

# Advanced Data Analysis

DATA 71200

Class 4

# Schedule

**6-Jun      Representing Data**

7-Jun      Async: DataCamp

8-Jun      Evaluation Methods

9-Jun      Async: DataCamp

# Reading for today

- ▶ **Ch 4: "Representing Data/Engineering Features" in Guido, Sarah and Andreas C. Muller. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc. 213–55.**

# Inspecting Data to Gain Insights

- ▶ **Review from last week**

- Data size and type
- Summary statistics
- Histograms
- Scatter Matrix

# Representing Data

- ▶ **Continuous versus categorical**
  - One-Hot Encoding
  - Binning
- ▶ **Transformations**
- ▶ **Automatic feature selection**
- ▶ **Utilizing expert knowledge**

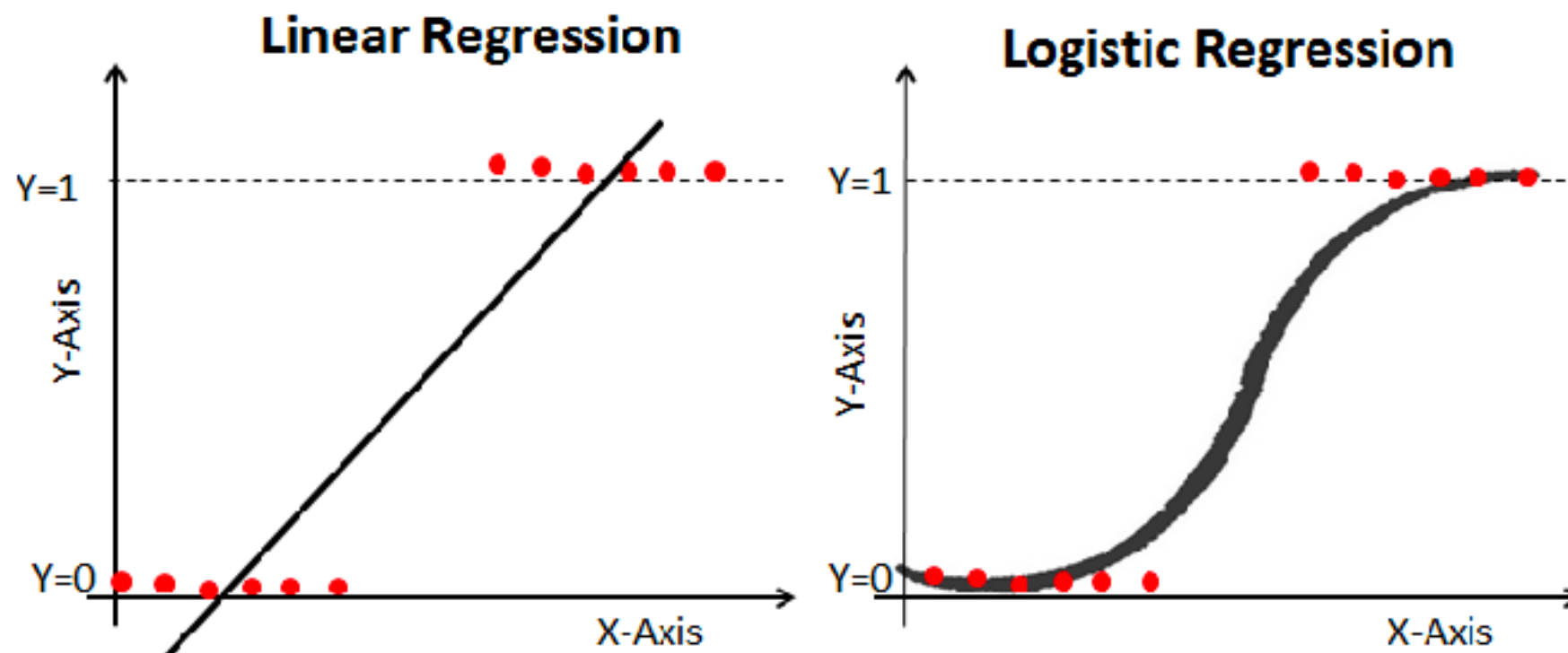
# Some Terminology

## ▸ (Linear) Regression

- Continuous predictive model created by estimating a linear relationship between features

## ▸ Logistic Regression

- Predictive model of the probability of a certain class



# Some Terminology

## ► **Regularization**

- Adds an extra term to the cost function
- Can be applied to linear and logistic regression
- Can also be used for feature selection
- Lasso (least absolute shrinkage and selection operator) regression is another form, referred to as L1
- Ridge is a form of regularization, referred to as L2

# Some Terminology

- ▶ **Lasso Regression (L1)**

- reduces the coefficients of the least important variables to zero (removing them completely by the model)

- ▶ **Ridge Regression (L2)**

- addresses *multicollinearity* (linear relationships between parameters) and having more parameters than observations



# Continuous Versus Categorical

- ▶ **(Linear) Regression - predicts continuous values**
- ▶ **Classification - predicts categorical, or discrete, values**
- ▶ **Continuous versus categorical distinction also holds for input features**

# One-Hot Encoding

- ▶ **Split the different categories in their own variable**
- ▶ **E.g., a single variable for color where the values are the strings “blue”, “red”, “yellow” would be encoded as**

	Blue	Red	Yellow
Blue	1	0	0
Red	0	1	0
Yellow	0	0	1

← **Variables**

**Values** ↑

*Categorical data can also be encoded as numbers*

	Categorical Feature	Integer Feature
0	socks	0
1	fox	1
2	socks	2
3	box	1

**1      2      3      4      5      6**  
**0      1      2      3      4      5**

**[ [ 0 .   0 .   1 .   1 .   0 .   0 . ]**  
**[ 0 .   1 .   0 .   0 .   1 .   0 . ]**  
**[ 0 .   0 .   1 .   0 .   0 .   1 . ]**  
**[ 1 .   0 .   0 .   0 .   1 .   0 . ] ]**

**box   fox   socks   0      1      2**

# In-Class Activity 1

- ▶ **Apply one-hot encoding to the ocean\_proximity value in the California Housing dataset that we looked at last class**
  - Using `pd.dummies` and/or `OneHotEncoder` from `scikitlearn`
  - `housing['ocean_proximity'].values.reshape(-1,1)`

# Binning

- ▶ **Discretizing continuous data into numerical bins can be useful when small differences in value are not significant**
- ▶ **E.g., for numerical grade data (out of 100), it may be more useful to give a model how many scores fall into ranges of 5 rather than the continuous data**

82	83	92	93	72	73	87	86	99	97	98	51	52	82	81	87	91	92	61	67	
										↓										
50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99											
2	0	1	2	2	0	4	3	4	2											

# In-Class Activity 2

- ▶ **Apply binning to the housing\_median\_age value in the California Housing dataset that we looked at last class**
  - `housing['housing_median_age'].values.reshape(-1, 1)`
  - Plot both the original data and the binned data
- ▶ **Explore binning with other features**

# Transformations

- ▶ **Squaring and cubing is useful for linear regression models**
- ▶ **Logarithms and exponentials are useful for representing your data with a Gaussian distribution, which is useful for mean-based models**

# In-Class Activity 3

- ▶ **Apply the following transformations to housing\_median\_age in the California Housing dataset that we looked at last class**
  - Squaring (\*\*2)
  - Cubing (\*\*3)
  - np.log
  - np.exp
- ▶ **Plot histograms and scatter matrices to explore the resultant data (for \*\*2, \*\*3, and np.log)**



# Automatic Feature Selection

- ▶ **Regularization can be used to assess the relative importance of features in the performance of a model**
  - Although this can't tell you anything about features you don't include
- ▶ **Recursive feature elimination (RFE) starts with all features and removes the poorly performing ones**
- ▶ **You can also start with one feature and build up a model**

# Utilizing Expert Knowledge

- ▶ **Domain knowledge can be useful for recognizing patterns in data that may be beneficial or detrimental to the model**
- ▶ **This can inform decisions about which features to include and how to represent them**

# DataCamp for tomorrow

- ▶ *Introduction to Python (If Needed)*
- ▶ AI Fundamentals
  - Introduction to AI
- ▶ Data Manipulation with pandas
  - Transforming Data
  - Aggregating Data
  - Slicing and Indexing
  - *Creating and Visualizing DataFrames (Optional)*
- ▶ *Writing Efficient Code with pandas (Optional)*

# Reading for Wednesday

- ▶ Ch 5: “Model Evaluation and Improvement” in Guido, Sarah and Andreas C. Muller. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

# Project 1

- ▶ **Due June 13**
- ▶ **Keep exploring potential datasets**
  - [kaggle.com](https://www.kaggle.com)
  - [archive.ics.uci.edu/ml/datasets.php](https://archive.ics.uci.edu/ml/datasets.php)
  - [libguides.nypl.org/eresources](https://libguides.nypl.org/eresources)
  - [opendata.cityofnewyork.us/data/](https://opendata.cityofnewyork.us/data/)
- ▶ **The data set will need to be labeled as you are going to use it for both supervised and unsupervised learning tasks**
- ▶ ***We will go over using IMPUTER to address missing values on Wednesday***