TRANSCRIPT: DATA PREPROCESSING

1.2 Introduction

Data preprocessing is the phase where a machine learning engineer spends most of his coding time in a machine learning project. Any model will read numerical data. However, raw data may have non numerical data, missing values etc. In this phase raw data is transformed to such a format which can be fed to the code of machine learning model. Processed data is also known as cleaned data.

We shall learn the following in this section.

1.3 Types of Variable

It is important to understand the types of features or variables. We encounter 2 kinds of variables. They are numerical and categorical. Numerical variable is one whose value is a number. It can take infinite number of values from the set of real numbers. It is ready to be fed to a machine learning model.

Categorical variable is one which can be divided into finite number of groups or categories. The groups can have numerical labels or literal labels but do not have mathematical meaning. Techniques like one hot encoding has to be implemented to derive mathematical meaning from categorical variable. We shall discuss one hot encoding in coming slides.

Further numerical variables are classified as Discrete or Continuous.

Discrete variables are numerical variables that have a countable number of values. A discrete variable is always numerical. For example, the number of customer complaints or the number of flaws or defects.

Continuous variables are numerical variables that have an infinite number of values. A continuous variable can be numeric or date or time. For example, the length of a part or the date and time a payment is received.

There are two types of categorical variable, nominal and ordinal. A nominal variable has no in built ordering to its categories. For example, gender is a categorical variable having three categories (male or female or other) with no in built ordering to the categories. An ordinal variable has a clear ordering. For example, temperature as a variable with three orderly categories (low, medium and high).

Now, we shall see a data set with different types of variables. Imagine a data set where the height of a person is dependent variable and is dependent upon age, gender and weight of a person. Please refer table 1 where dependent variables of such data-set is shown. Thus age, gender and weight are features or dependent variables. Weight is a numerical variable as it can take any value from real numbers. Similarly age is a numerical variable which can theoretically take any value from the set of natural numbers. However, gender is a categorical variable which can be male or female or other. Thus gender can have only 3 values. These values need to be changed to mathematical values which can be fed to a machine learning model.

1.4 Types of Variable

Raw data may have missing values. These missing values may arise due to data corruption, non availability of data, error in data acquisition methodology, etc. An example of missing value is given in table 1. Table 1 has 3 columns representing 3 variables or features which are weight, age and gender. In the weight column we can see that data in last but one column is missing.

We cannot feed data with missing values to a machine learning model. So, we need to fill this missing value. This process is called missing value imputation. Two common strategy towards missing value imputation are mean and median imputation. In mean imputation mean of available data values is filled in the missing cell. In case of median imputation median of available data values is filled in the missing cell. Thus the empty cell will be filled with 53.29 using mean imputation, and 50 using median imputation.

1.5 Missing value imputation

Raw data may have missing values. These missing values may arise due to data corruption, non availability of data, error in data acquisition methodology, etc. An example of missing value is given in table 1. Table 1 has 3 columns representing 3 variables or features which are weight, age and gender. In the weight column we can see that data in last but one column is missing.

We cannot feed data with missing values to a machine learning model. So, we need to fill this missing value. This process is called missing value imputation. Two common strategy towards missing value imputation are mean and median imputation. In mean imputation mean of available data values is filled in the missing cell. In case of median imputation median of available data values is filled in the missing cell. Thus the empty cell will be filled with 53.29 using mean imputation, and 50 using median imputation.

1.6 Feature scaling

Variables that are measured at different scales do not contribute equally to the model fitting & model learning, and might end up creating a bias. Bias refers to one feature becoming more important just because its magnitude is more than other feature.

Many machine learning algorithms perform better when numerical input variables are scaled to a standard range. This includes algorithms that use a weighted sum of the input, like linear regression, and algorithms that use distance measures, like k-nearest neighbors.

The process of bringing features or dependent variables to same scale is called feature scaling. 2 most important methods of feature scaling are Standard scaling and Normalization.

In standard scaling each value under a feature is scaled by subtracting the mean and dividing the result by variance. Let us understand using the data in table 1. Let us denote weight feature by x i, where i refers to numbering of data point and ranges from 1 to n, and n is the number of data points. As total number of data points is 10 in this case, so i will range from 1 to 10. The mean μ of feature weight is calculated using formula 1, and standard deviation is calculated using formula 2. Finally standard scaling is done using formula 3 on all the weight values. Similarly age feature is standardized.

In normalization, each value under a feature is scaled between -1 to 1 by subtracting minimum value, and dividing the result by difference of maximum and minimum values. Let us denote weight feature by x i, where i refers to numbering of data point and ranges from 1 to n, and n is the number of data points. As total number of data points is 10 in this case, so i will range from 1 to 10. Minimum value of weight is denoted by x i min and is the minimum among all the values under weight feature. Maximum value of weight is denoted by x i max and is the maximum among all the values under weight feature. Finally normalization is done using formula 6 on all the weight values. Similarly age feature is normalized.

Standardization and normalization calculations done in excel sheet can be downloaded from resources.

1.7 One hot encoding

In table 1, we can see that there is one categorical variable column named gender. The values under this column are male or female or other, and this does not convey any mathematical meaning, and so cannot be fed directly to a machine learning model. In order to make it mathematically meaningful, we use one hot encoding. Let us discuss it.

We have three values under gender. So, a 3 dimensional representation is created for gender. Suppose the 1st dimension of this 3 dimensional representation is male, 2nd is female and 3rd is other. As the 1st data point is female it is represented as 0 1 0. The 2nd dimension representing female is given a value 1 while other dimensions have 0 value. Similarly for 2nd data point the representation is 1 0 0. Here the dimension representing male is 1 and others are 0. Similarly for 6th data point the representation is 0 0 1. This encoding method is called one hot encoding as value corresponding to the active dimension is hot or 1, and only one dimension is active. Thus the name one hot encoding. The complete one hot encoded representation is shown in table 2 on next slide.

1.8 One hot encoding

Table 2 gives the complete one hot encoding for gender. It may be noted that three new columns representing the three dimensional representation for gender appear in table 2, while the original gender column is not there. The names of new columns will be as per the implementation by the software package in use.

Another example of one hot encoding is shown in table 3.