Лекция

Нормализация и стандартизация данных

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.

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 определённого вида обработки. В отличии от нормализации, стандартизация может применяться и к категориальным данным.

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. Диапазоны изменения признаков «Возраст» и «Доход» различаются в тысячи раз.

Возникает нарушение баланса между влиянием входных переменных, представленных в разных масштабах, на выходную переменную.

«Сырые» данные имеют разный масштаб и разное распределение по каждому признаку. Например, какой-то химический показатель смеси может иметь значения в диапазоне от 0.0001 до 0.2, а другой показатель от –100 до 100.

Стандартизация данных – это процесс приведения вектора каждого признака к виду, при котором его математическое ожидание станет нулевым, а дисперсия – единичной.

Нормализация данных – это процесс масштабирования вектора каждого признака, при котором вектор будет иметь единичную норму (при этом есть разные способы оценки\подсчета нормы).

Существует несколько методов нормализации.

Десятичное масштабирование (decimal scaling)

В данном методе нормализация производится путём 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.

n=3, получим x′i=xi/103.

Каждое наблюдаемое значение делим на 1000 и получаем: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701.

Недостаток 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.

Минимаксная нормализация

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

Эту формулу можно обобщить на привидение 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.

Вектор L1 Норма

Длина вектора может быть вычислена с использованием нормы L1, где 1 - верхний индекс L, например, L ^ 1.

Обозначения для нормы L1 вектора: || v || 1, где 1 - индекс. Таким образом, эту длину иногда называют нормой такси или нормой Манхэттена.

Норма L1 рассчитывается как сумма абсолютных значений вектора, где абсолютное значение скаляра использует обозначение | a1 |.

Норма - это вычисление манхэттенского расстояния от начала векторного пространства.

Норма L1 вектора может быть вычислена в NumPy с помощью функции norm () с параметром для указания порядка нормы, в данном случае 1.

```
# l1 norm of a vector
from numpy import array
from numpy.linalg import norm
a = array([1, 2, 3])
print(a)
l1 = norm(a, 1)
```

```
print(l1)
```

определяется вектор 1 × 3, затем вычисляется норма вектора L1.

```
[1 2 3]
6.0
```

Норма L1 часто используется при подборе алгоритмов машинного обучения в качестве метода регуляризации, например метод, позволяющий сохранять коэффициенты модели малыми, и, в свою очередь, модель менее сложной.

Вектор L2 Норма

Длина вектора может быть вычислена с использованием нормы L2, где 2 - верхний индекс L, например, L ^ 2.

Норма L2 вычисляет расстояние векторной координаты от начала векторного пространства. Как таковая, она также известна как евклидова норма, поскольку она рассчитывается как евклидово расстояние от начала координат. Результатом является положительное значение расстояния. Норму L2 вектора можно рассчитать в NumPy с помощью функции norm () с параметрами по умолчанию.

```
# l2 norm of a vector
from numpy import array
from numpy.linalg import norm
a = array([1, 2, 3])
print(a)
l2 = norm(a)
print(l2)
```

Сначала определяется вектор 1 × 3, затем вычисляется норма вектора L2.

При выполнении примера сначала печатается определенный вектор, а затем норма L2 вектора.

```
[1 2 3]

3.74165738677
```

Как и норма L1, норма L2 часто используется при подборе алгоритмов машинного обучения в качестве метода регуляризации, например метод, позволяющий сохранять коэффициенты модели малыми и, в свою очередь, модель менее сложной.

Норма L2 чаще используется, чем другие векторные нормы в машинном обучении.

Вектор Макс Норм

Длина вектора может быть рассчитана с использованием максимальной нормы, также называемой максимальной нормой.

Максимальная норма вектора называется L ^ inf, где inf - верхний индекс и может быть представлен символом бесконечности. Обозначения для максимальной нормы: || x || inf, где inf - индекс.

```
maxnorm(v) = ||v||inf
```

Максимальная норма вычисляется как возвращающая максимальное значение вектора, отсюда и название.

```
||v||inf = max(|a1|, |a2|, |a3|)
```

Максимальная норма вектора может быть вычислена в NumPy с помощью функции norm () с параметром порядка, установленным в inf.

```
# max norm of a vector
from numpy import inf
from numpy import array
from numpy.linalg import norm
a = array([1, 2, 3])
print(a)
maxnorm = norm(a, inf)
print(maxnorm)
```

Сначала определяется вектор 1 × 3, затем вычисляется максимальная норма вектора.

При запуске примера сначала печатается определенный вектор, а затем максимальная норма вектора

```
[1 2 3]
3.0
```

Максимальная норма используется в качестве регуляризации в машинном обучении, например, в весах нейронных сетей, называемой максимальной нормализацией.