Why Feature Correlation Matters .... A Lot!

Do Storks Deliver Babies? Although It's been proven theoretically in the context of correlation and causality, this articles explores correlation and how it differs from causality.

Will Badr Follow
Jan 18, 2019 · 5 min read ★

Photo by israel palacio on Unsplash

Machine Learning models are as good or as bad as the data you have. That's why data scientists can spend hours on pre-processing and cleansing the data. They select only the features that would contribute most to the quality of the resulting model. This process is called **"Feature Selection". Feature Selection** is the process of selecting

the attributes that can make the predicted variable more accurate or eliminating those attributes that are irrelevant and can decrease the model accuracy and quality.

Data and feature correlation is considered one important step in the feature selection phase of the data pre-processing especially if the data type for the features is continuous. so **what is data correlation?**

Data Correlation: Is a way to understand the relationship between multiple variables and attributes in your dataset. Using Correlation, you can get some insights such as:

- One or multiple attributes depend on another attribute or a cause for another attribute.
- One or multiple attributes are associated with other attributes.

So, why is correlation useful?

- Correlation can help in predicting one attribute from another (Great way to impute missing values).
- Correlation can (sometimes) indicate the presence of a causal relationship.
- Correlation is used as a basic quantity for many modelling techniques

Let's get a closer look at what this means and how correlation can be useful. There are three types of correlations:

Positive Correlation: means that if feature **A** increases then feature **B** also increases or if feature **A** decreases then feature **B** also decreases. Both features move in tandem and they have a linear relationship.

Negative Correlation (Left) and Positive Correlation (Right)

Negative Correlation: means that if feature **A** increases then feature **B** decreases and vice versa.

No Correlation: No relationship between those two attributes.

Each of those correlation types can exist in a spectrum represented by values from 0 to 1 where slightly or highly positive correlation features can be something like 0.5 or 0.7. If there is a strong and perfect positive correlation, then the result is represented by a correlation score value of 0.9 or 1.

If there is a strong negative correlation, it will be represented by a value of -1.

If your dataset has perfectly positive or negative attributes then there is a high chance that the performance of the model will be impacted by a problem called — "Multicollinearity". **Multicollinearity** happens when one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy. This can lead to skewed or misleading results. Luckily, decision trees and boosted trees algorithms are immune to multicollinearity by nature. When they decide to split, the tree will choose only one of the perfectly correlated features. However, other algorithms like Logistic Regression or Linear Regression are not immune to that problem and you should fix it before training the model.

How Can I Deal With This Problem?

There are multiple ways to deal with this problem. The easiest way is to delete or eliminate one of the perfectly correlated features. Another way is to use a dimension reduction algorithm such as Principle Component Analysis (PCA).

Spearman VS Pearson Correlation Matrix:

Spearman and Pearson are two statistical methods to calculate the strength of correlation between two variables or attributes. **Pearson Correlation Coefficient** can be used with continuous variables that have a linear relationship. Here is an example:

```python
from scipy.stats import pearsonr
import matplotlib.pyplot as plt

#Create two random variable
X = [1,2,3,4,5,6,7,8,9,10,11,12,13]
Y = [7,8,8,11,10,11,12,15,20,14,16,15,19]

#plot the variables to show linearity
plt.scatter(X,Y)
plt.show()

```

pearson.py hosted with ❤ by **GitHub** **view raw**

The output of the above code

To print the Pearson coefficient score, I simply run `pearsonr(X,Y)` and the results are: `(0.88763627518577326, 5.1347242986713319e-05)` where the first value is the Pearson Correlation Coefficients and the second value is the P-value. 0.8 means that the variables are highly positively correlated.

If the variables have a non-linear relationship, you can measure the correlation using **Spearman Correlation Coefficient.** It can also be used with ordinal categorical

variables. You can get the Spearman Coefficient Score by running:

```
scipy.stats.spearmanr(X,Y)
```

Now, this might sound complicated especially with high-dimensional datasets. In that case, it is better to visualize the correlation in a matrix. Here is how you can do that using pandas, I am using Porto Seguro's Safe Driver Prediction Dataset from Kaggle:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

train = read_csv("./train.csv")

def correlation_heatmap(train):
    correlations = train.corr()

    fig, ax = plt.subplots(figsize=(10,10))
    sns.heatmap(correlations, vmax=1.0, center=0, fmt='.2f',
                square=True, linewidths=.5, annot=True, cbar_kws={"shrink": .70})
    plt.show();

correlation_heatmap(train)
```

heatmap.py hosted with ❤ by **GitHub** **view raw**

As you can see in the above matrix, there is a high correlation between ps_reg_03 and ps_reg_02 variables and also between ps_car_12 and ps_car_13 and so on.

There is also another popular method called — Kendall's Tau Coefficient which is also based on variable ranks but unlike Spearman's coefficient, it does not take into account the difference between ranks. Since the focus of this article is on Pearson and Spearman Correlation, Kendall method is outside of this article's scope.

Misconception (Do Storks Deliver Babies?):

Correlation is often interpreted as causation which is a big misconception. Correlation between variables does NOT indicate causation. Any highly correlated variable should be examined and thought of carefully. Here is a (humorous) German paper that used correlation to prove the theory that babies are delivered by Storks. The study shows a significant correlation between the increase in the stork population around the city and the increase in deliveries outside city hospitals

Source: http://web.stanford.edu/class/hrp259/2007/regression/storke.pdf

The chart on the left shows an increase in the number of storks (bold black line) and a decrease in the number of hospital deliveries. On the other hand, the chart on the right shows that a number of out-of-hospital deliveries (white square marks) follow the increasing pattern in the number of storks. Although the study is not meant to prove (the baby-stork theory) scientifically, it shows that a relationship may appear to be causal through having a high correlation. This can due to some unobserved variables. For example, the population increase can be another causal variable.

In Conclusion: Correlations are very useful in many applications, especially when conducting regression analysis. However, it should not be mixed with causality and misinterpreted in any way. You should also always check the correlation between different variables in your dataset and gather some insights as part of your exploration and analysis.

Machine Learning Data Science Statistics Analytics Data