Расстояние Хэмминга (Hamming distance)

> Синонимы: Кодовое расстояние
> Разделы: Метрики

В теории информации и анализе данных расстояние Хэмминга между двумя строками или векторами одинаковой длины — это количество позиций, в которых соответствующие символы различны. Таким образом, для векторов $X = (x_1, x_2, \ldots, x_n)$ и $Y = (y_1, y_2, \ldots, y_n)$ оно может быть записано в виде:

$$d_H(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i),$$

где $\delta(x_i, y_j) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases}$

Оно измеряет минимальное количество перестановок, необходимое для замены одной строки на другую, или, что то же самое, минимальное количество ошибок, которые могли бы преобразовать одну строку в другую.

Справедливо, $d_H(X, Y) \leqslant n$, т.е. расстояние Хэмминга всегда меньше длины векторов (строк), между которыми оно измеряется.

В более общем контексте расстояние Хэмминга — это одна из нескольких строковых метрик для измерения расстояния редактирования между двумя последовательностями. Оно названо в честь американского математика Ричарда Хэмминга.

Например, пусть $X = (01011)$, а $Y = (11010)$, тогда $d_H(X, Y) = 2$.

С помощью расстояния Хемминга можно представлять степень близости друг к другу категориальных величин. Например, закодируем с помощью унитарного кода слова:

Слово	**Код**
Красный	100
Синий	010
Зеленый	001

Несложно увидеть, что расстояние между словами будет равно 2. Если для двух строк $d_H = 1$, то говорят, что они являются соседними.

Определение степени близости категориальных значений с помощью расстояния Хэмминга открывает возможность для их использования в машинном обучении, в частности, в алгоритмах классификации и кластеризации.