Calculating correlation values for categorical data

Science, Computer Science

Calculating correlation values for categorical data In order to find the correlation values for the fields in our data set, The Pearson Correlation Coefficient was used. This requires that the data in both fields be quantitative. But what if we were looking to calculate the correlation on two given fields that were say, numerical and categorical, or even both categorical. The Point Biserial coefficient is a special case of The Pearson Correlation Coefficient; it is a branch of PCC although they are mathematically equivalent.

It is used when one field has quantitative data and the other has categorical values, specifically categorical data that can only be one of two options for example gender. To calculate the PBC the data is divided between the two values of the dichotomous data, where the two values of this field are given the values 0 and 1. The distribution of the data will in general show the frequencies for each value and can be used to show how well two fields are correlated.

Spearman’s Rank Order Coefficient is a method of estimating correlation between data that is nominal and importantly must be ordered. It checks how well the relationship between the two fields can be described using a monotonic function Another method for calculating the correlation is the Chi squared Test, this requires data to be classified and frequencies worked out in a table. From this table the correlations can be determined using the Chi Square Test, this works on any pair of nominal or categorical fields