Постановка задачи машинного обучения

Задача обучения по прецедентам

X — множество *объектов*;
Y — множество *ответов*;
$y\colon X \to Y$ — неизвестная зависимость (target function).

Дано:
$\{x_1, \dots, x_\ell\} \subset X$ — *обучающая выборка* (training sample);
$y_i = y(x_i),\ \ i = 1, \dots, \ell$ — известные ответы.

Найти:
$a\colon X \to Y$ — алгоритм, решающую функцию (decision function), приближающую y на всём множестве X.

Как задаются ответы. Признаковое описание

$f_j\colon X \to D_j,\ \ j = 1, \dots, n$ — признаки объектов (features).

Типы признаков:

- $D_j = \{0, 1\}$ — *бинарный* признак f_j;
- $|D_j| < \infty$ — *номинальный* признак f_j;
- $|D_j| < \infty$, D_j упорядочено — *порядковый* признак f_j;
- $D_j = \mathbb{R}$ — *количественный* признак f_j.

Как задаются ответы. Признаковое описание

$f_j \colon X \to D_j$, $j = 1, \dots, n$ — признаки объектов (features).

Типы признаков:

- $D_j = \{0, 1\}$ — *бинарный* признак f_j;
- $|D_j| < \infty$ — *номинальный* признак f_j;
- $|D_j| < \infty$, D_j упорядочено — *порядковый* признак f_j;
- $D_j = \mathbb{R}$ — *количественный* признак f_j.

Вектор $\bigl(f_1(x), \dots, f_n(x)\bigr)$ — *признаковое описание* объекта x.

Матрица *«объекты–признаки»* (feature data)

$$F = \bigl\| f_j(x_i) \bigr\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

Как задаются ответы. Типы задач

Задачи классификации (classification):

- $Y = \{-1, +1\}$ — *классификация* на 2 класса.
- $Y = \{1, \dots, M\}$ — на M непересекающихся классов.
- $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться.

Задачи восстановления регрессии (regression):

- $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$.

Задачи ранжирования (ranking, learning to rank):

- Y — конечное упорядоченное множество.

Предсказательная модель

Модель (predictive model) — параметрическое семейство функций

$$A = \{a(x) = g(x, \theta) \mid \theta \in \Theta\},$$

где $g\colon X \times \Theta \to Y$ — фиксированная функция,
Θ — множество допустимых значений параметра θ.

Пример.
Линейная модель с вектором параметров $\theta = (\theta_1, \dots, \theta_n)$, $\Theta = \mathbb{R}^n$:

$g(x, \theta) = \sum\limits_{j=1}^{n} \theta_j f_j(x)$ — для регрессии и ранжирования, $Y = \mathbb{R}$;

$g(x, \theta) = \operatorname{sign} \sum\limits_{j=1}^{n} \theta_j f_j(x)$ — для классификации, $Y = \{-1, +1\}$.

Этапы обучения и применения модели

Этап обучения (train):
Метод обучения (learning algorithm) $\mu\colon (X \times Y)^\ell \to A$
по выборке $X^\ell = (x_i, y_i)_{i=1}^{\ell}$ строит алгоритм $a = \mu(X^\ell)$:

$$\boxed{\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_1(x_\ell) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}} \xrightarrow{\mu} a$$

Этап применения (test):
алгоритм a для новых объектов $x'_1, \dots, x'_k$ выдаёт ответы $a(x'_i)$.

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_k) \end{pmatrix}$$

Функционалы качества

$\mathscr{L}(a, x)$ — *функция потерь* (loss function) — величина ошибки алгоритма $a \in A$ на объекте $x \in X$.

Функции потерь для задач классификации:

- $\mathscr{L}(a, x) = \big[a(x) \neq y(x)\big]$ — индикатор ошибки;

Функции потерь для задач регрессии:

- $\mathscr{L}(a, x) = \big|a(x) - y(x)\big|$ — абсолютное значение ошибки;
- $\mathscr{L}(a, x) = \big(a(x) - y(x)\big)^2$ — квадратичная ошибка.

Эмпирический риск — функционал качества алгоритма a на X^ℓ:

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}(a, x_i).$$

Сведение задачи обучения к задаче оптимизации

Минимизация эмпирического риска (empirical risk minimization):

$$\mu(X^\ell) = \arg\min_{a \in A} Q(a, X^\ell).$$

Пример: *метод наименьших квадратов* ($Y = \mathbb{R}$, $\mathscr{L}$ квадратична):

$$\mu(X^\ell) = \arg\min_{\theta} \sum_{i=1}^{\ell} \big(g(x_i, \theta) - y_i\big)^2.$$

Понятие обобщающей способности (generalization performance):

- найдём ли мы «закон природы» или *переобучимся*, то есть подгоним функцию $g(x_i, \theta)$ под заданные точки?
- будет ли $a = \mu(X^\ell)$ приближать функцию y на всём X?
- будет ли $Q(a, X^k)$ мало́ на новых данных — *контрольной выборке* $X^k = (x'_i, y'_i)_{i=1}^k$, $y'_i = y(x_i)$?