Занятие 2. Обучение с учителем. Линейная регрессия

Машинное обучение и большие данные, 4 семестр

Иван Евгеньевич Бугаенко,
ассистент каф. ПМиФИ

Основные понятия в ML

x – объект, $\mathbb{X}$ – множество объектов ($x \in \mathbb{X} \subseteq \mathbb{R}^d$)

$y = y(x)$ – ответ (метка), $\mathbb{Y}$ – множество ответов ($y: \mathbb{X} \to \mathbb{Y}$)

$x^i = \left(x_1^i, x_2^i, \ldots, x_d^i\right)$ – признаковое описание x^i объекта (i-го объекта)

Если мы собрали M объектов ($i = \overline{1, M}$), то:

$$\chi = (x^i \quad y_i)_{i=1}^M = \begin{pmatrix} x_1^1 & x_2^1 & \ldots & x_d^1 & y_1 \\ x_1^2 & x_2^2 & \ldots & x_d^2 & y_2 \\ x_1^3 & x_2^3 & \ldots & x_d^3 & y_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1^M & x_2^M & \ldots & x_d^M & y_M \end{pmatrix}$$ – датасет (обучающая выборка)

Классы задач обучения с учителем

Если $\mathbb{Y} = Y \subseteq \mathbb{R}^n$, то задача поиска алгоритма $a: \mathbb{X} \to \mathbb{Y}$ называется ***задачей регрессии***

Если $|\mathbb{Y}| = \text{const}$, то задача поиска алгоритма $a: \mathbb{X} \to \mathbb{Y}$ называется ***задачей классификации***

Что такое задача обучения с учителем?

$x \in \mathbb{X} \subseteq \mathbb{R}^d \quad | \quad y: \mathbb{X} \to \mathbb{Y},\ y = y(x) \quad | \quad \chi = (x^i \quad y_i)_{i=1}^M$

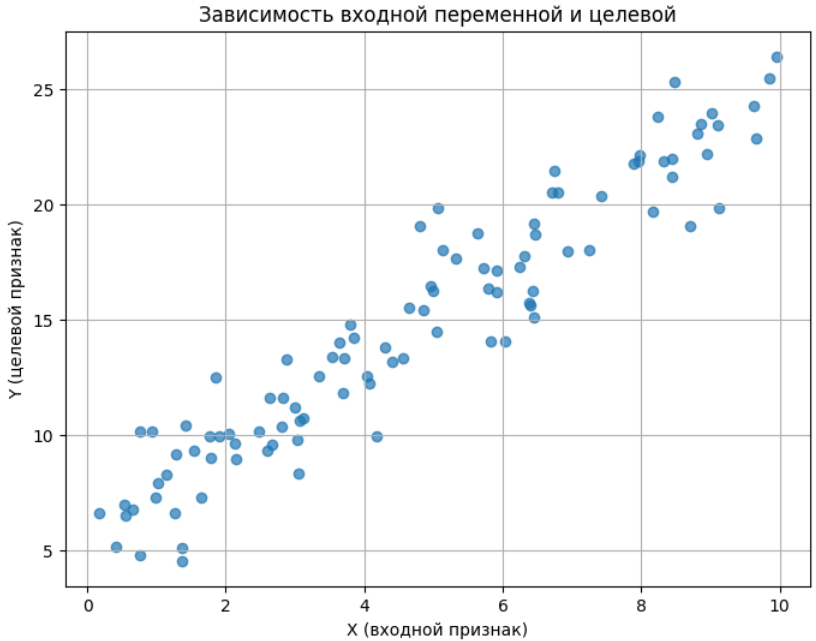
Вводятся следующие объекты:

- $a: \mathbb{X} \to \mathbb{Y}$ – алгоритм, $\hat{y} = a(x),\ \ \hat{y} \in \mathbb{Y}$
- $L\big(a(x), y(x)\big) = L(\hat{y}, y)$ – функция потерь/ошибки (loss function)
- $Q(a, \chi) = \sum_{x \in \chi} L\big(a(x), y(x)\big)$ – ошибка алгоритма a на реализации выборки χ

Задача ML: необходимо найти такой $a(x)$, что $Q(a, \chi) \to \min$

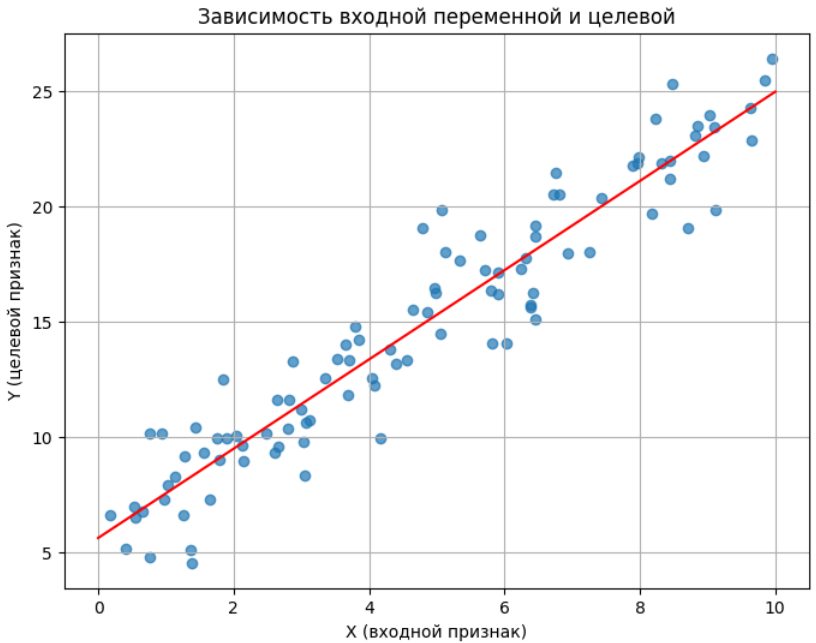
В машинном обучении различается большое количество ***семейств алгоритмов***

Линейная регрессия



Линейная регрессия

$$a(x^j) = \omega_0 + \omega_1 x_1^j$$



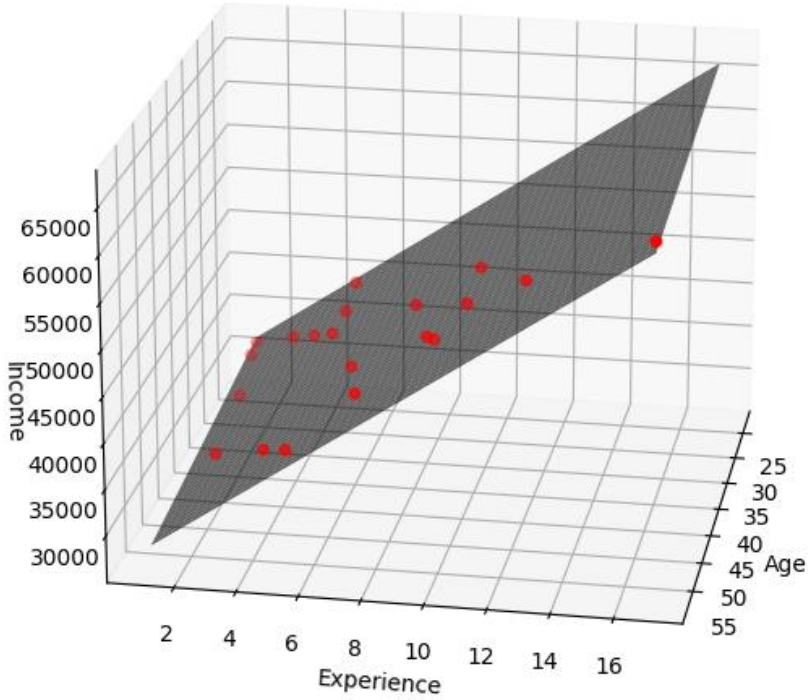
Задача регрессии. Линейные модели

Простейшим и базовым семейством алгоритмов являются ***линейные модели***:

$$a(x^j) = \omega_0 + \omega_1 x_1^j + \cdots + \omega_d x_d^j = \omega_0 + \sum_{i=1}^{d} \omega_i x_i^j ,$$

где ω_0 – свободный коэффициент (смещение, bias), x_i^j – i-ый признак j-го объекта, ω_i – вес i-го признака

Многомерная линейная регрессия



$$a(x^j) = \omega_0 + \omega_1 x_1^j + \omega_2 x_2^j$$