

# Обучение с учителем

Санкт-Петербургский государственный университет  
Кафедра статистического моделирования

16 сентября 2025, Санкт-Петербург

**Машинное обучение** — это раздел искусственного интеллекта, в котором разрабатываются методы и алгоритмы, позволяющие компьютерам обнаруживать закономерности в данных и делать прогнозы без явных инструкций.

**Обучение с учителем** — один из способов машинного обучения, в ходе которого для каждого примера в обучающем наборе известно, какой результат является правильным.

**Пример задач:**

- Регрессия: предсказание стоимости недвижимости, количества продаж некоторого товара, погоды.
- Классификация: предсказание ценовой категории товара, типа изображения, болеет ли человек или нет.

Дано:

- ❶ Пространство объектов  $X$  — множество описаний объектов (например, фотографии, тексты, таблицы с признаками).
- ❷ Пространство ответов  $Y$  — множество меток или значений, которые нужно предсказывать (например, классы «кот»/«собака» или числовые значения цен).
- ❸ Обучающая выборка  $D = \{(x_i, y_i)\}_{i=1}^n$ , где
  - $x_i \in X$  — набор значений признаков (регрессоров).
  - $y_i \in Y$  — целевая переменная или метка, которую мы хотим научиться предсказывать.
  - $n$  — количество индивидов (записей) в выборке.

Задача: построить такую функцию (модель)

$$a : X \longrightarrow Y,$$

чтобы ее предсказания  $\hat{y} = a(x)$  были как можно ближе к истинным ответам  $y$ .

Чтобы оценить, насколько хорошо модель предсказывает ответы, используется **функция потерь**  $L(y, \hat{y})$ . Она показывает, насколько велико расхождение между истинными значениями  $y$  и его предсказаниями  $\hat{y}$ .

**Примеры:**

- Для задачи **регрессии** наиболее распространенной функцией потерь является среднеквадратичная ошибка (MSE):

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- Для задачи **бинарной классификации**, если  $\hat{y}$  представляет собой вектор вероятностей принадлежности к положительному классу, используется кросс-энтропия:

$$L(y, \hat{y}) = - \sum_{i=1}^n [y_i \ln \hat{y}_i + (1 - y_i) \ln(1 - \hat{y}_i)].$$

Пусть целевая переменная  $y$  принимает значения  $\{-1, 1\}$ . Хотим обучить линейную модель так, чтобы плоскость, которую она задает, как можно лучше отделяла объекты одного класса от другого.

Линейный классификатор:

$$\hat{y} = a(x; w) = \text{sign}\langle x, w \rangle.$$

Функция потерь:

$$L(y, \hat{y}) = \sum_{i=1}^n \mathbb{I}[y_i \langle x_i, w \rangle < 0] \longrightarrow \min_w .$$

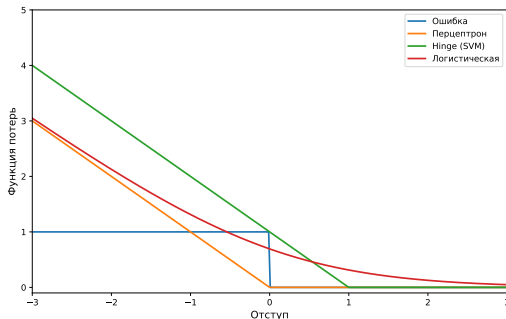
Величина  $M_i = y_i \langle x_i, w \rangle$  называется **отступом** (margin) классификатора. Абсолютная величина отступа говорит о степени уверенности классификатора.

**Проблема:** функция  $\mathbb{I}[M < 0]$  кусочно-постоянная, следовательно функцию потерь невозможно оптимизировать градиентными методами, поскольку во всех точках производная равна нулю.

**Решение:** можно мажорировать эту функцию более гладкой функцией и минимизировать функцию потерь с этой мажорирующей функцией.

# Линейная классификация. Функции потерь

- 1 Перцептрон:  $L(M) = \max(0, -M)$  — отступы учитываются только для неправильно классифицированных объектов пропорционально величине отступа.
- 2 Hinge (SVM):  $L(M) = \max(0, 1 - M)$  — объекты, которые классифицированы правильно, но не очень «уверенно», продолжают вносить свой вклад в градиент.
- 3 Логистическая:  $L(M) = \ln(1 + e^{-M})$ .



# Логистическая регрессия

Посмотрим на задачу классификации как на задачу предсказания вероятностей (например, предсказание «кликабельности» рекламного баннера).

**Принцип работы:** научить линейную модель предсказывать значения  $z \in \mathbb{R}$  (логиты), а затем преобразовывать их в вероятности с помощью сигмоиды:

$$z_i = \langle x_i, w \rangle = \ln \frac{p_i}{1 - p_i}, \quad p_i = \frac{1}{1 + e^{-\langle x_i, w \rangle}} = \sigma(\langle x_i, w \rangle).$$

Функция правдоподобия для распределения Бернулли:

$$p(y \mid \mathbf{X}, w) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}.$$

Прологарифмируем:

$$\sum_{i=1}^n \left[ y_i \ln(\sigma(\langle x_i, w \rangle)) + (1 - y_i) \ln(1 - \sigma(\langle x_i, w \rangle)) \right].$$



Теперь пусть  $y \in \{-1, 1\}$ . Тогда, поскольку  $\sigma(z) = 1 - \sigma(-z)$ , логарифм правдоподобия можно представить в следующем виде:

$$\begin{aligned}\ln p(y \mid \mathbf{X}, w) &= - \sum_{i=1}^n \left[ \mathbb{I}[y_i = 1] \sigma(z_i) + \mathbb{I}[y_i = -1] (1 - \sigma(z_i)) \right] \\ &= - \sum_{i=1}^n \ln \sigma(y_i \langle x_i, w \rangle) \\ &= \sum_{i=1}^n \ln (1 + e^{-M})\end{aligned}$$

Таким образом, функцию потерь в логистической регрессии можно представить в виде функции от отступа.