How to Calculate the Correlation Between Ordinal Input Features and Nominal Output

In data science and machine learning, understanding the relationship between your input features and the output variable is crucial for making informed predictions and decisions. However, when you have an ordinal input feature (i.e., a variable with a meaningful order, but not a measurable distance between categories) and a nominal output (i.e., categorical data without a specific order), determining the strength and direction of this relationship can be a bit more complex.

But don't worry—there are still plenty of techniques to help you measure and understand this connection. In this blog, we will explore how you can assess the correlation between an ordinal input feature and a nominal output, breaking down the process step-by-step.

What Are Ordinal and Nominal Variables?

Before diving into the correlation methods, let’s briefly define what we mean by ordinal and nominal variables:

- **Ordinal Variable:** An ordinal variable has categories with a meaningful order or ranking but no specific, consistent difference between the ranks. For example, "low," "medium," and "high" ratings are ordinal variables because they have a natural order, but the difference between "low" and "medium" may not be the same as between "medium" and "high."
- **Nominal Variable:** A nominal variable, on the other hand, consists of categories without any order or ranking. For example, “red,” “blue,” and “green” are nominal variables because these categories don’t have an inherent order.

When the input feature is ordinal, and the output is nominal, you need a method that accounts for the ordering of the input while respecting the unordered nature of the output.

1. Chi-Square Test of Independence

One common way to examine the relationship between an ordinal input and a nominal output is to use the **Chi-square test of independence**. This statistical test determines whether two categorical variables (in this case, ordinal and nominal) are independent or related.

While the Chi-square test is typically used for two nominal variables, it can also be used to explore the relationship between an ordinal feature and a nominal output by discretizing the ordinal variable into a set of categories. By creating a contingency table that cross-tabulates the ordinal input and the nominal output, you can evaluate whether the distribution of the input feature differs across the categories of the nominal output.

Steps to Apply the Chi-Square Test:

1. **Discretize the Ordinal Feature**: If your ordinal feature consists of ordered categories (e.g., "low," "medium," and "high"), you can treat them as categorical levels.
2. **Build a Contingency Table**: Count how many data points fall into each combination of ordinal feature levels and nominal output categories.
3. **Perform the Chi-Square Test**: The Chi-square statistic compares the observed counts with expected counts, testing if the features are related or independent.

Example:

Suppose you're analyzing customer satisfaction (nominal: "satisfied," "neutral," "unsatisfied") based on their income level (ordinal: "low," "medium," "high"). By discretizing the ordinal income categories and applying the Chi-square test, you can assess whether income level is associated with satisfaction level.

2. Kruskal-Wallis H Test

If your ordinal input has more than two levels (such as ratings from "poor" to "excellent") and your nominal output is still categorical, you might want to consider the **Kruskal-Wallis H test**. This non-parametric test is used to compare more than two independent groups to determine whether there are statistically significant differences in the distribution of the ordinal input between the categories of the nominal output.

The Kruskal-Wallis test works similarly to the **ANOVA** test but doesn't require the data to follow a normal distribution. It's particularly useful when you have an ordinal feature and a nominal output with multiple categories.

Steps to Apply Kruskal-Wallis H Test:

1. **Rank the Ordinal Data**: If the ordinal feature has multiple levels (e.g., "poor," "fair," "good," "excellent"), assign numerical ranks to the data points.
2. **Group the Data by Nominal Output**: Separate the data into groups based on the categories of the nominal output.
3. **Perform the Kruskal-Wallis Test**: The test evaluates whether there is a statistically significant difference in the ranks of the ordinal input across the categories of the nominal output.

Example:

Let’s say you're studying employee performance (ordinal: "low," "medium," "high") based on different departments (nominal: "Sales," "Marketing," "Engineering"). The Kruskal-Wallis H test can help you determine whether performance ratings differ significantly across departments.

3. Ordinal Logistic Regression

If your nominal output has only two categories (binary), you can apply **logistic regression** techniques, but with a twist: **Ordinal logistic regression**. This approach is specifically designed to model the relationship between an ordinal input variable and a binary outcome. However, for a nominal output with more than two categories, this method might not be directly applicable.

Ordinal logistic regression estimates the probability of an event occurring, given the ordinal input feature. It assumes a proportional odds model, meaning that the odds of being in a higher category of the ordinal input relative to the categories of the nominal output are constant.

Steps to Apply Ordinal Logistic Regression:

1. **Prepare Your Data**: Ensure that your ordinal feature is encoded appropriately and your nominal output is also in a suitable format.
2. **Fit the Model**: Train the ordinal logistic regression model using your ordinal input and nominal output.
3. **Interpret the Coefficients**: The coefficients indicate the effect of the ordinal feature on the likelihood of different nominal outcomes.

Example:

Imagine you're predicting the likelihood of a customer purchasing a product (nominal: "yes," "no") based on their loyalty level (ordinal: "low," "medium," "high"). Ordinal logistic regression can model how loyalty influences purchasing behavior.

4. Spearman's Rank Correlation

For ordinal data, **Spearman's rank correlation** can be useful if your nominal output is binary (two categories). This non-parametric measure assesses how well the relationship between two variables can be described using a monotonic function, meaning that as one variable increases, the other tends to increase (or decrease).

Though Spearman's rank correlation is typically used for two ordinal variables, it can be adapted to assess the relationship between an ordinal input and a binary nominal output by comparing the ranks of the ordinal data and the categories of the nominal data.

Steps to Apply Spearman’s Rank Correlation:

1. **Rank the Ordinal Data**: Convert the ordinal feature to numerical ranks.
2. **Calculate the Correlation**: Compute the Spearman correlation coefficient between the ranks of the ordinal input and the binary nominal output.

Example:

Consider predicting whether a student will pass or fail an exam (binary nominal: "pass," "fail") based on their study hours (ordinal: "few," "moderate," "many"). By ranking study hours and calculating Spearman's rank correlation, you can evaluate how strongly the study hours are related to exam outcomes.

Conclusion

In summary, when dealing with ordinal input features and nominal output, there are several methods at your disposal to measure correlation. The appropriate technique depends on factors like the number of levels in the ordinal input and the categories in the nominal output. Methods like the Chi-square test, Kruskal-Wallis test, ordinal logistic regression, and Spearman's rank correlation all provide ways to quantify the relationship, allowing you to make better decisions based on your data.

Choosing the right method will help you draw meaningful insights from your data, ensuring that you can model and predict your nominal output accurately, even when the input is ordinal.