Udemy

Normalleştirme ile standartlaştırma arasındaki fark?

Normalleştirme veriyi belli bir aralığa indirgiyor. Fakat standartlaştırma veriyi bozmadan ortalama değerden ne kadar bozulmuş onu bize gösteriyor.

**İstatistiksel Normalleştirme (Statistical Normalisation)**

İstatistiksel normalleştirme, özellikle, veri madenciliği (data mining) gibi bilgisayar bilimlerinin istatistiksel veri işleme alanlarında kullanılan bir yöntemdir. Yöntemin amacı, veriler arasında farklılığın çok fazla olduğu durumlarda verileri tek bir düzen içerisinde ele almaktır.

Diğer bir kullanılışı ise farklı ölçekleme sisteminde bulunan verilerin birbiri ile karşılaştırılabilmesidir. Buradaki amaç, matematiksel fonksiyonlar kullanarak, farklı sistemlerde bulunan verileri, ortak bir sisteme taşımak ve karşılaştırılabilir hale getirmektir.

Bu anlamda kullanılan normalleştirme fonksiyonları (normalization functions) aşağıda açıklanmıştır.

**1.    Asgari – Azami Normalleştirmesi (Min-Max Normalisation)**

Bu yöntemde, bir grup verinin içerisindeki en büyük ve en küçük değerler ele alınır. Diğer bütün veriler, bu değerlere göre normalleştirilir. Buradaki amaç en küçük değeri 0 ve en büyük değeri 1 olacak şekilde normalleştirmek ve diğer bütün verileri bu 0-1 aralığına yaymaktır.

Yukarıda verilen formüle göre her değerin normalleştirilmiş değeri hesaplanır. Bu hesaplamanın çalışmasını bir örnek sayı dizisi üzerinde gösterelim. Örneğin sayı dizimi aşağıdaki şekilde verilmiş olsun:

5, 8,9,11,20, 22,24,25,27,29

Yukarıdaki sayıların normalleştirilmiş halleri aşağıda verilmiştir:

|  |  |
| --- | --- |
| 5 | 0 |
| 8 | 0,125 |
| 9 | 0,166666667 |
| 11 | 0,25 |
| 20 | 0,625 |
| 22 | 0,708333333 |
| 24 | 0,791666667 |
| 25 | 0,833333333 |
| 29 | 1 |

Yukarıda görüldüğü üzere en küçük sayı 0 ve en büyük sayı 1 olarak normalleştirilmiştir. Bu sayılar dışındaki sayılar ise yer aldıkları aralığa göre değerler atanır. Örneğin sayı dizimi aşağıdaki şekilde olsaydı:

10,15,16,17,50,51,55,56,60

Bu durumda normalleştirilmiş değerler aşağıdaki şekilde olacaktı:

|  |  |
| --- | --- |
| 10 | 0 |
| 15 | 0,1 |
| 16 | 0,12 |
| 17 | 0,14 |
| 50 | 0,8 |
| 51 | 0,82 |
| 55 | 0,9 |
| 56 | 0,92 |
| 60 | 1 |

Görüldüğü üzere bu yeni normalleştirme sonucunda da sayılarımız 0 ile 1 arasında değerler aldılar. Şimdi bu iki farklı sayı grubunu karşılaştırabiliriz. Örneğin ilk sayı dizimizde bulunan 25, normalleştirme sonucunda 0,83 değerini almıştır. İkinci sayı grubumuzda bulunan 51 ise, normalleştirilmiş değer olarak 0,82 değerini almıştır. Bu duruma ilk sayı dizisinde bulunan 25’in sayılar arasındaki konumu, ikinci dizide bulunan 51 ile yakındır denilebilir (en azından ikinci dizide bulunan 17 sayısına göre daha yakındır çünkü ikinci dizideki 17 sayısı 0,14 gibi çok daha farklı bir değerdir).

Örnekte de görüldüğü üzere iki farklı dizide bulunan sayılar, birbiri ile karşılaştırılmak istendiğinde, ortak bir sayı sistemine çevrilerek içinde bulundukları diğer sayılara göre, göreceli olarak bir değer atanmış ve karşılaştırılmıştır. İşte normalleştirme tam olarak budur

**2. Standart Skor (Standard Score):**

Diğer bir normalleştirme yöntemidir. Bir önceki yöntemde, sayılar en yüksek ve en düşük değerlere göre normalleştirilmişti. Bu yöntemde ise ortalama değer (mean value) ve [standart sapma (standard deviation)](http://bilgisayarkavramlari.sadievrenseker.com/2012/01/29/2011/05/30/standart-sapma-standard-deviation-stdev/) değerleri göz önüne alınır.  Sistemde kullanılan standart sapmaya atfen, standart skor (standard score) olarak da isimlendirilir. Oldukça popüler normalleştirme yöntemlerinden birisidir. Formülü aşağıdaki şekilde yazılabilir:

Yukarıdaki denklemden de anlaşılacağı zere, bir değerin normalleştirilmesi sırasında, ortalama değere (μ) olan uzaklığı alınarak standart sapma değerine (σ) bölünür. Konunun anlaşılması için, bir önceki örnekte kullandığımız sayı dizisini yeniden işlemeye çalışalım:

|  |  |
| --- | --- |
| 5 | -1,367526918 |
| 8 | -1,025645188 |
| 9 | -0,911684612 |
| 11 | -0,683763459 |
| 20 | 0,341881729 |
| 22 | 0,569802882 |
| 24 | 0,797724035 |
| 25 | 0,911684612 |
| 29 | 1,367526918 |

Yukarıdaki değerleri inceleyecek olursak, 5 ve 29 sayılarının normalleştirme sonucunun mutlak değeri aynı çıkmıştır. Bunun sebebi, iki sayının ortalama değer olan 17’ye olan uzaklıklarının eşit olmasıdır ( 5-17 = -12 ve 29-17 = 12) sadece yönleri farklıdır ve bu durum normalleştirme sonucunda görülmüştür.

Gelelim ikinci dizimizin normalleştirme sonucuna:

|  |  |
| --- | --- |
| 10 | -1,251528679 |
| 15 | -1,016867051 |
| 16 | -0,969934726 |
| 17 | -0,9230024 |
| 50 | 0,625764339 |
| 51 | 0,672696665 |
| 55 | 0,860425967 |
| 56 | 0,907358292 |
| 60 | 1,095087594 |

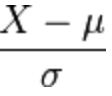
Bir önceki normalleştirme sonucunda en büyük ve en küçük sayıların sadece yönü farklıyken bu sefer değerlerde farklılık oldu. Bunun sebebi iki sayının ortalama değere olan uzaklığının farklılaşmasıdır. Bu ikinci dizideki ortalama değer 36.667 olduğu için iki değer farklı olarak normalleştirilmiştir.

Bir önceki normalleştirme olan asgari / azami normalleştirmede bu fark görülememekteydi. Diğer bir deyişle asgari / azami normalleştirme, en büyük sayıyı hep 1 ve en küçük sayıyı hep 0 olarak kabul edip, diğer sayıları bu 0-1 aralığına yerleştirmekteydi, z-skor normalleştirmesinde ise sayılar ortalama değere göre uzaklıklarına göre normalleştirilmektedir. Ayrıca standart sapmaya bölünerek, sayılar arasındaki hareketlilik (değişim hızı) ortalamaya olan uzaklığı normalize etmektedir.

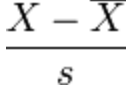
**Normalizasyon yöntemleri**

Gelen bir yorum üzerine bu kısmı eklemeyi gerekli gördüm. Kabaca normalizasyonda kullanılan bazı yöntemleri aşağıdaki şekilde sıralamak mümkündür:

1. Standart Skorlama (Standard Score) : Bir popülasyondaki parametrelerin bilinmesi durumunda hata değerlerini normalleştirmek için kullanılır. Normal dağılımı esas alır ve bu dağılıma uyan popülasyonlarda başarılı çalışır. Basitçe her değerin ortalamadan olan uzaklıklarının standart sapmaya oranı ile bulunur ve aşağıdaki şekilde formülü verilebilir:

[](http://bilgisayarkavramlari.sadievrenseker.com/wp-content/uploads/2012/01/normallesme1.png)

2. Student’s t test(Öğrencinin t testi): popülasyonun parametrelerinin bilinmediği durumlarda kullanılan ve dolayısıyla bir beklendik değerle hesaplanan ve bu beklendik değerle veriler arasındaki farkı / hatayı (tortu, residual) bularak bu değere göre normalleştiren yöntemdir. Her değerin beklendik değere olan uzaklığı ve beklendik değerlerin farklarının ortalamasına oranı ile bulunur. Aşağıdaki şekilde formülü verilebilir (s ile gösterilen değer tortu değeridir).

[](http://bilgisayarkavramlari.sadievrenseker.com/wp-content/uploads/2012/01/normallesme2.png)

3. Standartlaştırılmış an (Standardized moment): Genelde standart moment olarak Türkçede geçen yöntem, aslında momentlerin standartlaştırılması için k. momentin standart sapmaya oranı ile bulunur.

[normallesme3](http://bilgisayarkavramlari.sadievrenseker.com/wp-content/uploads/2012/01/normallesme3.png)

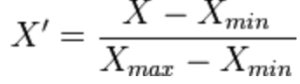
Merkezi moment değeri ise aşağıdaki şekilde hesaplanabilir (E ile gösterilen fonksiyon beklendik değer (Expected value) olmak üzere):

[normallesme4](http://bilgisayarkavramlari.sadievrenseker.com/wp-content/uploads/2012/01/normallesme4.png)

4. Varyans çarpanı (coefficient of variance): Bir popülasyondaki farklılıkları göstermek için kullanılan çarpandır. Basitçe bir popülasyonun ortalama değerinin standart sapmasına oranı olarak hesaplanır:

[normallesme5](http://bilgisayarkavramlari.sadievrenseker.com/wp-content/uploads/2012/01/normallesme5.png)

5. Özellik ölçeklemesi (Feature Scaling) veya Min-Max Normalleştirmesi : Bir popülasyondaki en yüksek ve en düşük değere görece olarak her değerin konumunu hesaplamak için kullanılır:

[](http://bilgisayarkavramlari.sadievrenseker.com/wp-content/uploads/2012/01/normallesme6.png)