Введение в Линейную регрессию

Преподаватель: Герард Костин

План

Введение

- Постановка задачи машинного обучения
- Задача регрессии
- Линейная регрессия
- Метод наименьших квадратов
- Мультиномиальная регрессия
- Кривые обучения

Постановка Задачи

- **Пусть**
  - X — множество описаний объектов,
  - Y — множество допустимых ответов.
- **Существует неизвестная целевая зависимость $y*$**
  - отображение $y* : X \to Y$
  - значения $y*$ известны только на объектах конечной обучающей выборки X^n
  - $X^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- **Требуется построить алгоритм a, который приближал бы неизвестную целевую зависимость как на элементах выборки X^n, так и на всём множестве X.**

Функционал качества и функция потерь

- **Вводится функция потерь $L(y, y')$, характеризующая величину отклонения ответа $y = a(x)$ от правильного ответа y' на произвольном объекте $x \in X$.**
- **Типичный выбор функции потерь:**
  - В задачах классификации $L(y, y') = [y \neq y']$ (т.е. число ошибок классификации);
  - В задачах регрессии $L(y, y') = (y - y')^2$.
- **Функционал качества**
  - характеризует среднюю ошибку (эмпирический риск) алгоритма a на произвольной выборке X^n
  - $Q(a, X^n) = \dfrac{1}{n} \sum_{i=1}^{n} L(a(x_i), y^*(x_i))$.
- **Метод минимизации эмпирического риска**
  - Требуется найти алгоритм a^*, минимизирующий среднюю ошибку на обучающей выборке:
  - $a^* = \arg\min_{a \in A} Q(a, X^n)$.

Признаки

Типы признаков

"бинарный" признак: ;

"номинальный" признак: — конечное множество;

"порядковый" признак: — конечное упорядоченное множество;

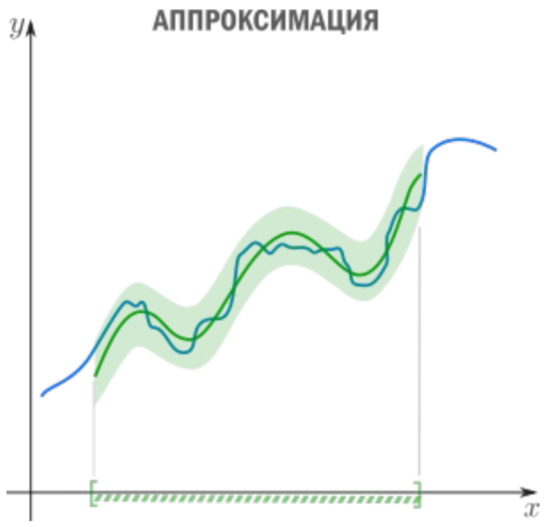
"количественный" признак: — множество действительных чисел.

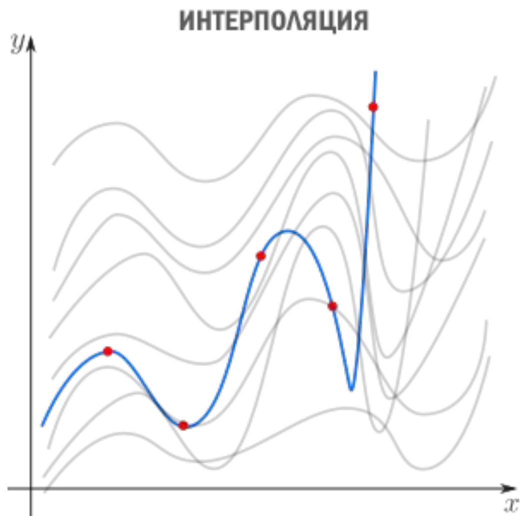
Регрессия?

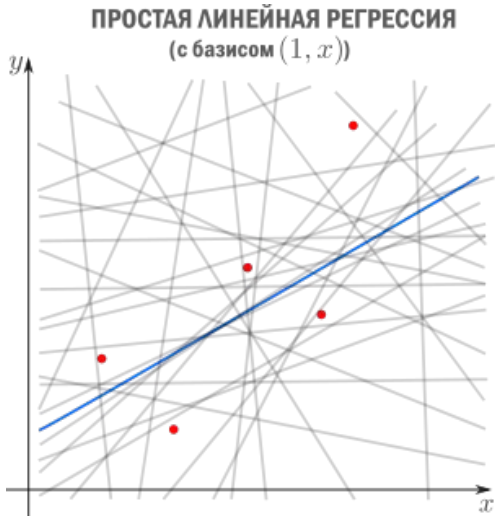
Интерполяция — способ выбрать из семейства функций ту, которая проходит через заданные точки.

Аппроксимация — способ выбрать из семейства «простых» функций приближение для «сложной» функции на отрезке, при этом ошибка не должна превышать определенного предела.

Регрессия — способ выбрать из семейства функций ту, которая минимизирует функцию потерь. Последняя характеризует насколько сильно пробная функция отклоняется от значений в заданных точках.







Регрессия

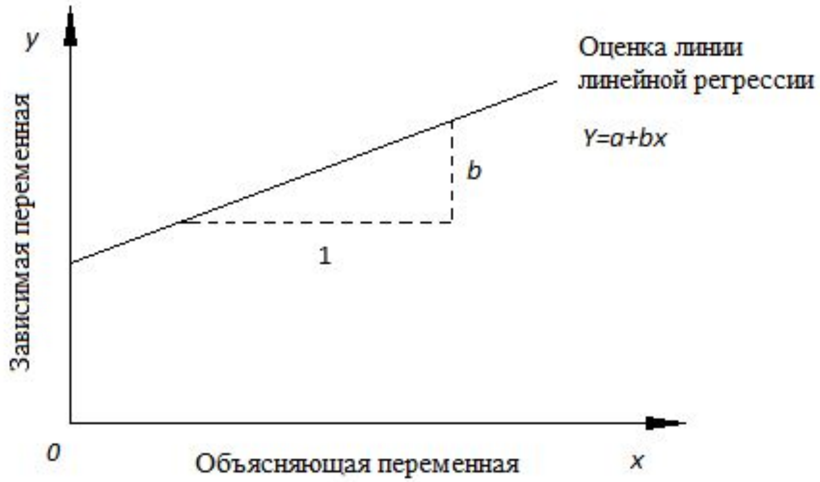
Уравнение, которое оценивает линию простой линейной регрессии:

$$Y=a+bx.$$

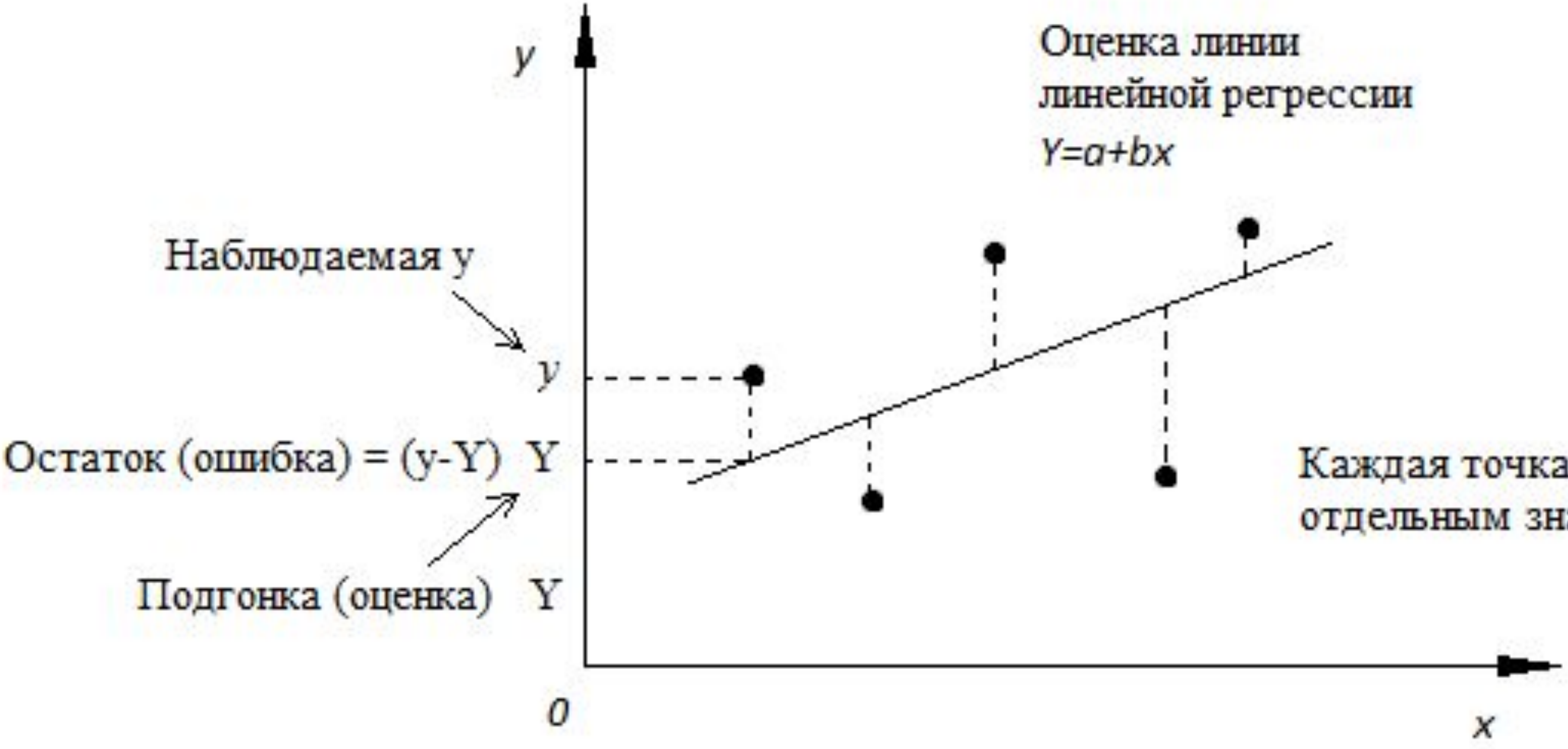
x называется независимой переменной или предиктором.

Y – зависимая переменная или переменная отклика.

- *a* – свободный член (пересечение) линии оценки; это значение *Y*, когда *x=0* .
- *b* – угловой коэффициент или градиент оценённой линии; она представляет собой величину, на которую *Y* увеличивается в среднем, если мы увеличиваем *x* на одну единицу.
- *a* и *b* называют коэффициентами регрессии оценённой линии, хотя этот термин часто используют только для *b*.

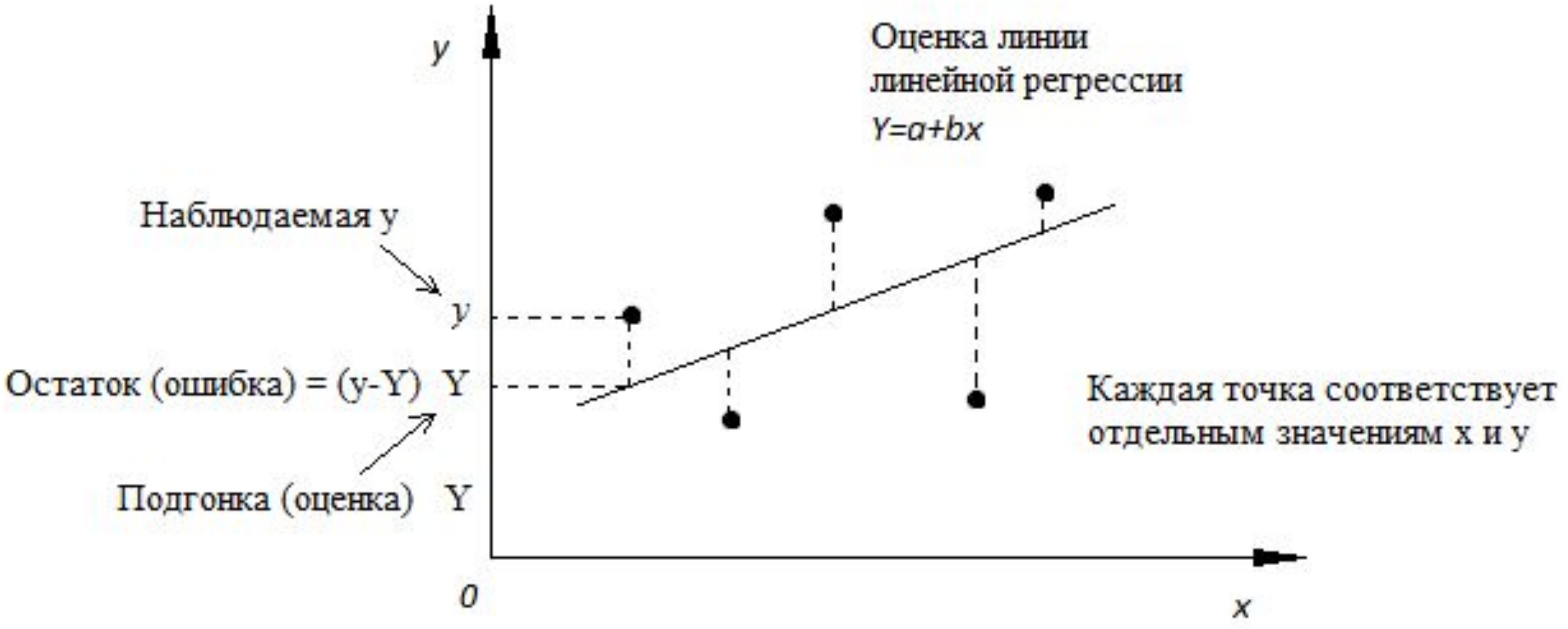


Метод наименьших квадратов



Метод наименьших квадратов

$$\mathrm{SSE}(a,b) = \mathrm{SS}_{res[iduals]} = \sum_{i=1}^{N} \text{отклонение}_i{}^2 = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - a - b \cdot x_i)^2,$$



Метод наименьших квадратов

$$\mathrm{SSE}(a,b) = \mathrm{SS}_{res[iduals]} = \sum_{i=1}^{N} \text{отклонение}_i{}^2 = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - a - b\cdot x_i)^2,$$

$$\frac{\partial}{\partial a}\mathrm{SSE}(a,b) = -2\sum_{i=1}^{N}(y_i - a - bx_i),$$

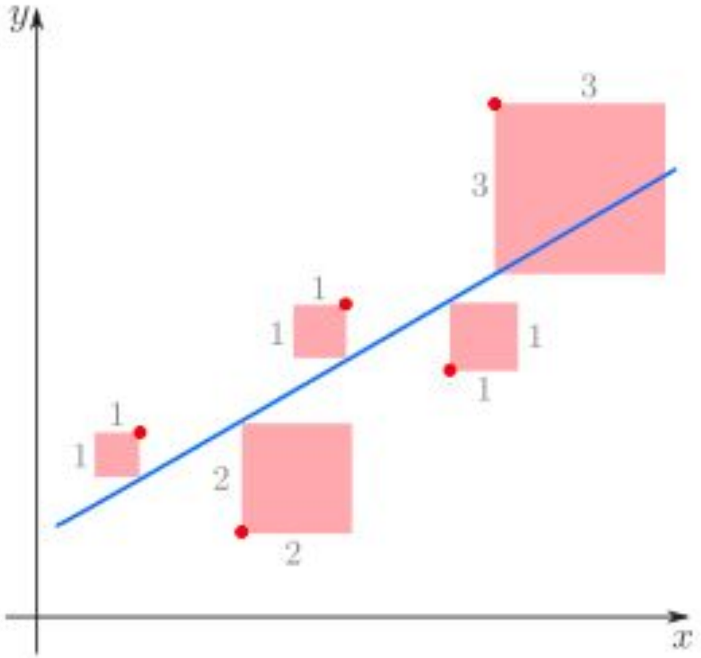
$$\frac{\partial}{\partial b}\mathrm{SSE}(a,b) = -2\sum_{i=1}^{N}(y_i - a - bx_i)x_i.$$

$$0 = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i),$$

$$0 = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i)x_i,$$

$$\hat{a} = \frac{\sum_i y_i}{N} - \hat{b}\frac{\sum_i x_i}{N},$$

$$\hat{b} = \frac{\frac{\sum_i x_i y_i}{N} - \frac{\sum_i x_i \sum_i y_i}{N^2}}{\frac{\sum_i x_i^2}{N} - \left(\frac{\sum_i x_i^2}{N}\right)^2}.$$



Мультиномиальная регрессия

- **Дано**
  - $X = x^{(1)}, \ldots, x^{(N)}$
  - $x^{(k)}$ - вектор размерности D, поэтому размерность $X : (N \times D)$
  - $Y = y^{(1)}, \ldots, y^{(N)}; y^{(k)} \in R$

- **Для некоторого примера $x^{(k)}$:**
  - $a(x^{(k)}) = \hat{y}^{(k)} = w_1 x_1^{(k)} + w_2 x_2^{(k)} + \ldots + w_D x_D^{(k)} = w^T x^{(k)}$

- **Целевая функция:**
  - $$L = \sum_{i}^{N} (y_i - \hat{y}_i)^2$$

- **Найти:** $w = ?$

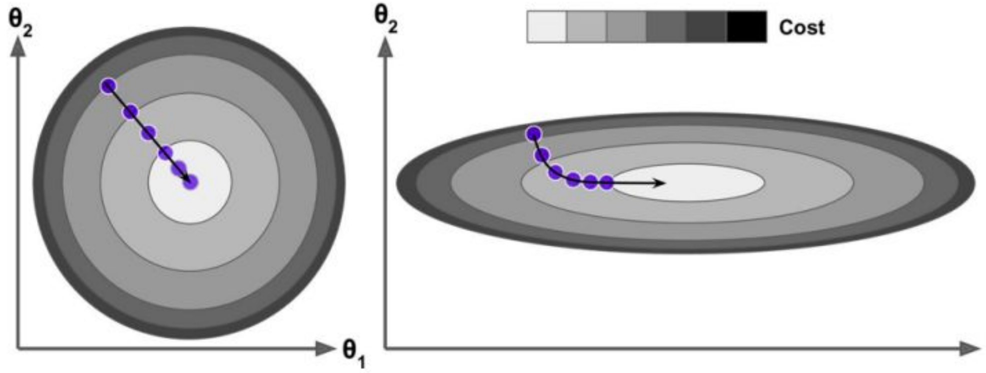
Мультиномиальная регрессия

- $\frac{\partial L}{\partial w_j} = 0$
- $\frac{\partial L}{\partial w_j} = \sum_i 2(y_i - w^T x_i)(-\frac{\partial (w^T x_i)}{\partial w_j})$
- $\frac{\partial w^T x_i}{\partial w_j} = x_{ij}$
- $\frac{\partial L}{\partial w_j} = \sum_i 2(y_i - w^T x_i)(-x_{ij}) = 0$
- $\frac{\partial L}{\partial w_j} = \sum_i 2(y_i - w^T x_i)(-x_{ij}) = 0$
- $\sum_i y_i x_{ij} = \sum_i (w^T x_i) x_{ij}$
- $x_{:,j}^T y = w^T (X^T x_{:,j})$ — **при фиксированном** j
- **для всех** j **сразу:** $X^T y = w^T (X^T X)$
- $(w^T (X^T X))^T = (X^T X) w$
- $X^T y = (X^T X) w$
- **решение:** $w = (X^T X)^{-1} X^T y$

$$X^+ = (X^\top X)^{-1} X^\top$$

Метод градиентного спуска

- $\mathbf{w}^{(next)} = \mathbf{w} - \eta \nabla_w MSE(\mathbf{w})$
- $\nabla_w MSE(\mathbf{w}) = (\frac{\partial(MSE(\mathbf{w}))}{\partial w_1}, \frac{\partial(MSE(\mathbf{w}))}{\partial w_2}, \ldots, \frac{\partial(MSE(\mathbf{w}))}{\partial w_D})^T$
- $\frac{\partial(MSE(\mathbf{w}))}{\partial w_j} = \frac{2}{n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)}) x_j^{(i)}$



Мультиномиальная регрессия

- **Дано**
  - $X = x^{(1)}, \ldots, x^{(N)}$
  - $x^{(k)}$ - вектор размерности D, поэтому размерность $X : (N \times D)$
  - $Y = y^{(1)}, \ldots, y^{(N)}; y^{(k)} \in R$

- **Для некоторого примера $x^{(k)}$:**
  - $a(x^{(k)}) = \hat{y}^{(k)} = w_1 x_1^{(k)} + w_2 x_2^{(k)} + \ldots + w_D x_D^{(k)} = w^T x^{(k)}$

- **Целевая функция:**
  - $L = \sum_{i}^{N} (y_i - \hat{y}_i)^2$

- **Найти: $w = ?$**

Кривые Обучения