GENERAL ASSEMBLY

INTRODUCTION TO CLASSIFICATION: K-NEAREST NEIGHBORS

Matt Speck, Data Science Immersive, h/t Joseph of Iowa

AGENDA

- What is Classification?
- Introduction to K-Nearest Neighbors (KNN)
- KNN Examples/Applications
- Coding Implementation

WHAT IS CLASSIFICATION?

- Class guesses?

WHAT IS CLASSIFICATION?

- Classification methods are used to predict what ***class*** data points will fall into.

- How is this different from regression?

WHAT IS CLASSIFICATION?

- Classification methods are used to predict what ***class*** data points will fall into.
- How is this different from regression?
- **Regression** is used to predict *quantitative* targets
- **Classification** is used to predict *qualitative* targets

WHAT IS CLASSIFICATION?

- Check: I have a home with X bedrooms, Y sq ft, Z lot size. What is the price of this home? Is this a classification or regression problem?

- Check: I have an unknown fruit that is 5.5 inches long, 2 inches in diameter, and yellow. What is this fruit? Is this a classification or regression problem?

- Check: I have a person who works in X industry, has Y years of experience, and lives in Z district. Does this person make above the median US annual salary? Is this a classification or regression problem?

WHAT IS CLASSIFICATION?

- Let’s see a bunch of these classification methods in action. Back to the repo…

INTRODUCTION TO K-NEAREST NEIGHBORS

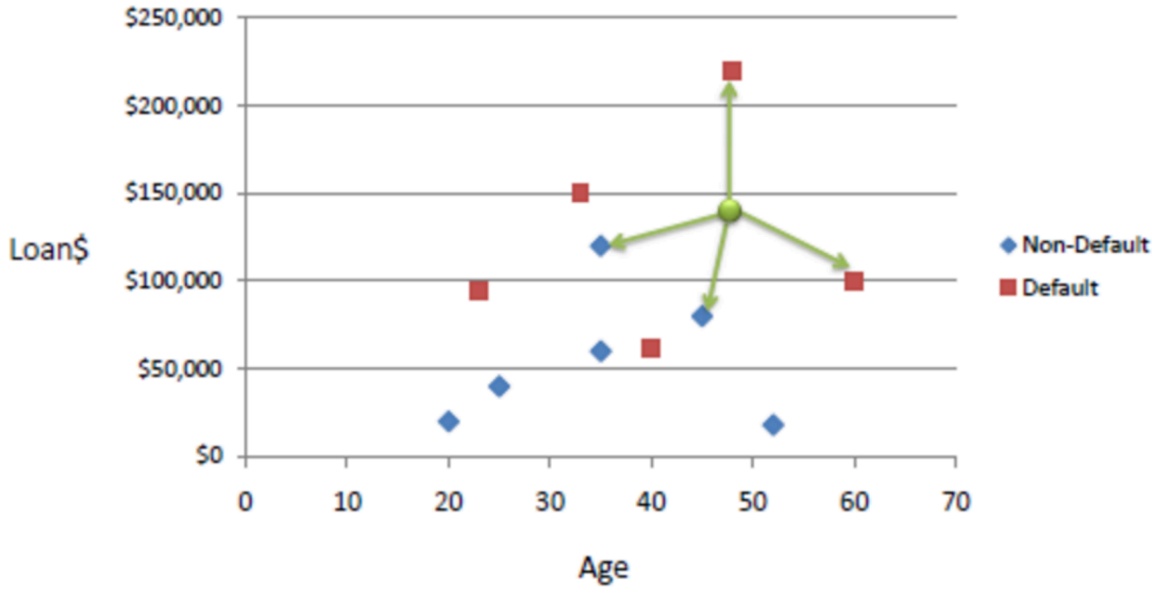
- KNN is a **non-parametric, lazy** learning algorithm that predicts outcomes based on the **similarity (near-ness)** of inputted features to the training set
- Non-parametric: Makes no assumptions about the underlying distribution of our data
- Lazy: Training phase is minimal – KNN uses all (or nearly all) of the training data
- Based on feature similarity: How closely out-of-sample features resemble our training set determines how we classify a given data point
- Because of this above, KNN can be thought to be a spatial algo

INTRODUCTION TO K-NEAREST NEIGHBORS

- procedure KNN(x)
- Begin looping through all known data points in training data, find the closest k points to x
- assign f(x) = majority classification among the k closest points
- end

EXAMPLES AND APPLICATIONS

- Consider determining if an individual is going to default on their loan. Age and Loan are the two numerical variables (predictors) and Default is the target



SELECTING OUR VALUE OF K

- How does K affect our bias-variance tradeoff?
- http://scott.fortmann-roe.com/docs/BiasVariance.html

ADVANTAGES AND DRAWBACKS

- **ADVANTAGES:**
- Simple to understand and explain
- Model training is fast
- Can be used for classification and regression
- Non-linear
- **DRAWBACKS:**
- Must store all training data
- Prediction (testing) phase can be slow when n is large
- Sensitive to irrelevant features
- Accuracy is (generally) not competitive with best supervised learning models