Формальная постановка задачи машинного обучения

Машинное обучение – "наука о том, как восстановить функцию по точкам".

Пусть:
X – множество объектов
Y – множество ответов
$y : X \to Y$ – неизвестная зависимость (target function)

Дано:
$\{x_1, ..., x_l\} \subset X$ – обучающая выборка (training sample)
$y_i = y(x_i),\ i = 1, ..., l$ – известные ответы

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

Признаковое описание объектов
$f_j : X \to D_j,\ j = 1, ..., n$ – признаки объектов (features)

$D_j = \{0, 1\}$ – 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

Представление обучающей выборки – матрица "объекты-признаки"(номера столбцов – номера признаков, номера строк – номера объектов)

Разновидности ответов – типы задач
$Y = \{-1, +1\}$ – задача **классификации** (classification) на 2 класса
$Y = \{1, ..., M\}$ – задача **классификации** на M непересекающихся классов
$Y = \{0, 1\}^M$ – задача **классификации** на M классов, которые могут пересекаться
$Y = \mathbb{R}$ или $Y = \mathbb{R}^m$ – задача **восстановления регрессии** (regression)
Y – конечное упорядоченное множество – задача **ранжирования** (ranking)

Предсказательная модель
Модель (predictive model) – параметрическое семейство функций:

$$A = \{a(x) = g(x, \theta) \mid \theta \in \Theta\} \tag{1}$$

Функционалы качества
Один из методов решения задач машинного обучения – сведение их к задачам оптимизации, т.е. выбор оптимального вектора θ через максимизацию точности предсказываемых ответов. Эту точность показывают функционалы качества.

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

Для задач классификации:
$\mathcal{L}(a, x) = [a(x) \neq y(x)]$ – индикатор ошибки

Для задач регрессии:
$\mathcal{L}(a, x) = |a(x) - y(x)|$ – абсолютное значение ошибки
$\mathcal{L}(a, x) = (a(x) - y(x))^2$ – квадратичная ошибка

Эмпирический риск – функционал качества алгоритма a на X^l

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} (a, x_i) \tag{2}$$

Решение задачи – минимизация эмпирического риска (empirical risk minimization):

$$\mu(X^l) = \arg\min_{a \in A} Q(a, X^l) \tag{3}$$

Переобучение

– это когда найденный алгоритм хорошо работает на обучающей выборке и плохо на тестовой.

Эмпирические оценки переобучения:

$HO(\mu, X^l, X^k) = Q(\mu(X^l), X^k)$ – Эмпирический риск на тестовых данных (hold-out)
$LOO(\mu, X^l) = \frac{1}{L} \sum_{i=1}^{L} \mathcal{L}(\mu(X^L \setminus \{x_i\}), x_i), \quad L = l + 1$ – Скользящий контроль (leave-one-out)
$CV(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} Q(\mu(X_n^l), X_n^k), \quad X^L = X_n^l \sqcup X_n^k, \quad L = l + k$ – Кросс-проверка (cross-validation) по N разбиениям

Выбор μ, для которого такая оценка минимальна, может снять состояние переобучения.