Advanced Data Analysis

DATA 71200

Class 5

Course Schedule

4-Mar	**Representing Data**
11-Mar	Evaluation Methods
18-Mar	Supervised Learning (k-Nearest Neighbors, Linear Models) *Project 1 Due*
25-Mar	Supervised Learning (Naive Bayes Classifiers and Decision Trees)
1-Apr	Supervised Learning (Support Vector Machines and Uncertainty estimates from Classifiers)
7-Apr	Unsupervised Learning (Dimensionality Reduction & Feature Extraction, and Manifold Learning) *Project 2 Due*

Assignments for this week

- **DataCamp**
  - Preprocessing for Machine Learning in Python
    - Introduction to Data Preprocessing
    - Standardizing Data
    - *Feature Engineering* (March 11)
    - *Selecting features for modeling* (March 11)

- **Reading**
  - 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.

DATA 71200: Project 1 (Due March 18)

The goal for this assignment is for you to create a usable dataset from an open-source data collection that you will use for a supervised classification task in Project 2 and with unsupervised learning in Project 3.

Step 1: Find and download a dataset. Here are some potential places to look

- Amazon's AWS datasets: https://aws.amazon.com/opendata/public-datasets/
- Data Portals: http://dataportals.org/
- Kaggle datasets: http://kaggle.com
- NYPL digitizations: http://libguides.nypl.org/eresources
- NYC Open Data: http://opendata.cityofnewyork.us/data/
- Open Data Monitor: http://opendatamonitor.eu/
- QuandDL: http://quandl.com/
- UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php

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

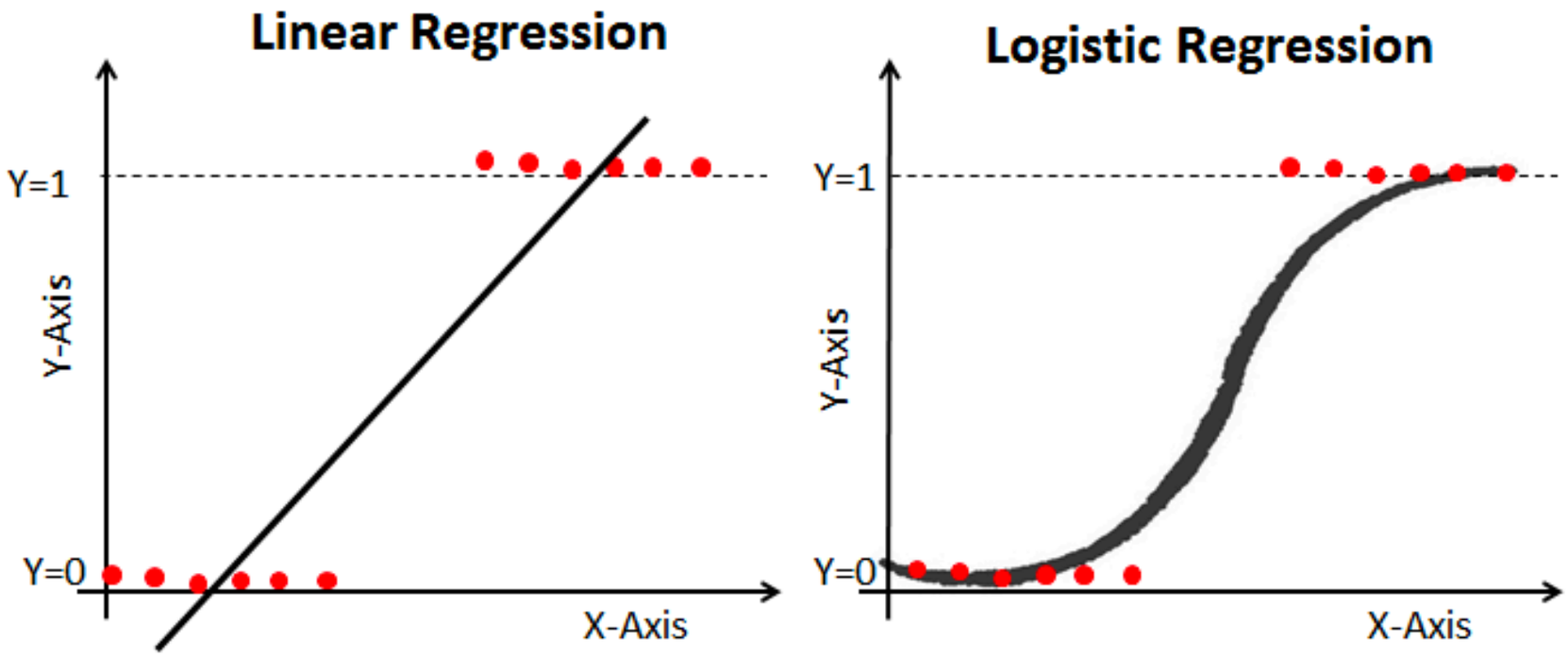

Image: https://medium.com/@hpsuresh12345/logistic-regression-60694a973bee

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 at L2

Some Terminology

- **Ridge Regression**
  - Predictive model that addresses *multicollinearity* (linear relationships between parameters) and having more parameters than observations

Continuous Versus Categorical

- Regression - predicts continuous values
- Classification - predicts categorical, or discrete, values
- Continuous versus categorical distinct 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

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

Binning

- **Discretizing continues 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**

Assignments for next week

- **DataCamp**
  - Preprocessing for Machine Learning in Python
    - Feature Engineering
    - Selecting features for modeling
    - Putting it all together
- **Reading**
  - Ch 5: “Model Evaluation and Improvement” in Guido, Sarah and Andreas C. Muller. (2016). Introduction to Machine Learning with Python, O’Reilly Media, Inc.