Linear Regression

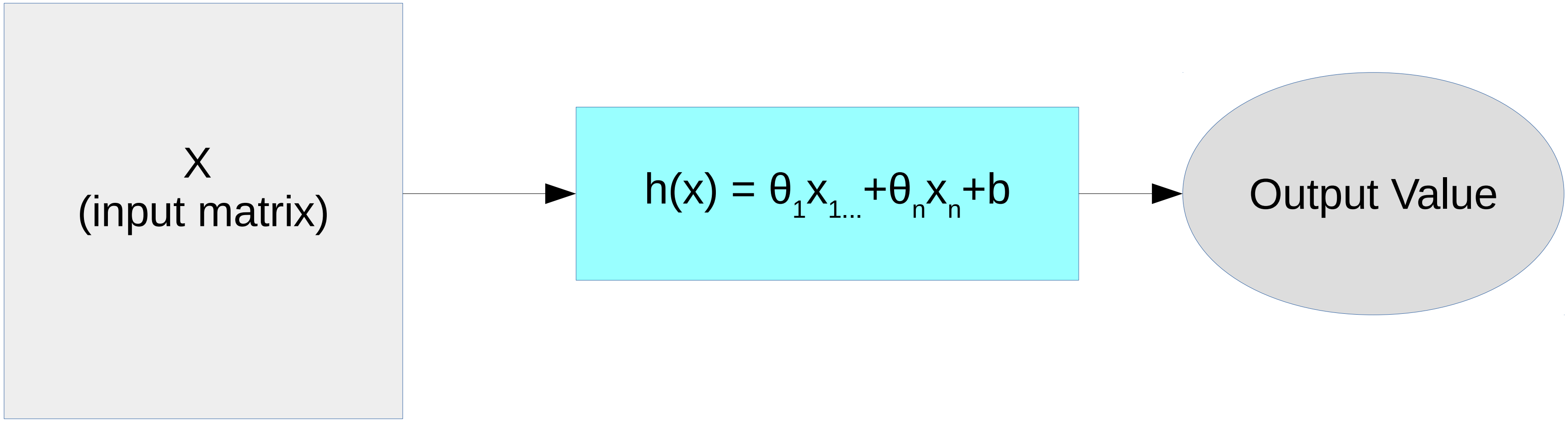
UIS CS570:Essentials of Data Science
Mike Bernico

The 5 Questions

- What does it do?
- What are the inputs?
- What are the outputs?
- How can we measure performance?
- How can it fail?

What Does it Do?

- Using a linear model, predicts a continuous output value based on inputs.



X
(input matrix)

$h(x) = \theta_1 x_{1\ldots} + \theta_n x_n + b$

Output Value

What are the Inputs

Scaled Numerical Values

X_m = .5, .2, .7

- Two Problems:
  - What does scaled mean?
  - How do you convert categorical variables to numbers?

'Scaled'

- All variables should be of the same scale or magnitude.
- Use Z score

$$z = \frac{x - \mu}{\sigma}$$

'Numercial'

- We already know how to handle text
- Continuous Variables are easy enough
- What about Categorical Values?

- Consider the Sex Category from Titanic

  Female = 1, Male = 0

  - Transforming the variable into 'indicator of female'

Handling Complex Categories

$\text{pClass} \in 1, 2, 3$

pClass_1	pClass_2	pClass_3
0	1	0

pClass_2	pClass_3
1	0

'one hot encoding'

What are the Outputs

- A continuous value that we want to predict

- Regression, not classification

Performance

- Hold Out Set (Test/Train Split)

- For test, we know the dependent variable but we will hide it from the computer. Let's call it y

- We will show the computer X, and let it predict y. Let's call this predicted value $\hat{y}$
- We will compare y and $\hat{y}$

- The metrics we will use to compare the two will be RMSE and The Coefficient of Determination.

How Can It Fail?

- Assumes the Model is Linear
  - We can adjust this by added polynomial terms manually.
  - But this opens up the potential for overfitting
    - Watch the Video on Bias/Variance!
    -
- Assumes i.i.d
  - Independent and identically distributed
  - Gauss-Markov assumptions for BLUE