CM2

January 29, 2021

1 Correlation Coefficient and Nature of data

1.1 Required Libraries

[1]:
```
import pandas as pd
```

1.2 Heart Dataset

[2]:
```
df_heart = pd.read_csv("heart_disease_missing.csv")
```

1.2.1 Correlation Coefficient

[3]:
```
df_heart_subset = pd.read_csv("heart_disease_missing.csv",␣
↪usecols=['thalach','exang','slope','oldpeak','target'])
df_heart_subset.corr(method="pearson")
```

[3]:

	thalach	exang	oldpeak	slope	target
thalach	1.000000	-0.360246	-0.351900	0.463824	0.415354
exang	-0.360246	1.000000	0.279862	-0.314675	-0.450321
oldpeak	-0.351900	0.279862	1.000000	-0.652509	-0.454241
slope	0.463824	-0.314675	-0.652509	1.000000	0.419238
target	0.415354	-0.450321	-0.454241	0.419238	1.000000

To measure statistical relationship between features of our Heart dataset, Correlation coefficient plays an important role. Ususally if it's higher than 0.7 then it is said to be strongly related.

In our case, Many features are weakly related to eachother. As shown above, Given features are related to each other by around 0.4 (similarly or dissimilarly, ie. 0.4 or -0.4), which indicates moderate level relation between these features.

1.2.2 Nature of Data

[4]:
```
df_heart_mean = pd.read_csv("heart_disease_missing.csv",␣
↪usecols=['age','sex','fbs','exang','oldpeak','chol','trestbps','thalach'])
df_heart_mean.mean()
```

[4]:
```
age          54.311321
sex           0.688679
trestbps    131.784610
```

```
chol        244.133256
fbs           0.132075
thalach     149.647978
exang         0.344340
oldpeak       1.113106
dtype: float64
```

Mean of numerical values represents center of data, whereas mean of binary values gives us frequency of the data. In this heart dataset, sex,fbs,exang are binary features which indicates that dataset is having more male(~69%) than female, Most people are having their fasting sugar level <120 mg/dl and Exercise induced angina is less produced.

For numerical features like age,trestbps,chol,thalach,oldpeak we can get average values. i.e. Most of the people are about 54 years old. Avergae resting blood pressure is around 131 mm Hg and heart rate achieved during thalium stress test is about 150.

[5]:
```
df_heart.var()
```

[5]:
```
age           83.637217
sex            0.215416
cp             1.045583
trestbps     326.063277
chol        2157.069050
fbs            0.115175
restecg        0.286384
thalach      487.358850
exang          0.226840
oldpeak        1.577304
slope          0.388904
ca             1.079026
thal           0.362545
target         0.249374
dtype: float64
```

Variance is a measure of how far a set of data are spread out from their mean value. Sometimes, two datasets can have same mean but their data spread could be different. So, Variance along with Mean plays an important role in data analysis.

The more the value of variance (here age,trestbps,chol,thalach), the data is more scattered from its mean and if the value of variance is low or minimum, then it is less scattered from mean.

[6]:
```
df_heart.skew()
```

[6]:
```
age        -0.106027
sex        -0.820789
cp          0.461438
trestbps    0.672687
chol        0.333700
fbs         2.188903
```

```
restecg     0.140468
thalach    -0.394100
exang       0.659880
oldpeak     1.224053
slope      -0.604086
ca          1.377751
thal       -0.250145
target     -0.171644
dtype: float64
```

Skewness essentially measures the symmetry of the distribution. For features like age,cp,chol,restecg,thalach,thal skewness is between -0.5 and 0.5, which idicates that data are fairly symmetrical. Whereas for sex,trestbps,exang,slope skewness is between -1 and — 0.5 or between 0.5 and 1, which indicates that data are moderately skewed. And for fbs,oldpeak,ca skewness is greater than 1, so the data are highly skewed.

[7]:
```
df_heart.kurt()
```

[7]:
```
age        -0.561563
sex        -1.339028
cp         -1.240674
trestbps    0.603542
chol        0.254413
fbs         2.817791
restecg    -1.180532
thalach    -0.214108
exang      -1.579550
oldpeak     1.363172
slope      -0.567830
ca          1.020304
thal       -0.646726
target     -1.989397
dtype: float64
```

Kurtosis determines the heaviness of the distribution tails. As we can see, Many features(sex,cp,restecg,exang,target) are having kurt() values less than -1, which indicates that distribution is too flat for them, whereas few of them(fbs,oldpeak,ca) are greater than 1, indicating Peaked distribution. Others are having value near to 0, which shows that, they do not varies much from normal distribution.

1.3 Iris Dataset

[8]:
```
df_iris = pd.read_csv("iris_dataset_missing.csv")
```

1.3.1 Correlation Coefficient

[9]:
```
df_iris.corr(method="pearson")
```

[9]:

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.031792	0.880635	0.809915
sepal_width	-0.031792	1.000000	-0.285793	-0.267574
petal_length	0.880635	-0.285793	1.000000	0.958274
petal_width	0.809915	-0.267574	0.958274	1.000000

For Iris dataset, Features are strongly related to eachother, except couple of pairs. As most of the features are strongly related to each-other, this makes statistical analysis more useful for future tasks.

1.3.2 Nature of Data

[10]:
```
df_iris.mean()
```

[10]:
```
sepal_length    5.858909
sepal_width     3.059083
petal_length    3.812370
petal_width     1.199708
dtype: float64
```

Mean gives us the overview of data. Here, Average sepal_length for flowers is 6 cm, whereas sepal_width and petal_length are about 3 cm and petal_width of flowers is arounf 1 cm.

[11]:
```
df_iris.var()
```

[11]:
```
sepal_length    0.742420
sepal_width     0.207131
petal_length    3.216602
petal_width     0.619672
dtype: float64
```

for Iris dataset, Only Petal_length is having high variance, which means data is more scattered from mean. Whereas, for other features, it is not scattered much.

[12]:
```
df_iris.skew()
```

[12]:
```
sepal_length    0.401506
sepal_width     0.367708
petal_length   -0.255767
petal_width    -0.074751
dtype: float64
```

As skewness for each feature is between -0.5 to 0.5, Data is said to be fairly symmetrical.

For specific values, A negative skewness value in (petal_length, petal_width) indicates that tail is larger towards the left hand side of the distribution. Whereas, A positive skewness value in (sepal_length,sepal_width) indicates that tail is larger towards the right hand side of the distribution.

[13]:
```
df_iris.kurt()
```

[13]:
```
sepal_length   -0.544820
sepal_width     0.510490
petal_length   -1.389810
petal_width    -1.315451
dtype: float64
```

Kurtosis is one of the two measures that quantify shape of a distribution and determine the volume of the outlier. Here, Sepal related features are near to 0, varies less from normal distribution. Whereas, Petal related features are <-1, indicating relatively flat distribution (Also known as Platykurtic distribution).

1.4 References

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.kurt.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.mean.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.skew.html
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.var.html