
People with no idea about AI, telling me my AI will destroy the world
Me wondering why my neural network is classifying a cat as a dog..
Dog

Source

Welcome to Week 5 Lecture 2!

Data Science in Python & Machine Learning

Assignments Due This Week by Sunday night

1. Abalone Preprocessing Exercise (Core)
2. Pipelines Activity (Core)
3. Project 1 - Part 5 (Core)

Learning Goals:

1. Explain the three main types of variables we will be working with in this stack.
2. Convert data types into the correct format for machine learning
3. Implement data preparation in Python

Warm Up

What are the 3 primary kinds of variable features we learned about in the last stack.

Data Preprocessing

It must all be numbers.

Data Preparation

Interval/Continuous/Quantitative variables: Numbers that really mean numbers (int or float)

Examples:

- Values of length in cm: [2, 3.5, 8]
- Income in dollars: [57000, 38000]

Solution: No further preparation needed!

Note: Sometimes numeric data is in string form, such as ‘2’. This should be dealt with in data cleaning.

Data Preparation

Ordinal Categorical Variables: categories that can be ranked

Examples:

- low/medium/high,
- strongly disagree/disagree/agree/strongly agree
- $5-10, $11-15, $16-20, >$20

Solution: Convert to numbers representing the rank (sometimes your best judgment is required):

Example:

```
df[ordinal_column].replace({‘low’:0, ‘medium’:1, ‘high’:2}, inplace=True)
```

Data Preparation

Categorical Nominal variables: different groups with no order.

NOTE: Sometimes these are represented as integers in the data, but should be interpreted as categories.

Examples:

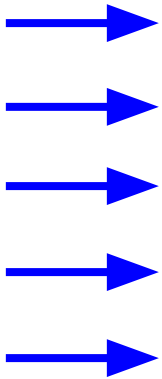
- Yes/No,
- Male/Female/Nonbinary,
- Fat free/Regular
- Teacher A, Teacher B, Teacher C, Teacher D,

Solution: One-hot encoding

One hot encode

Each category becomes its own column

Genre
Drama
Comedy
Drama
Documentary
SciFi



Comedy	Drama	Documentary	Scifi
0	1	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

One-Hot Encoding Can Cause Data Leakage

How can data from the test set leak into the training set with one-hot encoding?

Consider this:

What if when you split the data there are categories in a nominal column that are present in the test set, but not the training set?

If you encoded the column before splitting, there would be a column assigned to that category, even though it's not present in the training data!!

That's data leakage.

So, **we encode after splitting.**

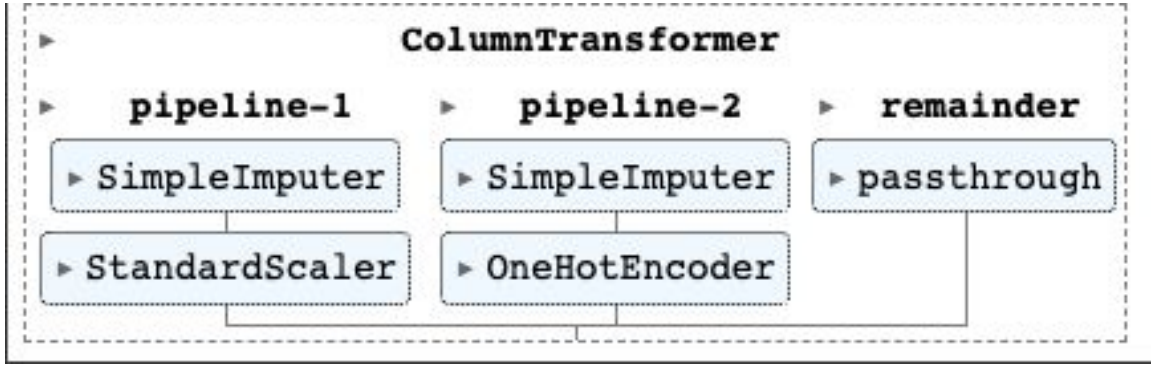
Transforming Data in Python:
SimpleImputer (imputes missing data)

Imputes missing values in many columns at once:

- ‘mean’, ‘median’, or ‘mode’ for numeric values
- ‘most frequent’, or constant for numeric and categorical values
- Can be used on a subset of rows

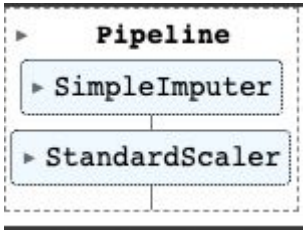
Make Column Transformer

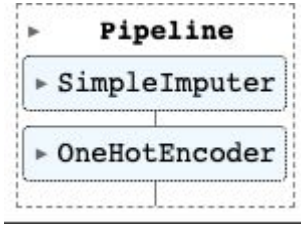
Make Column Transformer allows you to process numerical and categorical columns with the appropriate strategy. Can allow you to process different columns differently.

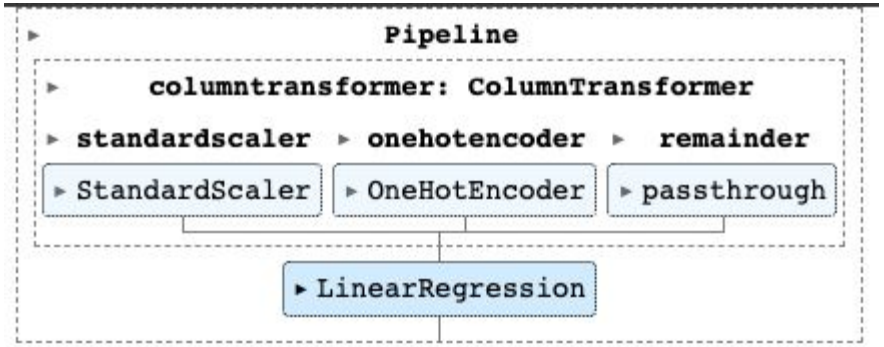


Pipelines

Pipelines allow you to chain together steps after a train test split that would normally require several lines of code without it.







Import Libraries

```
# Imports
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.compose import make_column_selector, make_column_transformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.pipeline import make_pipeline
from sklearn.impute import SimpleImputer
```

Pipelines and Tuples

```
num_selector = make_column_selector(dtype_include='number')
cat_selector = make_column_selector(dtype_include='object')

median_imputer = SimpleImputer(strategy='median')
freq_imputer = SimpleImputer(strategy='most_frequent')

scaler = StandardScaler()
ohe_encoder = OneHotEncoder(sparse=False, handle_unknown='ignore')

num_pipe = make_pipeline(median_imputer, scaler)
cat_pipe = make_pipeline(freq_imputer, ohe_encoder)

num_tuple = (num_pipe, num_selector)
cat_tuple = (cat_pipe, cat_selector)
```

Column Transformer

make_column_transformer(cat_tuple, remainder='passthrough')

By default ColumnTransformer **DROPS** any columns not specified in the steps.

```
[65]  1 col_trans = make_column_transformer(num_tuple, cat_tuple, remainder = 'passthrough')
```

Now you know what you need to know to prepare a dataset for modeling!

CodeAlong Notebook

CodeAlong DataSet

Challenge!

[Tonight's Challenge Notebook]

[Tonight's Challenge Dataset]