**DATA TRANSFORMATION**

* Normalization
* Handling Skewness
* Aggregation of attributes

**1- Normalization / Feature Scaling / Standardization**

What is feature scaling?

* Features on different scales
* Model with linear characteristics
* Center features around 0 and transform to unit variance
* Transforms to approximately normal distribution

**Why standardize variables?**

Many researchers have noted the importance of standardizing variables for multivariate analysis. Otherwise, variables measured at different scales do not contribute equally to the analysis. For example, in boundary detection, a variable that ranges between 0 and 100 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization in effect gives the variable with the larger range a weight of 100 in the analysis. Transforming the data to comparable scales can prevent this problem. Typical data standardization procedures equalize the range and/or data variability.

Many models assume that the distribution is normal, and that why if this is not the case we try to normalize the distribution as much as possible.

|  |  |  |  |
| --- | --- | --- | --- |
| Standardisation  (z-score scaling) | Min-Max scaling  (Normalisation) | Mean Normalisation | Unit vector |
|  |  |  |  |
| Z-score | 0-1 | -1 to 1 | 0-1 |

**Normalization (0-1 scaling / min-max scaling):**

Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

**Standardization (z-score scaling / z-score normalization):**

Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1.

sklearn.preprocessing.scale

Standardization assumes that your data has a Gaussian (bell curve) distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian.

**Which one**

For example, in clustering analyses, standardization may be especially crucial in order to compare similarities between features based on certain distance measures. Another prominent example is the Principal Component Analysis, where we usually prefer standardization over Min-Max scaling

When Scaling

Rule of thumb I follow here is any algorithm that computes distance or assumes normality, scale your features!!!

When to standardize: models (datacamp)

* Model in linear space
* Dataset features have high variance
* Dataset features are continuous and on different scales
* Linearity assumptions

|  |  |  |
| --- | --- | --- |
| algorithm |  | Note |
| KNN | + | with an Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally. |
| PCA | + | Scaling is critical, while performing Principal Component Analysis(PCA). PCA tries to get the features with maximum variance and the variance is high for high magnitude features. |
| Gradient Descent | + | This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven. |
| Tree based models | - | are not distance based models and can handle varying ranges of features. Hence, Scaling is not required while modelling trees. |
| Naïve Bayes | - | are by design equipped to handle this and gives weights to the features accordingly. Performing a features scaling in these algorithms may not have much effect. |
| LDA | - |

**2- Handling Skewness**

Skewness basically gives the shape of normal distribution of values.

|  |  |  |
| --- | --- | --- |
|  | | |
|  | Positively Skewed | Negatively Skewed |
| Transformations | Square root | Square |
| Cube root | |
| Logarithmic (base 10 / base e / base 2) (only with positive data) | |

Resolving outliers :

Outliers can be found using outliers() function from outliers package. This function returns the values at extreme distances from the mean. Once, these are found, we can handle them accordingly to reduce the skewness.

**3- Aggregation**

Aggregation of data refers to making subsets of data using various combinations of attributes, applying the statistical measures on them and reporting the results.

Notlarim:

Feature scaling, yani degiskenlerin olceklendirilmesi, tum algoritmalara uygulanmasi gereken bir metot degil. Hangilerinde yapmamiz gerekiyor sorusunun cevabi, degiskenlerle ilgili hesaplamalari mesafeler uzerinden yapan, yani degiskenlerin varyansini, standart sapmasini kullanarak hesap yapan algoritmalar icin gecerli. Tabi istisnasi, algoritma bu isi kendisi yapiyorsa bizim yapmamiza gerek kalmayacaktir.

Bu aciklamadan da anlasilacagi gibi, normalization, “standardization” ya da “future scaling” meselesi, degiskene ait degerlerin bibiri ile olan mesafelerinin orani muhafaza edilmek sartiyla olcegin daraltilmasina dayaniyor. Neden olcegi daraltma ihtiyaci hissediyoruz. Cunku algoritma bizim predictor olarak verdigimiz tum numerik degiskenleri kullanarak bir fonksiyon olusturacak ve bunu predicton yapmak icin kullanacak. Degiskenlerimizin olcekleri birbirinden cok farkli oldugunda, algoritmanin olusturacagi fonksiyondaki katsayi oranlari bu durumdan olumsuz olarak etkilenecek. Ornegin 2000 ile 10000 arasinda bir deger alan degiskenin katsayisi ile 2 ile 10 arasinda degerler alan bir degiskenin gercek etkileri esit olsa bile, olceklendirme yapmazsak, buyuk degerler alan degiskenin fonksiyona etkisi daha fazla olacak ve dolayisi ile tahmin etme gucunu azaltacaktir.



asd