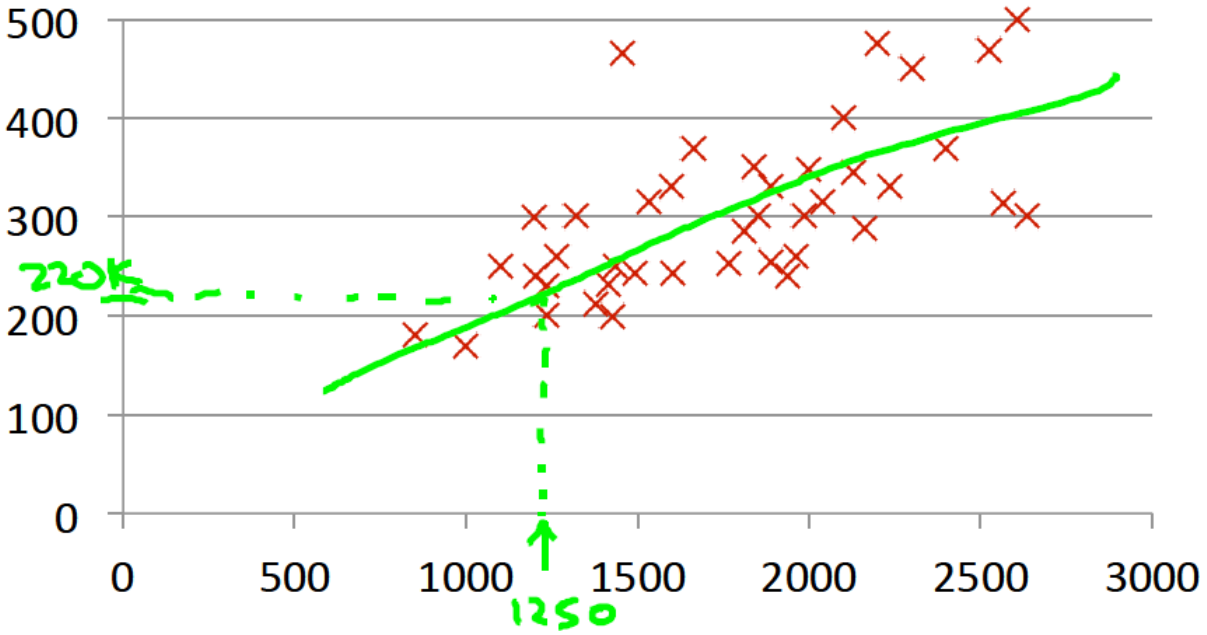
Лабораторная работа №12

Линейная регрессия.

Линейная регрессия позволяет предсказать значение зависимой переменной, исходя из значений одной или нескольких независимых переменных.

Например, пусть имеется зависимость между площадью дома и его ценой.



По оси x – площадь, по оси y – цена. Одной из задач может являться предсказание стоимости дома, площадь которого 1250 условных единиц. Значения, лежащие на оси x обычно называют признаками (features), а значения на оси y целевыми значениями (target).

Задачи подобного рода могут быть решены при помощи линейной регрессии. Нами будет рассмотрен самый простой вариант регрессии – одномерная линейная регрессия. Необходимо понимать, что для успешного предсказания необходимо, чтобы зависимость между данными носила выраженный линейных характер. Для случаев, когда зависимость не является линейной, существует специальный вариант регрессии – нелинейная регрессия. В реальных задачах чаще встречается случай зависимости нескольких переменных – многомерная регрессия, но она остается за рамками рассмотрения.

Чтобы решить задачу регрессии, необходимо высказать какую-то догадку относительно того, какая зависимость наблюдается между данными. Такая догадка называется гипотезой, и в случае одномерной линейной регрессии выражается в виде:

$$h(x) = \theta_0 + \theta_1 x$$

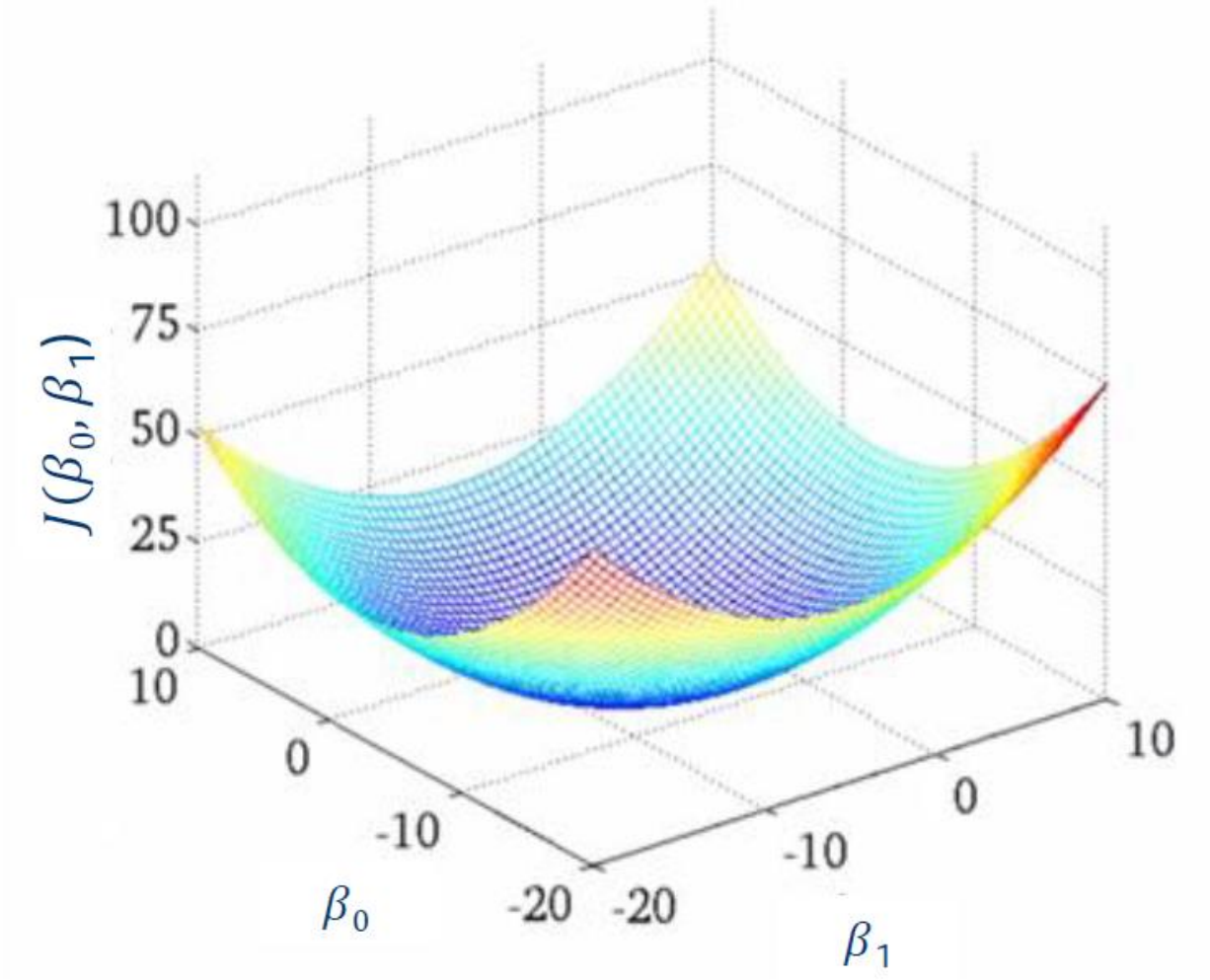
Величины θ_i называются параметрами модели и именно их необходимо определить.

Необходимо выбрать их так, чтобы значение $h(x)$ было максимально близко к y из обучающей выборки, т.е. из исходных данных.

Введем понятие функции стоимости (cost function).

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$$

m – количество пар (x,y) в обучающей выборке. Иногда параметры θ_0, θ_1могут быть обозначены другими символами, например:



Чтобы определить параметры θ_0, θ_1, необходимо минимизировать значение функции стоимости:

$$min_{\theta_0, \theta_1}\frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$$

Для минимизации может быть использован метод градиентного спуска:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Параметр α называется скоростью обучения и изменяется в пределах от 0 до 1.

Для линейной регрессии:

$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)$$

$$\theta_1 = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)\, x_i$$

В нашей задаче для вычисления θ_0, θ_1 будет использоваться конечное число итераций. В качестве начальных значений выбираются произвольные значения, например, 0, 0. На каждом шаге нужно вычислять значения θ_0, θ_1 и только потом изменять их. Например, для расчета θ_1 необходимо использовать старое значение θ_0.Чтобы знать, что вы находитесь на верном пути, необходимо понимать, что на каждом шаге итерации значение функции стоимости должно уменьшаться для новых значений θ_0, θ_1.