Информационная энтропия (Information entropy)

> Синонимы: Энтропия Шеннона
> Разделы: Метрики

В теории информации энтропия — это средняя скорость генерирования значений некоторым случайным источником данных. Величина информационной энтропии, связанная с определенным значением данных, вычисляется по формуле:

$$H = -\sum_{i=1}^{n} P_i log P_i,$$

где P_i — вероятность i-го состояния системы (значения принимаемого переменной), n — число состояний системы (значений, принимаемых переменной).

Когда источник данных генерирует значение, имеющее низкую вероятность (т.е. когда происходит маловероятное, неожиданное событие), с ним связана большая информация, чем с более вероятным событием. Количество информации, выражаемое событием, связанным с появлением определенного значения данных, можно рассматривать как случайную переменную, математические ожидание которой и равно информационной энтропии.

Таким образом, информационную энтропию можно рассматривать как меру неупорядоченности или неопределенности состояния некоторой системы, описываемой данными. В этом смысле она является прямым аналогом понятия энтропии, используемой в статистической термодинамике. Впервые понятие энтропии как меры информации было введено К. Шенноном в 1948 г.

Энтропия измеряется в битах, натах (natural units) или дитах (десятичных числах) в зависимости от основания логарифма, используемого при вычислении энтропии. Логарифм используется в связи с тем, что он аддитивен для независимых источников. Например, энтропия события — броска монеты равна 1 биту, а энтропия m бросков составит m бит.

В обычном представлении $log_2 n$ бит требуется для представления переменной, принимающей n значений, если n является степенью 2. Если значения равновероятны, то энтропия в битах будет равна значению n.

Если появление одного из значений более вероятно, чем других, то наблюдение, в котором это значение появляется, можно интерпретировать как менее информативное, чем то, в котором появляется более редкое значение.

В анализе данных и машинном обучении энтропия используется в алгоритмах классификации как мера классовой однородности подмножеств наблюдений, полученных в результате разбиения обучающего множества на классы. Чем выше однородность подмножества, т.е. чем больше примеров одного класса и меньше «примесь» примеров других классов, тем меньше энтропия и тем лучше результаты классификации.

Если все примеры, попавшие в подмножество, относятся к одному классу (т.е. вероятность в результате случайного выбора получить именно данный класс равна 1), то энтропия равна 0. Это очевидно, поскольку $log(1) = 0$ (см. формулу). В этом случае подмножество однозначно ассоциируется с классом.

Наихудший с точки зрения классификации случай, когда классы представлены в равных пропорциях и оказываются равновероятными. Тогда неопределенность выбора, а следовательно, и энтропия множества максимальны.

Таким образом, при решении задачи классификации, в процессе которой происходит последовательное разбиения исходного набора данных на подмножество по некоторому признаку., лучшим будет такое разбиение, которое обеспечит минимальную энтропию (или максимальную однородность по классам) результирующих подмножеств.

Данный подход известен как критерий прироста информации (или, что одно и то же, уменьшения энтропии) и используется в популярных алгоритмах построения деревьев решений, таких как ID3 и C4.5.