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?

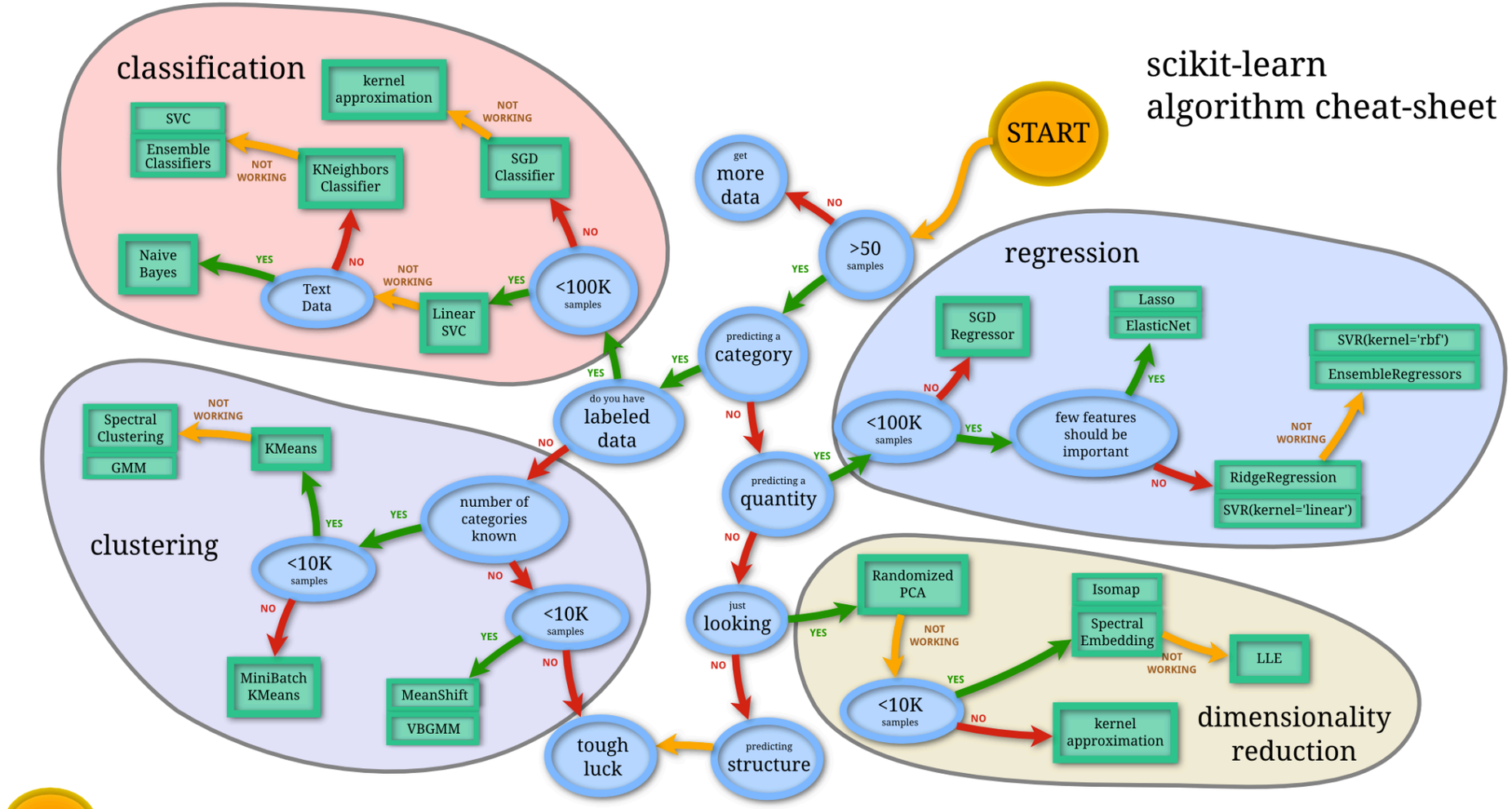
- Class guesses?

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

EXAMPLES AND APPLICATIONS

- A botanist has a meadow of flowers, but does not have the time to survey every single flower that is growing. He wants to find a quick way to sample them all.

- Benefits
- Simple to understand and explain
- Model training is fast
- Can be used for classification and regression

ADVANTAGES AND DRAWBACKS

Benefits

- Simple to understand and explain
- Model training is fast
- Can be used for classification and regression

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