Replace Discrete Values

Updated: June 27, 2015

Replaces discrete values from one column with numeric values based on another column

Category: Statistical Functions (https://msdn.microsoft.com/en-us/library/azure/dn905867.aspx)

Module Overview

You can use the **Replace Discrete Values** module to generate a probability score that can be used to represent a discrete value. This can be useful for understanding the information value of the discrete values.

You select a column that contains the discrete (or categorical) value, and then select another column to use for reference. Depending on whether the second column is categorical or non-categorical, the module computes one of the following values:

- The **conditional probability** for the second column given the values in the first column.
- The **mean** and **standard deviation** for each group of values in the first column.

The module outputs both a transformed dataset, and a function that you can save as a transform and apply to other datasets.

How to Configure Replace Discrete Values

1. For the **Discrete columns** option, click **Launch column selector** to choose one or more columns that contain discrete (or categorical) values

   The discrete columns you select here must be categorical. Use the Metadata Editor (https://msdn.microsoft.com/en-us/library/azure/dn905986.aspx) module to mark the columns if you get an error.

2. For the **Replacement columns** option, click **Launch column selector** to choose one or more columns that contain the values to use in computing a replacement score.

   You must select an equal number of columns to the discrete columns in the previous step, and they cannot be the same columns.

Tip

The module does allow you to select more than one pair of columns to analyze. However, in practice, when you choose multiple columns, they are matched by an internal heuristic, not by order of selection. To ensure that you get the right replacement values, we recommend that you select a pair of columns each time and use separate instances of **Replace Discrete Values** if analyzing multiple columns.

3. Depending on whether the second column is categorical or non-categorical, the module computes one of the following values:

    - If the second column is a categorical value, it computes the **conditional probability** for the second column given the values in the first column.

      For example, if you choose **occupation** (from the Census dataset) as the discrete column and choose **gender** as the replacement column. The module outputs:

        - **P(gender|occupation)**.

    - If the second column non-categorical values that can be converted to numbers (such as numeric or Boolean values **not** marked as categorical), it calculates the **mean** and **standard deviation** for each group of values in the first column.

      For example, if you select **occupation** as the **Discrete column** and the numeric column **hours-per-week** as the **Replacement column**, the new value columns would have these headings:

        - **Mean(hours-per-week|occupation)**

        - **Std-Dev(hours-per-week|occupation)**

> **Warning**
>
> You cannot choose the statistical function to apply. The module calculates the appropriate measure, based on the data type of the column selected for **Replacement column**.

4. Run the experiment.

5. The module outputs a transformed dataset in which the columns you selected as the **Replacement columns** have been replaced with the computed scores.

   The headings of the new columns indicate the source columns and the operation.

Options

Discrete columns

Use the Column Selector to choose columns to analyze. The columns must contain discrete values, and they must be already marked as categorical.

Replacement columns

Choose columns to replace with scores. These columns are used to computes the scores, but only the scores are output.

> **Tip**
>
> Although the original column values for the replacement columns are not included in the transformed dataset, the actual data columns in your dataset are not directly affected by the transformation; the new score columns are just added as metadata. You can always get the column values by using add columns and connecting the source dataset..

Examples

The sample experiments do not include any instances that are relevant to this module, but the usage of **Replace Discrete Values** can be illustrated by some simple examples:

Replacing a Categorical Value with a Probability Score

The following table contains a categorical column X, and a column Y with True/False values that are treated as categorical values. When you use **Replace Discrete Values**, it calculates a conditional probability score as shown in the third column.

X	**Y**	**P(Y|X)**
Blue	0	P(Y=0|X=Blue) = 0.5
Blue	1	P(Y=1|X =Blue) = 0.5
Green	0	P(Y=0|X=Green) = 2/3
Green	0	P(Y=0|X=Green) = 2/3
Green	1	P(Y=1|X=Green) = 1/3
Red	0	P(Y=0|X=Red) = .75
Red	0	P(Y=0|X=Red) = .75
Red	1	P(Y=1|X=Red) = .25

Red	0	P(Y=0|X=Red) = .75

Replacement based on noncategorical column Y

If the second column is numerical, **Replace Discrete Values** calculates *Mean(Y|X)* and *Std-Dev(Y|X)* instead of the conditional probability score.

This example is based on the sample Auto Prices dataset. To simplify validation, a small subset of columns is projected and only the top 30 rows are extracted by using the Head option of **Partition and Sample**. Then **Replace Discrete Values** is used to compute the mean values and standard deviation for curb weight given the categorical column, num-of-doors.

Body	Num-of-doors	Curb-weight	Mean(curb-weight|num-of-doors)	Std-Dev(curb-weight|num-of-doors)
std	two	2548	2429.785714	507.45699
std	four	2337	2625.6	493.409877
std	two	2507	2429.785714	507.45699
turbo	four	3086	2625.6 5	493.409877
std	four	1989	2625.6	493.409877
turbo		2191		
std	four	2535	2625.6	493.409877

You can verify the mean and standard deviation for each group of values by using the **AVERAGEIF** function in Excel.

Missing values and conditional probability

This example demonstrates how missing values (nulls) propagate to the results.

- If the discrete value column and the calculation lookup column contains any missing values, the missing values are propagated to the new column.
- If the discrete value column contains only missing values, the module cannot process the column and an error message appears.

X	Y	P(Y|X)

1	True	*P(Y=true|X=1) = 1/2*
1	False	*P(Y=false|X=1) = 1/2*
2	True	*P(Y=true|X=2) = 1/3*
2	False	*P(Y=false|X=2) = 1/3*
2	Null	*P(Y=null|X=2) = null*

Technical Notes

- You must ensure that any discrete columns you want to replace are categorical, or the module will return an error.

  You can do this by using Metadata Editor (https://msdn.microsoft.com/en-us/library/azure/dn905986.aspx).

- If the second column contains Boolean values, the True-False values are processed as numeric with FALSE and TRUE equivalent to 0 and 1 respectively.

- The formula for the standard deviation column calculates the population standard deviation. Therefore, *N* is used in the denominator instead of *(N - 1)*.

- If the second column contains noncategorical data (numeric or Boolean values), the module computes the mean and standard deviation of Y for the given value of X—that is, for each row in the dataset indexed by i:

  $Mean(Y \mid X)_i = Mean(Y \mid X = X_i)$

  $StdDev(Y \mid X)_i = StdDev(Y \mid X = X_i)$

- If the second column contains categorical data or values that are neither numeric nor Boolean, the module computes the conditional probability of Y for the given value of X—that is, for each row in the dataset indexed by i.

- Any Boolean values in the second column are processed as numeric data with FALSE and TRUE equivalent to 0 and 1 respectively.

- If there is a class in the discrete column, such that a row with a missing value is present in the second column, the sum of conditional probabilities within the class is less than one.

Expected Input

Name	Type	Description
Dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Input dataset

Module Parameters

Name	Range	Type	Default	Description
Discrete columns	Any	ColumnSelection		Selects the columns that contain discrete values
Replacement columns	Any	ColumnSelection		Selects the columns that contain the data to use in place of the discrete values

Outputs

Name	Type	Description
Supplemented dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Dataset with replaced data
Transform function	ITransform interface (https://msdn.microsoft.com/en-us/library/azure/dn905982.aspx)	Definition of the transform function, which can be applied to other datasets

Exceptions

For a complete list of error messages, see Machine Learning Module Error Codes (https://msdn.microsoft.com/en-us/library/azure/dn905910.aspx).

Exception	Description

Error 0001 (https://msdn.microsoft.com/en-us/library/azure/dn905993.aspx)	Exception occurs if one or more specified columns of the data set couldn't be found.
Error 0003 (https://msdn.microsoft.com/en-us/library/azure/dn906003.aspx)	Exception occurs if one or more of inputs are null or empty.
Error 0020 (https://msdn.microsoft.com/en-us/library/azure/dn906040.aspx)	Exception occurs if the number of columns in some of the datasets passed to the module is too small.
Error 0021 (https://msdn.microsoft.com/en-us/library/azure/dn905802.aspx)	Exception occurs if the number of rows in some of the datasets passed to the module is too small.
Error 0017 (https://msdn.microsoft.com/en-us/library/azure/dn906039.aspx)	Exception occurs if one or more specified columns have a type that is unsupported by the current module.
Error 0026 (https://msdn.microsoft.com/en-us/library/azure/dn906052.aspx)	Exception occurs when columns with the same name are not allowed.
Error 0022 (https://msdn.microsoft.com/en-us/library/azure/dn906050.aspx)	Exception occurs if the number of selected columns in the input dataset does not equal the expected number.

See Also

Statistical Functions (https://msdn.microsoft.com/en-us/library/azure/dn905867.aspx)
A-Z List of Machine Learning Studio Modules (https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx)