для всех свойств так, чтобы составленная Вами программа работала с массивами данных любых размеров (любым количеством свойств/примеров). Заметим, что каждому свойству соответствует один столбец матрицы X.

Указание: при нормализации свойств важно сохранять данные, используемые для нормализации – *средние значения* и *стандартные отклонения*. После обучения параметров модели мы сможем получить новые данные – рассчитать цены на интересующие нас квартиры. Эти новые значения **x** (площадь квартиры, количество комнат и т.д.) следует снова нормализовать, используя рассчитанные ранее средние значения и стандартные отклонения.

3.2 Метод градиентного спуска для нескольких переменных

Ранее Вы уже применяли метод градиентного спуска в процессе решения задачи линейной регрессии для одной переменной. Единственным отличием теперь будет введение дополнительного свойства (признака) в матрицу X. Функция гипотезы и итеративное правило обновления функции градиентного спуска останутся без изменений.

Теперь необходимо завершить функции вычисления стоимости и градиентного спуска для линейной регрессии с несколькими переменными в файлах *computeCostMulti.m* и *gradientDescentMulti.m*. Вы также можете использовать функции из предыдущих заданий, если они написаны в векторизованном виде (т.е. поддерживают многомерные входные данные). Убедитесь, что ваше решение поддерживает любое число свойств и векторизовано!

Сколько свойств содержится в векторе X можно определить при помощи функции ≪size (X, 2)≫.

Указание: в случае с несколькими переменными функция стоимости можно также записать в следующем векторизованном виде:

$$J(\theta) = \frac{1}{2m} (X\theta - \vec{y})^T (X\theta - \vec{y})$$

где

$$X = \begin{bmatrix} — \left(x^{(1)}\right)^T — \\ — \left(x^{(2)}\right)^T — \\ \vdots \\ — \left(x^{(m)}\right)^T — \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}.$$

Векторизованная форма более эффективна для MATLAB в вычислительном плане. Если Вы хорошо разбираетесь матричными операциями то Вам будет несложно доказать, что обе формы записи эквивалентны.

3.3. Выбор скорости обучения

В этой части задания необходимо выяснить скорость обучения для используемых данных, в которой результат сходится быстрее. Вы можете изменять скорость обучения в файле *ex1_multi_data.m.*

В файле *ex1_multi.m* необходимо запустить функцию *gradientDescent* около 50 раз при соответствующей скорости обучения. Ваша функция должна также возвращать массив с данными *J(θ)* в вектор *J*. После последней итерации программа построит график значений *J(θ)* в зависимости от числа итераций. Если выбранные величины скорости обучения совпали с представленными в руководстве (см. ниже), то зависимость расположится ближе к Рис. 6.

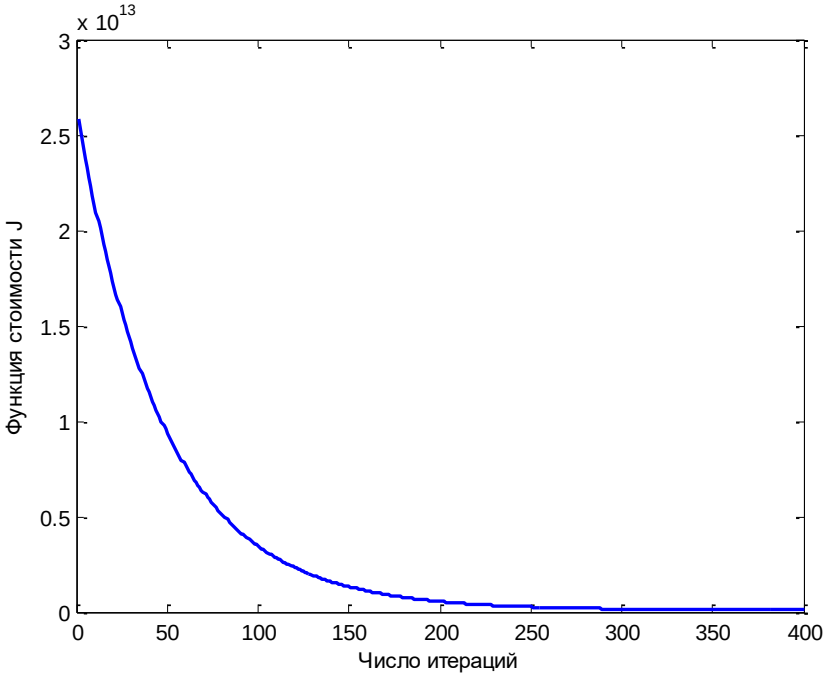


Рис. 6. Сходимость метода градиентного спуска с соответствующим диапазоном скоростей обучения

В противном случае, если значения *J(θ)* увеличиваются или расходятся (в бесконечность), – подберите другой диапазон скорости обучения и попробуйте снова. Рекомендуется попробовать значения скорости обучения *α* с уменьшением, скажем, в 3 раза (т.е. 0.3, 0.1, 0.03, 0.01 и т.д.).

Указание: Если величина *α* очень большая, то *J(θ)* может разойтись, что приведёт к чрезмерно большим значениям для вычислений. В этом случае MATLAB вернёт ≪NaN≫, что ознает ≪not a number≫ (не число). Это частая ситуация при наличии каких-либо операторов с $-\infty$ или $+\infty$.

Указание: Для сравнения влияния скорости обучения на сходимость можно построить *J(θ)* для нескольких *α* на одном графике. В MATLAB это можно выполнить, запустив градиентный спуск несколько раз с командой ≪hold on≫ между построениями. Например, если Вы используете 3 различных *α* (лучше больше) и сохранили значения J1, J2 и J3, то для построения этих точек на одном графике потребуется следующее:

```
plot(1:50, J1(1:50), 'b');
hold on;
plot(1:50, J2(1:50), 'r');
plot(1:50, J3(1:50), 'k');
```

Последние аргументы ≪b≫, ≪r≫, и ≪k≫ задают цвет точек для построения. Обратите внимание на кривые сходимости для разных *α*. С малым *α* скорость схождения до оптимального значения занимает больше времени относительно случая с большим *α*. Кроме того, с большим *α* система может не сойтись вовсе или разойтись!

Используя оптимальную скорость обучения из найденных, запустите *ex1_multi_data3.m* с тем чтобы найти окончательные параметры *θ*. Далее, используя полученные *θ*, спрогнозируйте стоимость трехкомнатной квартиры площадью 60 м^2. Вы примените это значение далее для проверки задания по *аналитическим уравнениям (системы нормальных уравнений)*. Не забудьте нормализовать свойства перед выполнением!

3.3 Система нормальных уравнений. МНК-решение

Замкнутая форма для линейной регрессии записывается как:

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

Использование этой формулы не требует нормализации свойств. Точный результат достигается без использования итераций: нет цикла как в градиентном спуске. Завершите код в *normalEqn.m* используя приведённую формулу, для вычисления *θ*. Здесь не требуется нормализация, однако это не значит, что не нужно добавить единичный столбец (учёт θ_0) к X.

Используйте этот метод для оценки стоимости трехкомнатной квартиры площадью 60 м^2, методом наименьших квадратов. Можете использовать предыдущее задание с использованием метода градиентного спуска для проверки (должно получиться сходное число).