NORMALIZATION

SIEYE RYU

When we analyze a bunch of data, we need to adjust values on different scales quite often. In this article, we study standard score (standardization), min-max normalization, mean normalization, scaling to unit length and quantile normalization.

1. Standard Score (Standardization)

Suppose that X is a finite subset of $\mathbb{R}$ and that $(X, 2^X, p)$ is a probability space. Let μ and σ denote the mean and the standard deviation of X, respectively:

$$\mu = \frac{1}{|X|} \sum_{x \in X} x$$

and

$$\sigma = \sqrt{\sum_{x \in X} p(x)(x-\mu)^2}.$$

From here on, we assume that p is a descrete uniform distribution, that is, $p(x) = \frac{1}{|X|}$ for all $x \in X$.

We define a function $z : X \to \mathbb{R}$ by

$$z(x) = \frac{x-\mu}{\sigma}$$

and denote the set of $z(x)$ by Z:

$$Z = \{z : z = z(x) \; \forall \, x \in X\}.$$

We call $z(x)$ a *z-score* of x.

If $q : Z \to [0,1]$ is defined by $q(z) = \frac{1}{|Z|}$, then $(Z, 2^Z, q)$ is also a probablity space. The set Z has the following properties:
(1) The mean of Z is 0.
(2) The standard deviation of Z is 1.
The first property can be verified by the following:

$$\frac{1}{|Z|} \sum_{z \in Z} z = \frac{1}{|X|} \sum_{x \in X} \frac{x-\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

and the second property follows from the following:

$$\sqrt{\sum_{z \in Z} q(z)(z-0)^2} = \sqrt{\frac{1}{|Z|} \sum_{z \in Z} z^2} = \sqrt{\frac{1}{|X|} \sum_{x \in X} \left(\frac{x-\mu}{\sigma}\right)^2} = 1.$$

The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1. In our case, if p is a normal distribution, then q becomes a standard normal distribution. When p is a normal distribution we can find the interval $[a, b]$ such that

$$p(a < x < b) = t \tag{1.1}$$

for any $t \in [0, 1]$. We first note that (1.1) is equivalent to

$$q\left(\frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}\right) = t.$$

Using standard deviation table, we can find the interval $[-c, c]$ such that

$$q(-c < z < c) = t.$$

Since $-c = \frac{a-\mu}{\sigma}$ and $c = \frac{b-\mu}{\sigma}$, it follows that $a = \mu - c\sigma$ and $b = \mu + c\sigma$.

It seems like data scientists often standardize data regardless of whether it is normally distributed. Apparently, standardization handles outliers so the prediction would not be affected by outliers much.

2. Min-Max Normalization

When we have several features which have distinct scales, we often need to make them have the same scales. Obviously, when all the features are equally important, our prediction would be more precise. Min-Max normalization rescales a range of values. For instance, when we have two features X_1 and X_2 such that

$$\min(X_1) = 0, \quad \max(X_1) = 0.1, \quad \min(X_2) = -1000, \quad \max(X_2) = 1500$$

we can rescale ranges so that the resulting features X_1' and X_2' satisfying

$$\min(X_1') = a, \quad \max(X_1') = b, \quad \min(X_2') = a, \quad \max(X_2') = b$$

for any real numbers a and b with $a < b$. Usually, $[a, b]$ is either $[0, 1]$ or $[-1, 1]$. We investigate how to convert the range into $[0, 1]$ and into an arbitrary interval $[a, b]$ for any $a < b$.

Suppose that X is a finite subset of $\mathbb{R}$. We denote the maximum and minimum of X by $X_{\max}$ and $X_{\min}$, respectively.

Let $x' : X \to \mathbb{R}$ be a map defined by

$$x'(x) = \frac{x - X_{\min}}{X_{\max} - X_{\min}}.$$

We denote the set of $x'(x)$ for $x \in X$ by X':

$$X' = \{x' : x' = x'(x) \ \forall x \in X\}.$$

Obvioiusly,

$$\max(X') = 1 \qquad \text{and} \qquad \min(X') = 0.$$

Let $x'' : X \to \mathbb{R}$ be a map defined by

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

We denote the set of $x''(x)$ for $x \in X$ by X'':

$$X'' = \{x'' : x'' = x''(x) \ \forall x \in X\}.$$

Obvioiusly,

$$\max(X'') = b \qquad \text{and} \qquad \min(X'') = a.$$

3. Mean Normalization

Depending on data set, mean normalization can make the implementation work a little bit better. You can see some examples on 'Mean Normalization.ipynb' in 'very basic' folder.

Suppose that X is a finite subset of $\mathbb{R}$. We denote the maximum, minimum and mean of X by X_{max}, X_{min} and μ respectively.

The rescaling formula is given by

$$x \quad \mapsto \quad \frac{x-\mu}{X_{\text{max}}-X_{\text{min}}}.$$

4. Scaling to Unit Length

The components of a feature vector can be scaled so that the resulting vector has a unit length:

$$x \quad \mapsto \quad \frac{x}{||x||}.$$

5. Quantile Normalization

We start this section with an example. The following table indicates how many ingredients x, y, z and w are included in certain cells A, B and C. Suppose that the equipments for the observations were all different. In the table, we can find out the proportion of x, y, z and w of each cell but we cannot compare A, B and C. In this case, we can use a quantile normalization to compare A, B and C.

	A	B	C
x	5	4	3
y	2	1	3
z	3	4	4
w	4	3	9

The minimum values of each column $\text{Min}(A), \text{Min}(B), \text{Min}(C)$ are replaced with the average of them, that is, 2, 1 and 3 in each column are replaced with $\frac{2+1+3}{3} = 2$. The second lowest values of each column are replaced with the average of them, that is, 3, 3, 3 in each column are replaced with 3. However, the minimum and the second lowest value of C are the same. We cannot decide which is lower. In this case, we replace 3 in column C with the mean of the replacements, that is, 5/2. Continuing this process, we obtain the following table.

	A	B	C
x	6	5	5/2
y	2	2	5/2
z	3	5	4
w	4	3	6

Now we calculate quantiles of two tables using linear interpolation. We adopt the default option of Pandas' quantile. (I hope I am correct!) We calculate the first quantile of A in the second table. Since the column A has four components, we need to consider the quantile of 5, that is, $\frac{1}{4} \times 5 = 1.25$. We denote a function defined by

$$(A, n) \quad \mapsto \quad \text{the n-th smallest component of A} \qquad (n = 1, 2, 3, 4)$$

by $\mathrm{Ord}(A, n)$ The $\lfloor 1.25 \rfloor$-th and the $\lceil 1.25 \rceil$-th smallest components of A are 2 and 3, respectively. So,

$$\mathrm{Ord}(A, \lfloor 1.25 \rfloor) = 2 \qquad \text{and} \qquad \mathrm{Ord}(A, \lceil 1.25 \rceil) = 3.$$

$\mathrm{Ord}(A, \lceil 1.25 \rceil) - [[\mathrm{Ord}(A, \lceil 1.25 \rceil) - \mathrm{Ord}(A, \lfloor 1.25 \rfloor)] \times (1.25 - \lfloor 1.25 \rfloor)] = 2.75$ Continuing this process, we have

	A	B	C
mean	3.75	3.75	3.75
std	1.71	1.5	1.66
min	2	2	2.5
25%	2.75	2.75	2.5
50%	3.5	4	3.25
75%	4.5	5	4.5
max	6	5	6

The quantiles of the original data are as follows.

	A	B	C
mean	3.5	3	4.75
std	1.29	1.41	2.87
min	2	1	3
25%	2.75	2.5	3
50%	3.5	3.5	3.5
75%	4.25	4	5.25
max	5	4	9

After quantile normalization, the differences between the quantiles of A, B and C become smaller.

References

[1] Normalization (statistics) on Wikipedia

[2] Feature scaling on Wikipedia

[3] Andrew Ng, Machine Learning, Stanford University https://www.coursera.org/lecture/machine-learning/implementational-detail-mean-normalization-Adk8G

[4] Quantile normalization on wikipedia