Нормализация данных (Data normalization)

> Синонимы: Нормирование данных, Нормировка данных
> Разделы: Алгоритмы

В машинном обучении нормализацией называют метод предобработки числовых признаков в обучающих наборах данных с целью приведения их к некоторой общей шкале без потери информации о различии диапазонов.

Иногда нормализацию данных называют стандартизацией, однако это неверно. Стандартизация это более широкое понятие и подразумевает предобработку с целью приведению данных к единому формату и представлению, наиболее удобному для использования определенного вида обработки. В отличии от нормализации, стандартизация может применяться и к категориальным данным.

Необходимость нормализации вызвана тем, что разные признаки обучающего набора данных могут быть представлены в разных масштабах и изменяться в разных диапазонах. Например, возраст, который изменяется от 0 до 100, и доход, изменяющийся от нескольких тысяч до нескольких миллионов. То есть диапазоны изменения признаков «Возраст» и «Доход» различаются в тысячи раз.

В этом случае возникает нарушение баланса между влиянием входных переменных, представленных в разных масштабах, на выходную переменную. Т.е. это влияние обусловлено не реальной зависимостью, а изменением масштаба. В результате, обучаемая модель может выявить некорректные зависимости.

Существует несколько основных методов нормализации.

Десятичное масштабирование (decimal scaling)

В данном методе нормализация производится путем перемещения десятичной точки на число разрядов, соответствующее порядку числа: $x_i^{'} = x_i/10^n$, где n — число разрядов в наибольшем наблюдаемом значении. Например, пусть имеется набор значений: -10, 201, 301, -401, 501, 601, 701. Поскольку n=3, то получим $x_i^{'} = x_i/10^3$. Иными словами, каждое наблюдаемое значение делим на 1000 и получаем: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701.

Минимаксная нормализация

Несложно увидеть недостаток предыдущего метода: результирующие значения всегда будут занимать не весь диапазон [0,1], а только его часть, в зависимости от наибольшего и наименьшего наблюдаемых значений. Если исходный диапазон мал (скажем, 400 — 500), то получим, что в результате десятичного масштабирование нормализованные значения будут лежать в диапазоне [0.4,0.5], т.е. его изменчивость окажется очень низкой, что плохо сказывается на качестве построенной модели.

Решить проблему можно путем применения минимаксной нормализации, которая реализуется по формуле:

$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$.

Эту формулу можно обобщить на приведение исходного набора значений к произвольному диапазону $[a, b]$:

$X' = a + \frac{X - X_{min}}{X_{max} - X_{min}}(b - a)$.

Наиболее часто используется приведение к диапазонам [0,1] и [-1,1]

Нормализация средним (Z-нормализация)

Недостаткам минимаксной нормализации является наличие аномальных значений данных, которые «растягивают» диапазон, что приводит к тому, что нормализованные значения опять же концентрируются в некотором узком диапазоне вблизи нуля. Чтобы избежать этого, следует определять диапазон не с помощью максимальных и минимальных значений, а с помощью «типичных» — среднего и дисперсии:

$x'_i = (x_i - \overline{X})/\sigma_x$.

Величины, полученные по данной формуле, в статистике называют Z-оценками. Их Абсолютное значение представляет собой оценку (в единицах стандартного отклонения) расстояния между x и его средним значением $\overline{X}$ в общей совокупности. Если z меньше нуля, то x ниже средней, а если z больше нуля, то x выше средней.

Отношение

В этом методе каждое значение исходных данных делиться на некоторое, заданное пользователем число, или на значение статистического показателя, вычисленного по набору данных, например, среднее, стандартное отклонение, дисперсию, вариационный размах и др.