GENERAL ASSEMBLY

INTRODUCTION TO CLASSIFICATION: K-NEAREST NEIGHBORS

Joseph Nelson, Data Science Immersive

AGENDA

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

WHAT IS CLASSIFICATION?

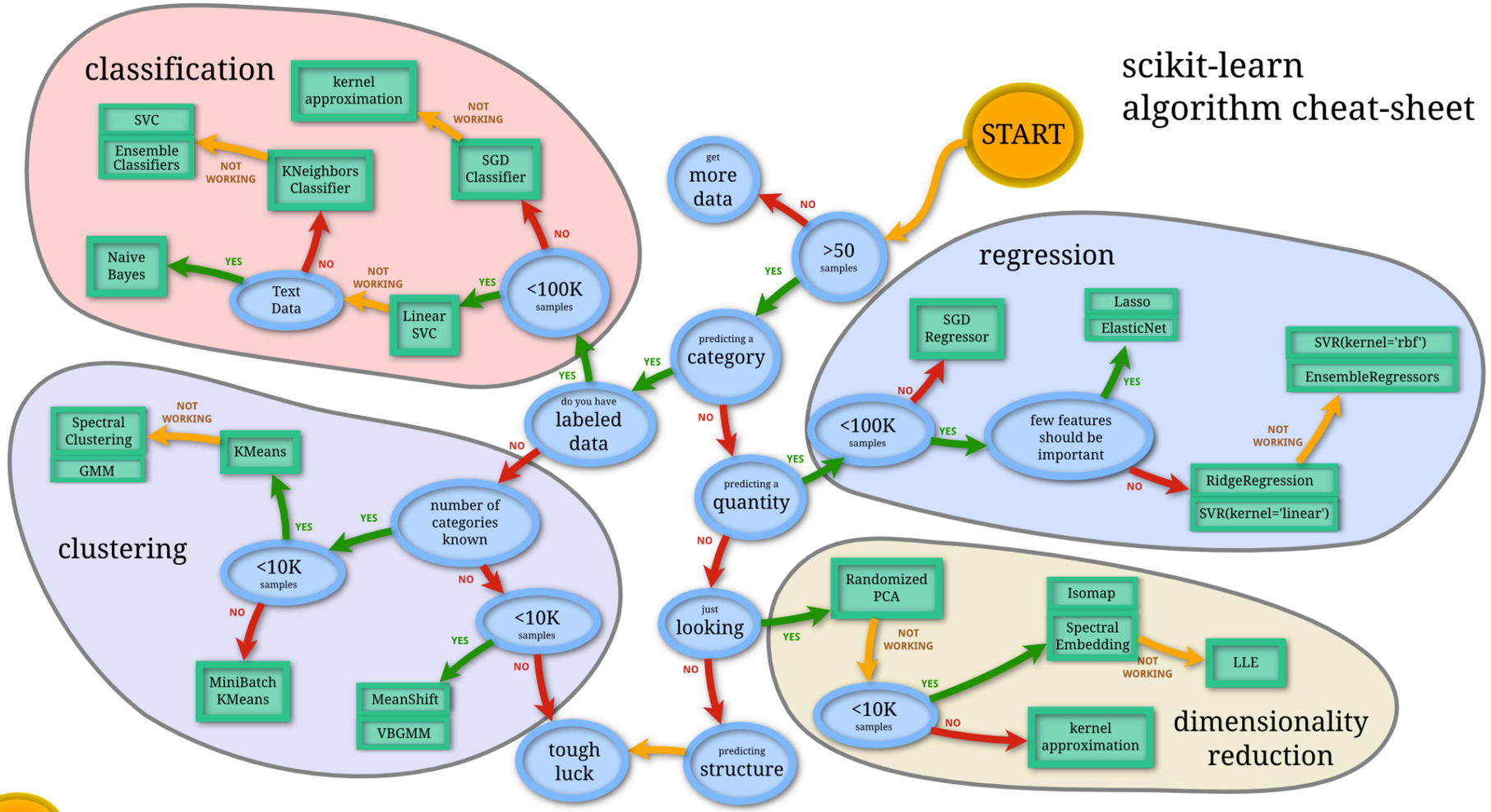
- Class guesses?

WHAT IS CLASSIFICATION?

- Regression is used to predict continuous values. Classification is used to predict which class a data point is part of (discrete value).
- Example 1: I have a home with X bedrooms, Y sq ft, Z lot size. What is the price of this home?
- Example 2: I have an unknown fruit that is 5.5 inches long, 2 inches in diameter, and yellow. What is this fruit?

- (Yes, that last slide was a pun)

WHAT IS CLASSIFICATION?



WHAT IS CLASSIFICATION?

- Let’s plot EIGHT different classification models. To the repo…

INTRODUCTION TO K-NEAREST NEIGHBORS

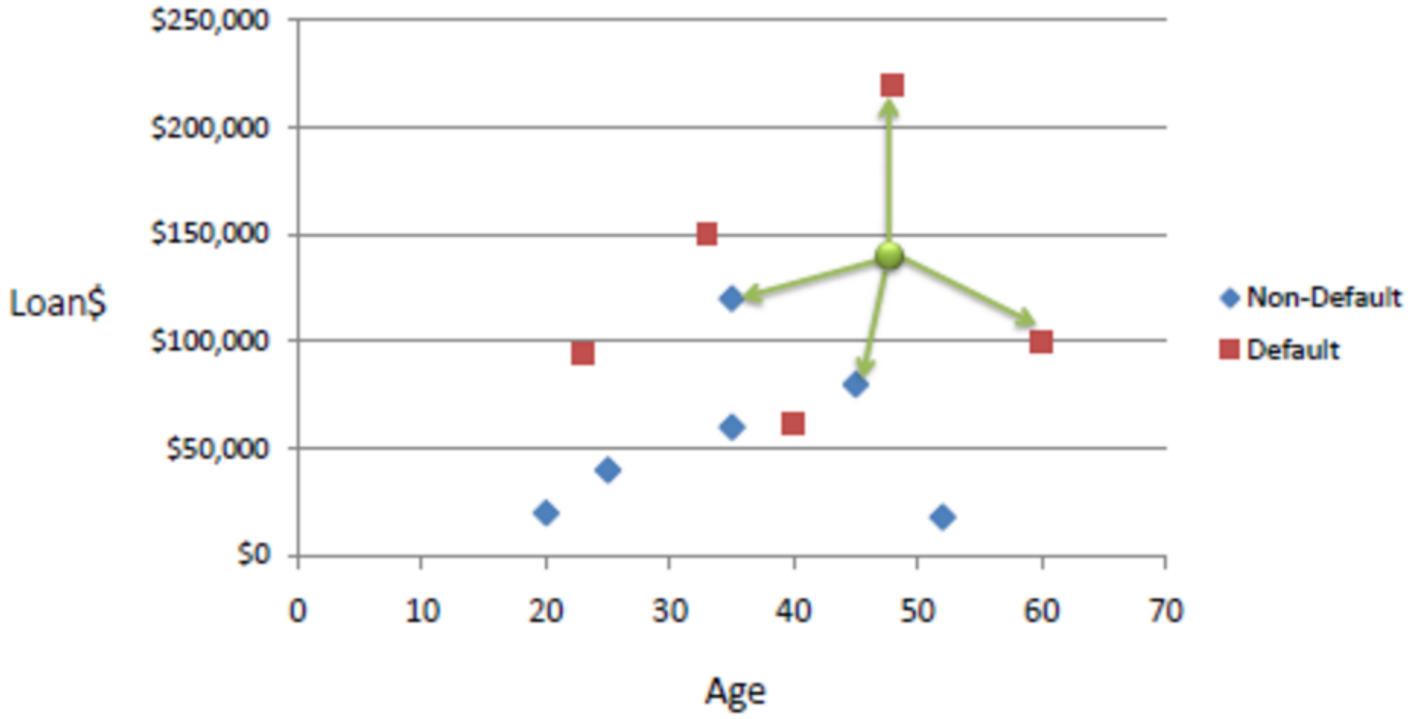
- KNN is an non-parametric lazy learning algorithm that predicts outcomes based on the similarity (near-ness) of inputted features to the training set
- Non-parametric: Makes assumptions about the underlying distribution of our data
- Lazy: Training phase is minimal – KNN uses all (or nearly all) of 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



EXAMPLES AND APPLICATIONS

- We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance. If K=1 then the nearest neighbor is the last case in the training set with Default=Y.

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y.

SELECTING A VALUE OF K

- How does our k affect our bias-variance tradeoff?

- http://scott.fortmann-roe.com/docs/BiasVariance.html

ADVANTAGES AND DRAWBACKS

Benefits

- Simple to understand and explain
- Model training is fast
- Can be used for classification and regression
- Non-linear, which may be common (imagine age vs income)

Drawbacks

Must store all of the training data

Prediction phase can be slow when n is large

Sensitive to irrelevant features

Sensitive to the scale of the data

Accuracy is (generally) not competitive with the best supervised learning methods