Data Pre-processing (Feature Engineering)

Applied Machine Learning for Educational Data Science

true

08/23/2021

Contents

[Updated: Mon, Aug 23, 2021 - 18:27:33]

Scales of Measurement and Types of Variables

It is important to understand the nature of variables and how they were measured and represented in a dataset. In social sciences, in particular psychology, there is a methodological consensus about the framework provided by Stevens (1946), also see Michell (2002) for an in-depth discussion. According to Stevens' definition, there are four levels of measurement: nominal, ordinal, interval, and ratio. Whether a variable is considered having a nominal, ordinal, interval, or ratio scale depends on the character of the empirical operations performed while constructing the variable.

- Nominal scale: Variables with a nominal scale cannot be be meaningfully added, subtracted, divided, or multiplied. Also, there is no hierarchical order among the assigned values. Most variables that contains labels for individual observations can be considered as nominal, e.g., hair color, city, state, ethnicity.

- Ordinal scale: Variables with an ordinal scale also represent labels; however, there is a meaningful hierarchy among the assigned values. For instance, if a variable is coded as Low, Medium, and High, they are simply labels but we know that High represents something more than Medium, and Medium represents something higher than Low (High > Medium > Low). On the other side, the distance between the assigned values do not necessarily represent the same amount of difference. Other examples of variables that can be considered as ordinal are letter grades (A-F), scores from likert type items (Strongly agree, agree, disagree, strongly disagree), education status(high school, college, master's, PhD), cancer stage (stage1, stage2, stage3), order of finish in a competition (1st, 2nd, 3rd).

- Interval scale: Variables with an ordinal scale represents quantities with equal measurement units but they don't have an absolute zero point. For instance, a typical example of an interval scale is temperature measured on the Fahrenheit scale. The difference between 20F and 30F is the same difference as the difference between 60F and 70F. However, 0F does not indicate no heat.

- Ratio scale: Variables with a ratio scale represents quantities with equal measurement units and have an absolute zero. Due to the nature of the existence of absolute zero point that represent 'nothing', ratio of measurements are also meaningful. Typical examples are height, mass, distance, length.

Below table provides a summary of properties for each scale.

	Indicating Difference	Indicating Direction of Difference	Indicating Amount of Difference	Has absolute zero
Nominal	X			
Ordinal	X	X		
Interval	X	X	X	
Ratio	X	X	X	X

In this class, we classify the variables in two types: **Categorical** and **Continuous**. The variables with a nominal or ordinal scale are considered as **Categorical** and the variables with an interval or ratio scale are considered as **Continuous**.

Processing Categorical Variables

One-hot encoding

Label encoding

Polynomial Contrasts

Processing Day and Time Data

Processing Continuous Variables

Centering and Scaling

Transformations

Basis Expansions and Splines

Linear Projections

Handling Missing Data

Data Leakage

Processing Text Data