DATASCI 510 Lesson 5

Data Preparation (Pre-processing)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Lesson 5 Agenda

- Announcements
- Transformation (Slides)
- Binning: Lesson_05_a_Binning.ipynb
- Break
- Indexing and Grouping: Lesson_05_b_Student.ipynb (1st Part)
- Category Columns: Lesson_05_b_Student.ipynb (2nd Part)
- Break
- One hot encoding: Lesson_05_b_Student.ipynb (3rd Part)
- Interview question

Announcements

- Create Study Groups
- Use of **Ed Discussion** will be graded!
  - Please submit at least two posts each week
  - The topics must be related to data science
  - Best topics pertain to the lecture or homework
- Should we continue with the **Interview Questions**?

Transformations of Numeric Variables (Normalizing and Binning)

Normalizing

NORMALIZATION

Overview

- Also referred to as "**scaling**" a variable
- Applies to numeric variables only
- Essential as part of data engineering
- Various ways of performing normalization
- Adjusts the scale and offset of a numeric column

NORMALIZATION

Min-max normalization method

- Often called feature scaling (https://en.wikipedia.org/wiki/Feature_scaling)
- Involves rescaling the variable from 0 and 1
- Is often favored because the range is always the same
- Is strongly affected by outliers and therefore often not recommended

NORMALIZATION

Z-normalization method

- Also referred to as standardization
- Ideal for variables following the normal distribution
- Involves changing the variable so that its mean is equal to 0.0 and its standard deviation equal to 1.0
- Outliers affect the overall normalization to a lesser extent

NORMALIZATION

Useful considerations when normalizing a variable

- Combining (linear) normalization methods is unnecessary, since it's just the final normalization that matters
- Binary variables can be normalized too, but in the case of min-max normalization it's unnecessary
- Variable values become comparable if one uses the same normalization method for all normalizations in a dataset
- When normalizing based on a sample, it is best to use the same values of min/max or μ/σ when you normalize the rest of the values of the variable
- Normalization can be reversed, if you have kept the parameters used for it

Numeric Binning

NUMERIC BINNING

Overview

- Involves grouping values of a numeric variable together and substituting them with a single value, usually a category
  - Groups = bins
- Loses some of the signal from the original variable
- Useful for replacing a continuous numeric variable with a categorical variable
  - Boundaries of each bin can be predefined or selected automatically

NUMERIC BINNING

Standard binning method (Equal-width binning)

1. Define the number of bins (N)
2. Find the bin width: W = (max(x) - min(x)) / N
3. For each bin:
    1. Calculate the boundaries low, high
    2. Find all the data points in x belonging to [low, high]
    3. Assign a unique bin label to these points

NUMERIC BINNING

Binning and histograms

Histograms are great for depicting what a variable's distribution looks like:

> A variable's histogram may help set binning limits

> The numpy histogram function can be used to determine boundaries:

- Try: plt.hist(x)

NUMERIC BINNING

Useful considerations when binning a variable

- Selecting an appropriate number of bins is very useful for meaningful results
- Usually various scenarios are tried before committing to a single one
- Binning is not reversible as a process

Summary of Normalization and Binning

- Normalization
  - Numeric to Numeric
  - Shifts and sets the scale
  - Reversible
  - *sklearn* package, *preprocessing* class, *StandardScaler* and *MinMaxScaler* functions
- Binning
  - Numeric to Categorical
  - Sets a categorical label
  - Irreversible
  - *numpy* package, *histogram* function, quantile, cut
- Comparison of various normalization methods in Python: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py

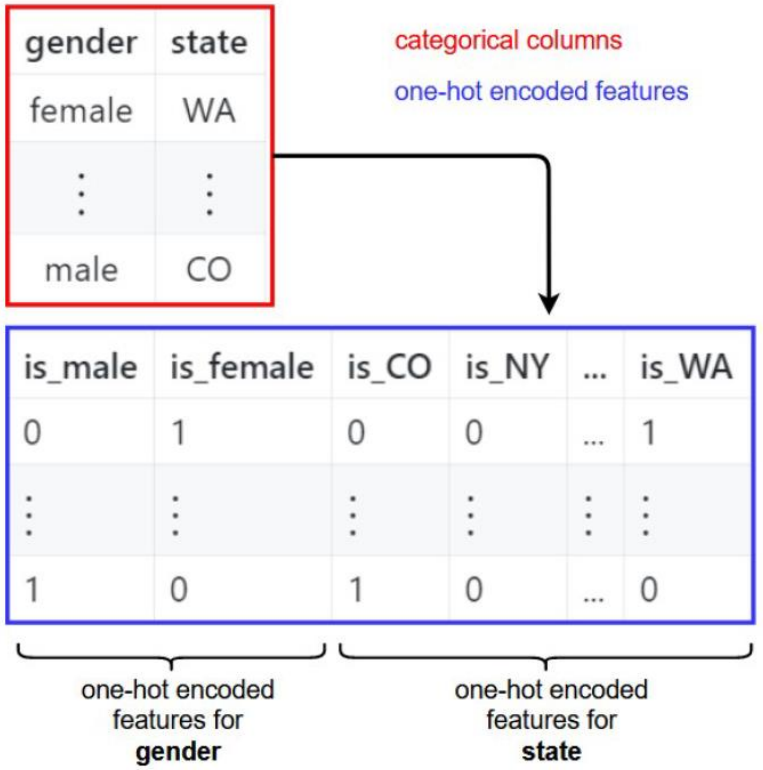
Transformations of Category Variables (Decode, Binning, One-hot Encoding)

One-Hot Encoding

One-Hot Encoding

one-hot encoding

- raw data has **two** categorical columns with 2×50 categories
- featurized data has 2×50 **binary** (numeric) columns
- **high cardinality** categorical features can make resulting data large (until you run out of memory)



One-Hot Encoding

one-hot encoding with `sklearn`

`fit` and `transform` is a common pattern in ML (even for data pre-processing steps like one-hot encoding)

```
from sklearn.preprocessing import OneHotEncoder

onehot = OneHotEncoder(sparse = False)

onehot.fit(data)

col_names = onehot.get_feature_names(data.columns)
data_onehot = pd.DataFrame(onehot.transform(data), columns = col_names)
```

Summary vs. Transformation

> Summary or Description
  - The description is a summarization which means a reduction in size
  - The description does not have the same length as the original variable (length is number of values in a variable or number of rows in a table)
  - E.g. One column is described by one scalar like a mean
  - E.g. Multiple columns are described by one scalar like the correlation coefficient
  - E.g. One column is described by multiple scalars like normalization constants (mean and standard deviation)

> Transformation
  - A transformation leads to one or more variables that have the same length as the original data.
  - E.g. One column transformed into another column (e.g. binning and normalization)
  - E.g. One column transformed into multiple columns (e.g. one-hot-encoding)
  - E.g. Multiple columns transformed to a single column (e.g. ratio, addition)

Technically (Mathematically) a description or summarization is also a transformation but colloquially we consider the above definitions and criteria

Some Transformation Examples (1)

> Type Casting

- object to string
- literal integers to numeric integers

> Type Promotion

- Boolean to integers
- Integer to float

> Type Coercion

- float to integer: Coercion is the conversion of values, like floats, to another type of values, like integers. Coercion allows for possible information loss. like changing the float 7.3 to the integer 7
- Literal number (“1”) to actual value (1)

> Categorical to Categorical

- Renaming Variables
- Category Binning or Consolidation: Combine categories to reduce cardinality. Category Consolidation takes a column with many categories and creates a new replacement column that has fewer categories (values)

Some Transformation Examples (2)

> Categorical to numeric

- One-hot encoding (binarization of categorical values) Takes one categorical column and creates multiple numeric (Boolean) columns. The Boolean values can be promoted to the integers 0 and 1, which are numeric.
- Target encoding involves replacing a categorical feature with average target value of all data points belonging to the category

> Numeric to Categorical

- Decode numeric ID to Category

> Numeric to Numeric

- Log transformation (non-linear transformation)
- Addition of 2 columns
- Normalization takes a numeric column and replaces it with a scaled version of the column. Normalization of data puts all features on a similar scale to prevent features from dominating because their numbers are larger

Binning

Open: Lesson_05_a_Binning.ipynb

Break

Data Preparation (Pre-processing)

Open: Lesson_05_b_Student.ipynb

Break

Data Preparation (Pre-processing)

Open: Lesson_05_b_Student.ipynb

Normalization

Open: Lesson_05_c_Normalization.ipynb