Preparing Data

Eng Teong Cheah
MVP Visual Studio &
Development Technologies

Agenda

Data preprocessing
Strategies for incomplete datasets

Data preprocessing

Cleaning Missing Data

Data scientists often check data for missing values and then perform various operations to fix the data or insert new values.

The goal such cleaning operations is to prevent problems caused by missing data that can arise when training a model.

Normalizing data

Normalization is a technique often applied as part of data preparation for machine learning.

The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

Normalization data

Normalization is also required for some algorithms to model the data correctly.

For example, assume your input dataset contains one column with values ranging from 0 to 1, and another column with values ranging from 10,000 to 100,000. The great difference of the numbers could cause problems when you attempt to combine the values as feature during modeling.

Group Data into Bins

The Group Data into Bins module supports multiple options for binning data. You can customize how the bin edges are set and how values are appointed into the bins.

Group Data into Bins

For example, you can:

- Manually type a series of values to serve as the bin boundaries.
- Calculate entropy scores to determine an information values for each range, to optimize the bins in the predictive model. + Assign values to bins by using quantiles, or percentiles ranks.

Group Data into Bins

- Control the number of values in each bin can also be controlled.
- Force an even distribution of values into the bins.

Group Categorical Values

The typical use for grouping categorical values is to merge multiple string values into a single new level.

For example, you might assign individual postal codes in a region to a single regional code, or group in multiple products under one category.

Strategies for incomplete datasets

Handling missing data

Real world data is usually missing values, which trip up a lot of machine learning algorithms. There are lots of tricks for dealing with these, but you have to be careful. The way in which you fill them can change the result dramatically. Being explicit and thoughtful about how you handle missing values will get you the very best results.

Clip values

To identify and optionally replace data values that are above or below specified threshold.

This is useful when you want to remove outliers or replace them with a mean, a constant, or other substitute value.

Working with imbalanced data

SMOTE is a better way of increasing the number of rare cases than simple duplicating existing cases.

SMOTE takes the entire dataset as an input, but it increases the percentage of only the minority cases.

Working with imbalanced data

For example, suppose you have an imbalanced dataset where just 1% of the cases have the target value A (the minority class), and 99% of the cases have the value B.

To increase the percentage of minority cases to twice the previous percentage, you would enter 200 for SMOTE percentage in the module's properties.

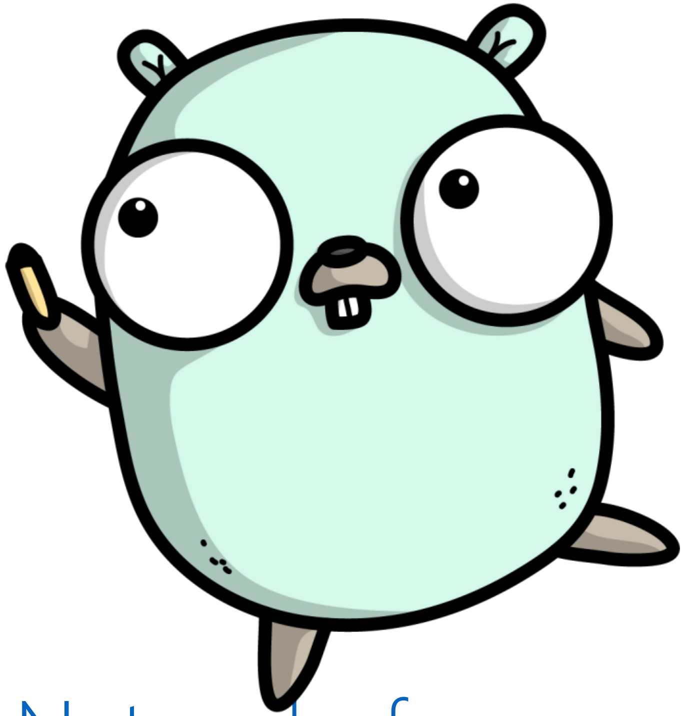
Demo

Preprocess data using Machine Learning Studio

Resources

TutorialsPoint

Microsoft Docs



Lecture Collection | Convolutional Neural Networks for Visual Recognition(Spring 2017)

Python Numpy Tutorial

Image Credits: @ashleymcnamara

Thank you

Eng Teong Cheah
Microsoft MVP Visual Studio & Development Technologies
Twitter: @walkercet
Github: https://github.com/ceteongvanness
Blog: https://ceteongvanness.wordpress.com/
Youtube: http://bit.ly/etyoutubechannel