LECTURE

Feature Engineering

Introducing Feature Engineering

Sean Kang

What is Feature Engineering

- Process of extracting useful features from raw data using math, statistics, or domain knowledge
- There are some automated tools to detect those meaningful data

Why is Feature Engineering Important

- To make our model more sophisticated to handle more complex data.
- Better understanding of underlying data
- Faster processing

What can we create

We can create feature X data from

- Quantitative data
- Categorical data
- Text data

Feature Engineering
Outlier Detection

Outlier Detection

Some data that is very different from most of the data.

Some of the data is real but there are chances that they are errors. Removing or fixing the data can improve the model creation.

* ie. Human age, 1000 years old
Methods: statistical: percentage, standard deviations, visual.

What is Percentile?

- This is different from percent
- These are IQ values. How to determine the 25 percentile, 75 percentile, etc.
- Looking at 0 and 100 would locate some outliers.

68	75	78	83	85	85	85	86	86	87
84	88	90	91	91	91	91	93	93	93
94	94	94	96	96	97	98	98	99	99
99	99	100	101	101	102	102	104	104	105
105	105	106	106	106	107	107	107	107	107
108	109	110	110	111	114	116	116	117	122
123	128	136	141						

How to do it in Python Numpy

Python3

```
import numpy as np

# 1D array
arr = [20, 2, 7, 1, 34]
print("arr : ", arr)
print("50th percentile of arr : ",
      np.percentile(arr, 50))
print("25th percentile of arr : ",
      np.percentile(arr, 25))
print("75th percentile of arr : ",
      np.percentile(arr, 75))
```

Using Data frame with Percentile (demo)

In [1]:
```
import pandas as pd

df = pd.read_csv("heights.csv")
df.head()
```

Out[1]:

	name	height
0	mohan	5.9
1	maria	5.2
2	sakib	5.1
3	tao	5.5
4	virat	4.9

Another Technique: One hot encoding

Building a matrix of on and off values

- The US, Japan, europe mpg dataset.
- Two approaches:
  - Pandas
  - SciKit Learn

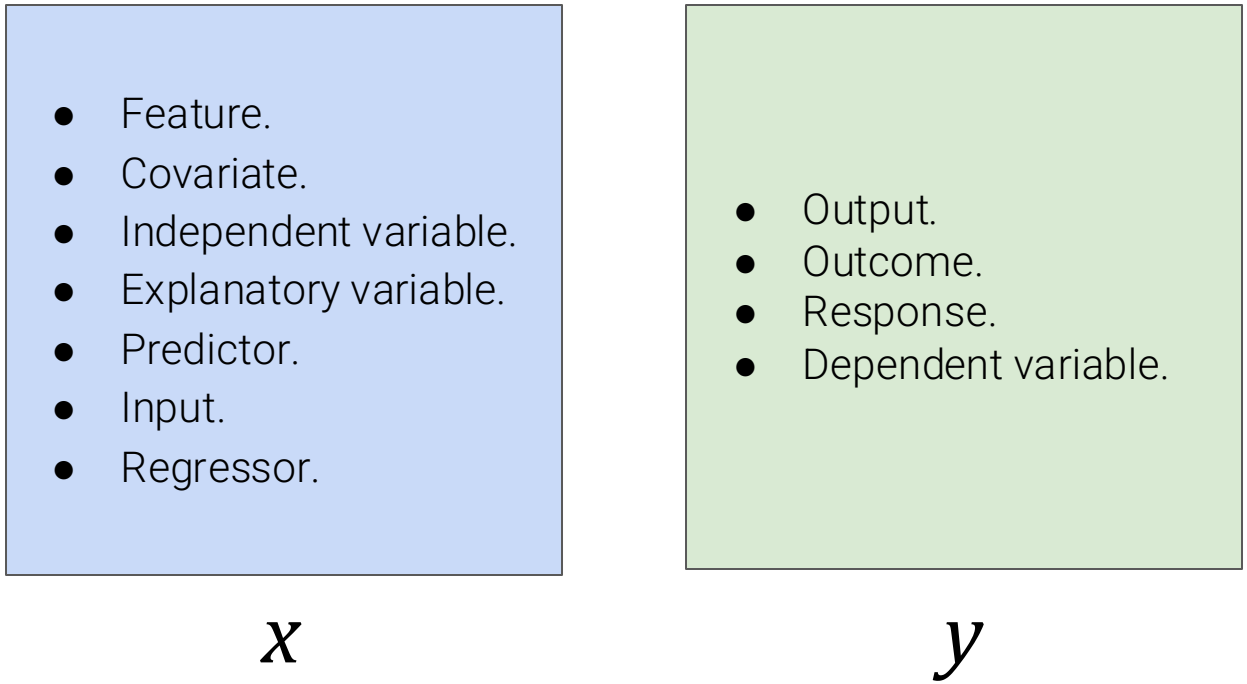
Handling Missing Values

- Using mean values for that column -- horsepower

Revisit: Multiple linear regression

Terminology

There are several equivalent terms in the regression context. You should be aware of them.



Adding independent variables

First, some terminology. For our purposes, all of these terms mean the same thing:

- Feature.
- Covariate.
- Independent variable.
- Explanatory variable.
- Predictor.
- Input.
- Regressor.

In the regression context, each of the above things has a "**weight**" assigned to it, given by the **parameter**. We also call these weights "**coefficients**." For instance, in $\hat{y} = \theta_0 + \theta_1 x$, we might say the "weight" associated with the constant/intercept term is θ_0, and the "weight" associated with the x term is θ_1

Multiple linear regression

Model 1: $\text{predicted PTS} = 3.98 + 2.4 \cdot \text{AST}$

different!

Model 2: $\text{predicted PTS} = 2.163 + 1.64 \cdot \text{AST} + 1.26 \cdot \text{3PA}$

These are different models! In general, $\hat{\theta}_j$ in one model will not be equal to $\hat{\theta}_j$ in another model.

- 2.4 is the slope of the relationship between AST and PTS, when only considering those two variables.
  - Parameters [3.98, 2.4] minimize average squared loss for Model 1.
- 1.64 is the slope of the relationship between AST and PTS, when also considering 3PA.
  - Parameters [2.163, 1.64, 1.26] minimize average squared loss for Model 2.

Summary

Summary

- We now know of three models, $\hat{y} = f_\theta(x)$.
  - The constant model, $f_\theta(x) = \theta$.
  - The simple linear regression model $f_\theta(x) = \theta_0 + \theta_1 x$.
  - The multiple linear regression model $f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$.
  - A model with optimal parameters is denoted $f_{\hat{\theta}}(x)$.
- We looked at the correlation coefficient, r, and studied its properties.
- We solved for the optimal parameters for the simple linear model by hand and by code, by minimizing average squared loss (MSE).

$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x} \qquad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

- We introduced the notion of a feature, and how we can have multiple in our models.
- We discussed the multiple R^2 coefficient and RMSE as methods of evaluating the quality of a linear model.