GENERAL ASSEMBLY

DECISION TREES AND RANDOM FORESTS

Doug Friedman

TODAY'S LEARNING OBJECTIVES

- Understand and build decision tree models for classification and regression with the sklearn library
- Understand and build random forest models for classification and regression
- Know how to extract the most important predictors in a random forest model

COURSE

PRE-WORK

PRE-WORK REVIEW

- Know how to build and evaluate (classification) models in sklearn
- Knowledge of resampling methods
- Understand the concepts of cross-validation and overfitting

OPENING

DECISION TREES AND RANDOM FORESTS

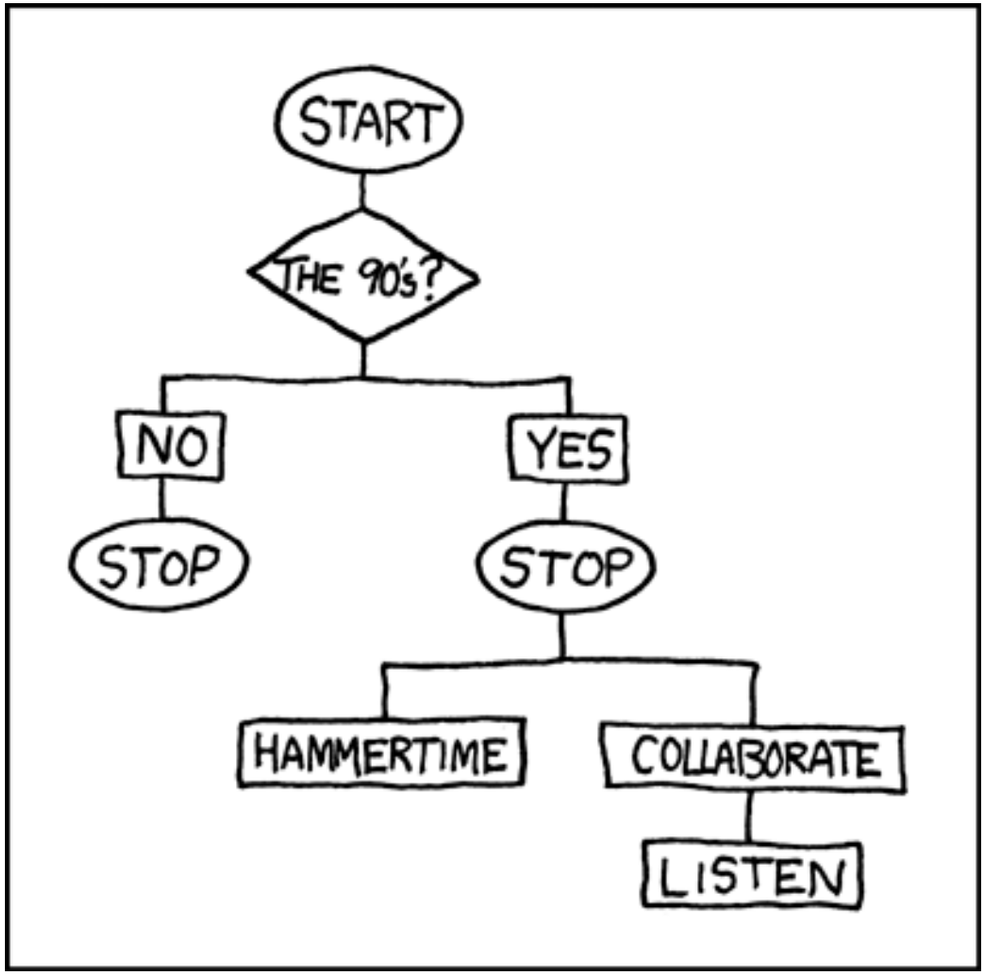
I LOVE (CLASSIFYING) THE 90s

[Verse 1]
All right stop, collaborate and listen
Ice is back I got a brand new invention
Something grabs a hold of me tightly
Flow like a harpoon daily and nightly
Will it ever stop? Yo - I don't know
Now turn off the lights (huh) and I'll glow
And to the extreme I rock a mic like a vandal
Light up a stage and wax a chump like a candle

Too Cold – Vanilla Ice (1998)

[Breakdown]
Stop!
Hammer time

U Can't Touch This – MC Hammer (1990)

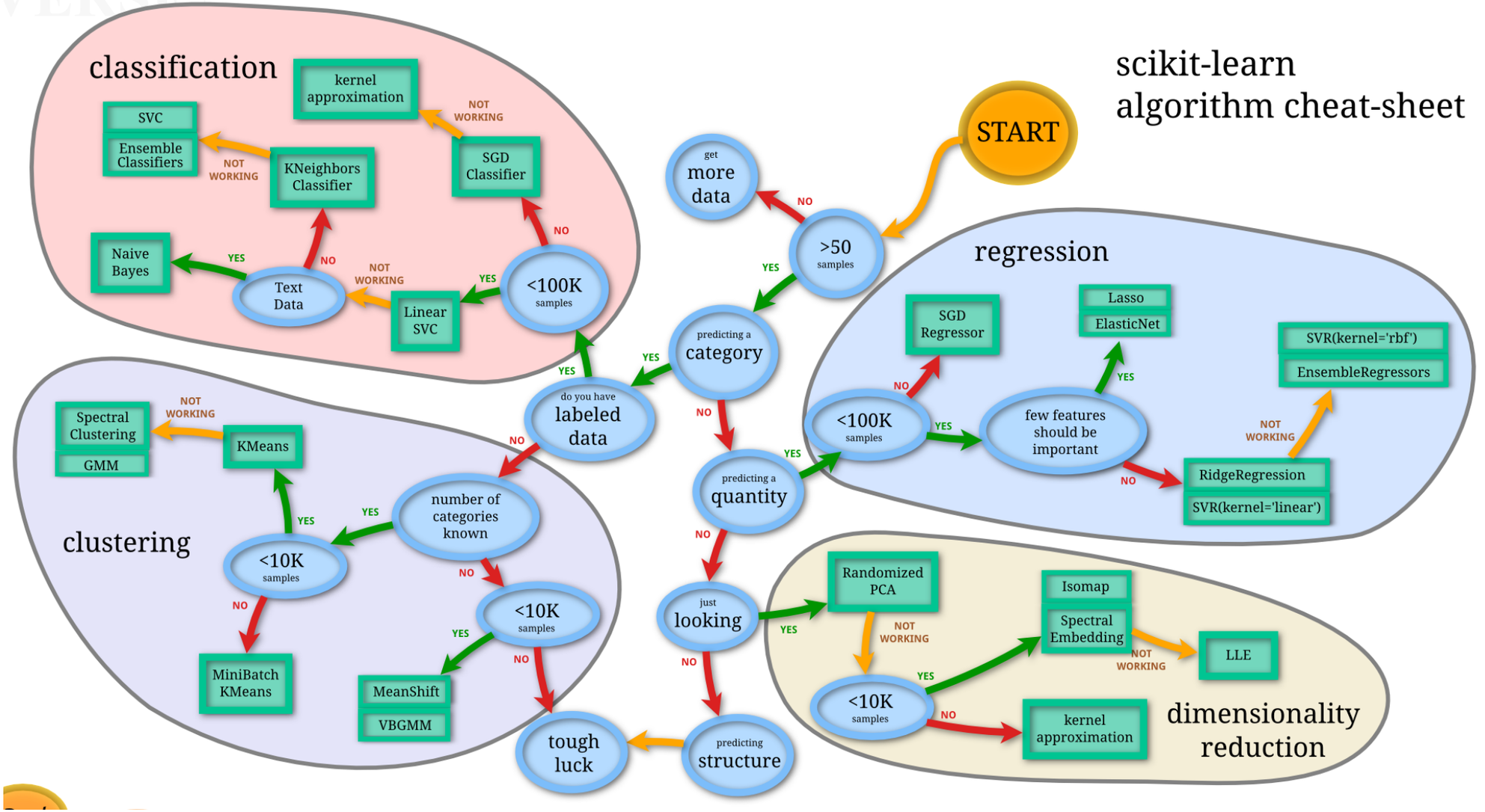


WHERE ARE WE IN THE DATA SCIENCE WORKFLOW?

- Data has been **acquired** and **parsed.**

- Today we'll **refine** the data and **build** models (We'll also use plots to **represent** the results).

WHERE ARE WE IN THE MACHINE LEARNING UNIVERSE?



GUIDED PRACTICE

VARIABLE IMPORTANCE

ACTIVITY: VARIABLE IMPORTANCE

ANSWER THE FOLLOWING QUESTIONS (10 minutes):

1. How do we identify the importance of different variables in regression models?

2. How do we identify the importance of different variables in classification models?

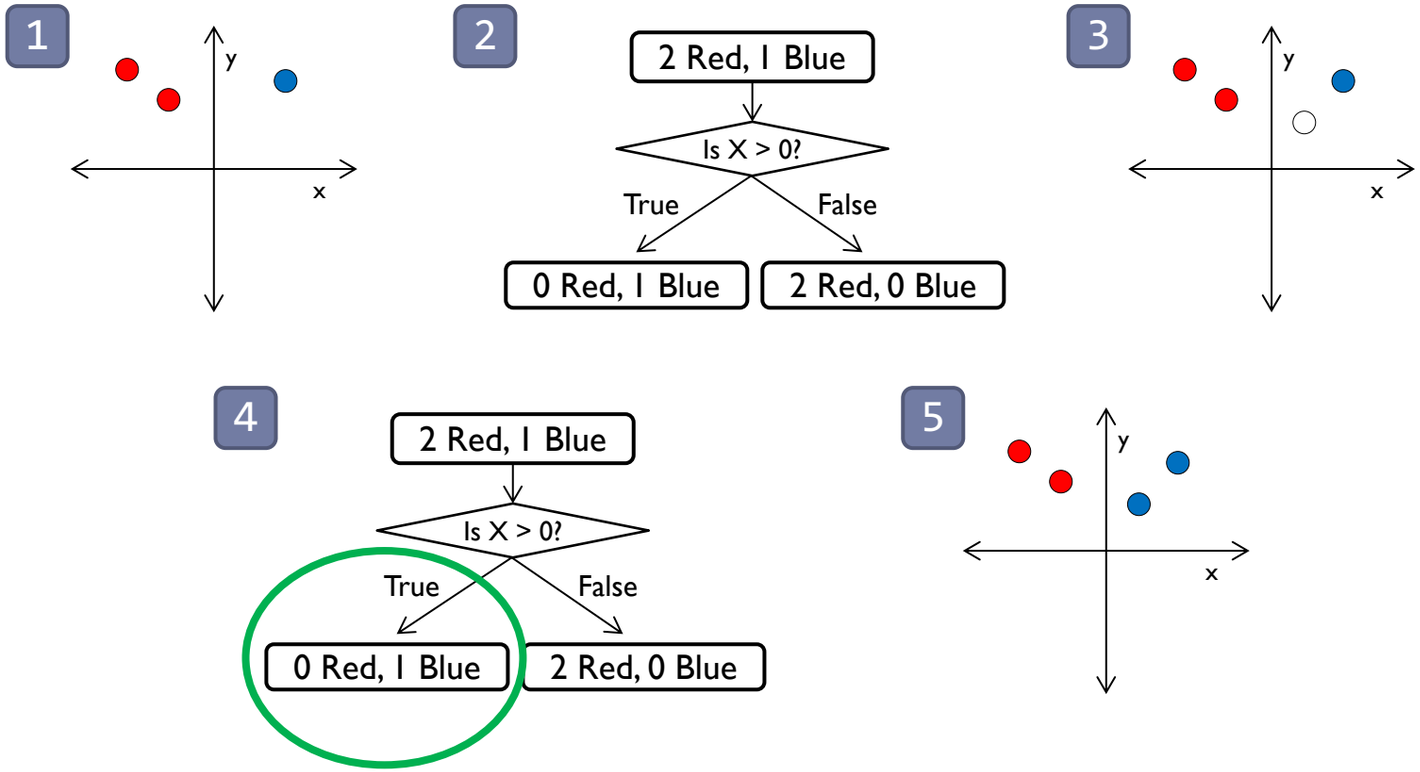
DELIVERABLE

Answers to the above questions

DECISION TREES

DECISION TREES

- Decision Trees are a machine learning model for regression and classification that develops *a series of yes/no rules* to explain the differences present in the outcome variable.

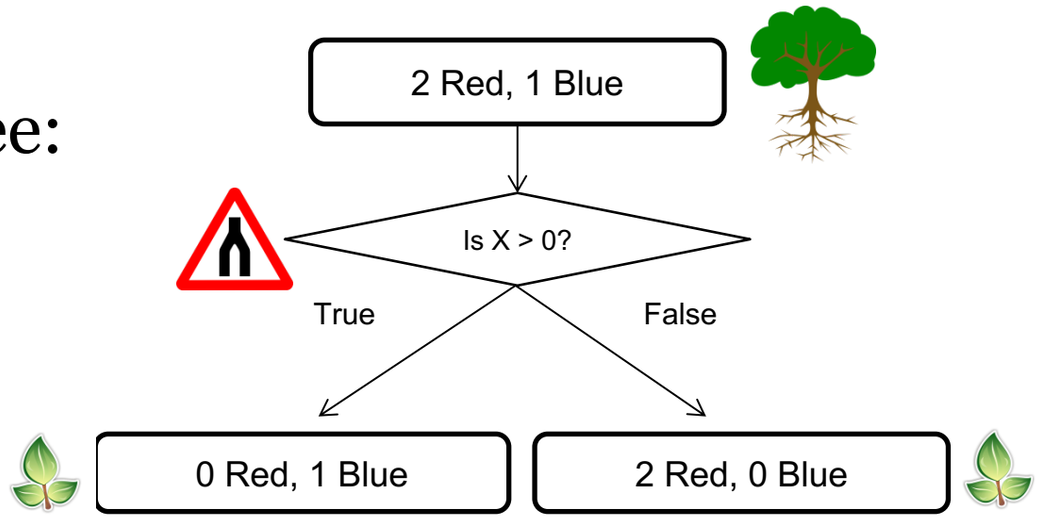


DECISION TREES

- When displayed, these series of rules appear as a tree with several branching paths or **splits**.

- The starting point of a decision tree is referred to as the **root** and subsequent branching points are called **nodes**. Nodes that do not split further are then called **leaves**.

- Using our example decision tree:



DECISION TREES

- The structure of a decision tree is determined by what yes/no rules will best predict the outcome variable.

- This is measured at each point of a decision tree by the **gini impurity** which measures the homogeneity of the outcome variable in a dataset from 0 (uniform) to 1 (inconsistent).

- Each rule in a decision tree decreases the gini impurity in the data until it approaches 0.

- For regression trees, MSE (or mean squared error) is *often* used in place of gini impurity.

GUIDED PRACTICE

GROW YOUR OWN TREES

ACTIVITY: GROW YOUR OWN TREES

DIRECTIONS (10 minutes):

1. Build decision trees for the two datasets on the next page.

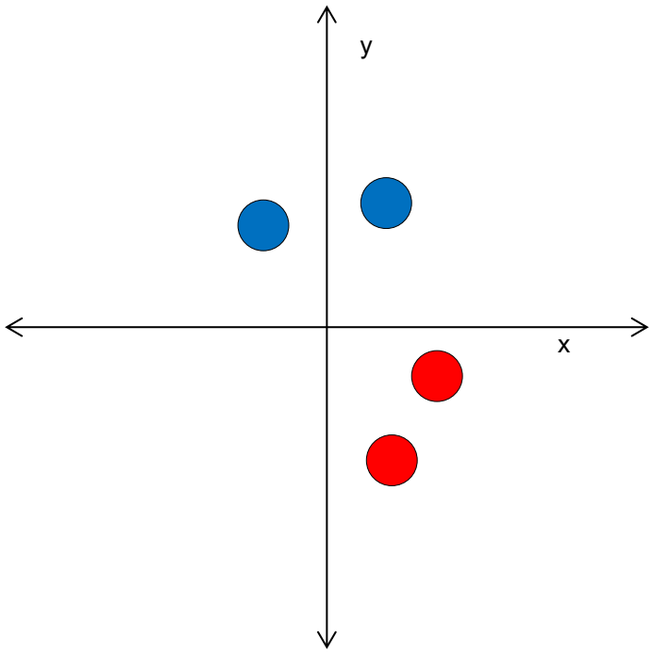
2. Once you're done, ask a neighboring team if they grew the same decision trees. Did they come up with a different result?

DELIVERABLE

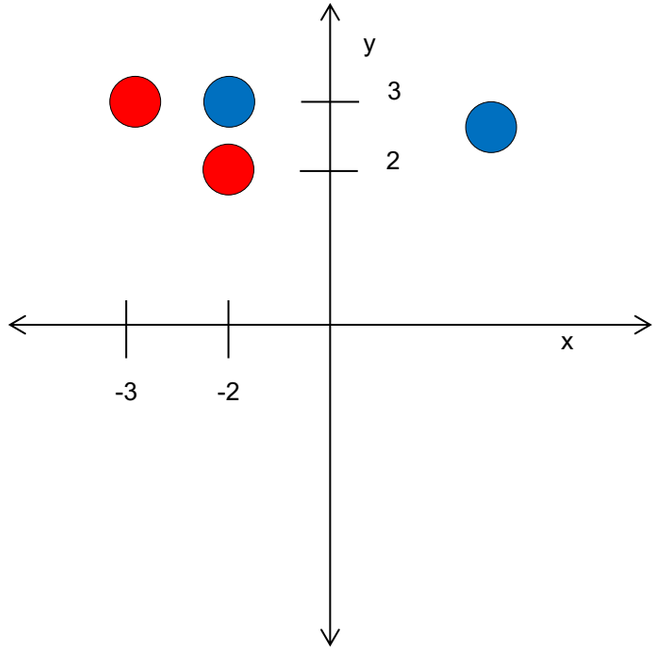
Answers to the above questions

ACTIVITY: GROW YOUR OWN TREES

Dataset A



Dataset B



Hint: You may want to use the tick marks!

DEMO

DECISION TREES

DECISION TREES

- Open up starter-code-12 and we'll take a look at how to build (and visualize) decision trees with the sklearn library.

GUIDED PRACTICE

TUNING TREES

ACTIVITY: TUNING TREES

DIRECTIONS (20 minutes):

Use the provided code to perform a grid search (with k-fold cross-validation) to identify the optimal parameters for a decision tree.

- Use the official sklearn documentation about decision trees to identify the correct range for your tuning parameters.
- Use either accuracy or the adjusted rand score to score your model.
- Visualize your final model using graphviz.

DELIVERABLE

Answers to the above questions

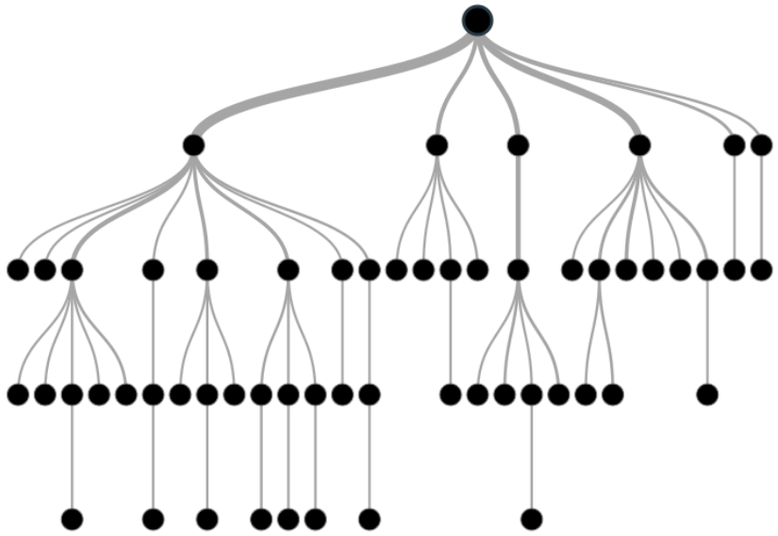
INTRODUCTION

PROS AND CONS OF DECISION TREES

PROS ~~AND CONS~~ OF DECISION TREES

- Decision trees are *non-linear* (a change in a predictor variable has a constant change on the output variable) which gives them more flexibility over linear models (e.g. linear regression).

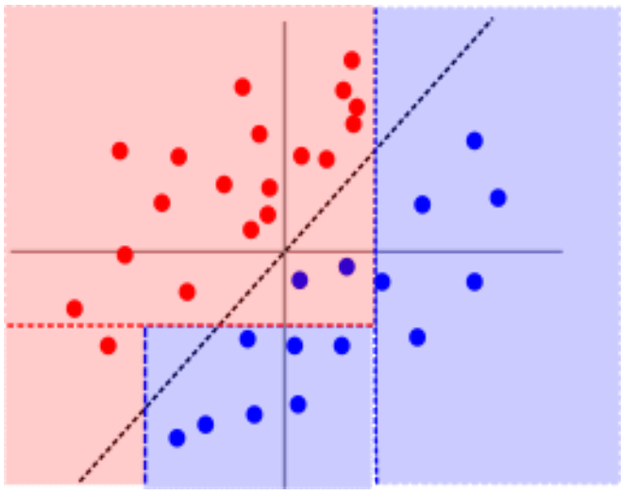
- Decision trees also produce easily interpreted visuals from which variable importance can be derived.

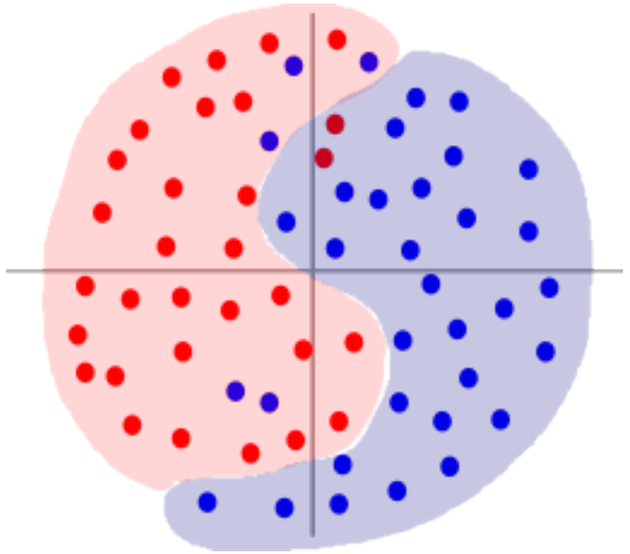


~~PROS AND~~ CONS OF DECISION TREES

- Decision trees are computationally intensive relative to other models, especially if you don't prune them.

- Decision trees are sometimes too flexible and can easily overfit your data. Cross-validation and tuning are key to keeping decision tree models generalizable.





GUIDED PRACTICE

TREE QUIZ

ACTIVITY: TREE QUIZ

ANSWER THE FOLLOWING QUESTIONS (10 minutes):

1. Why would a decision tree be liable to overfit?
2. What is the difference between agglomerative clustering and decision trees?

DELIVERABLE

Answers to the above questions

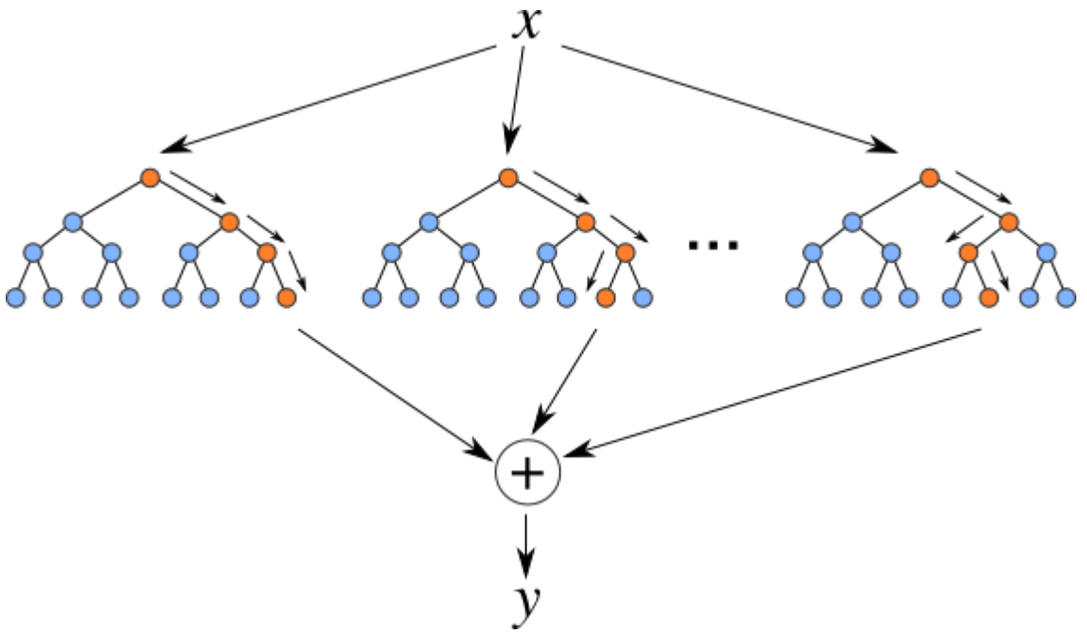
INTRODUCTION

RUNNING THROUGH THE RANDOM FORESTS

RUNNING THROUGH THE RANDOM FORESTS

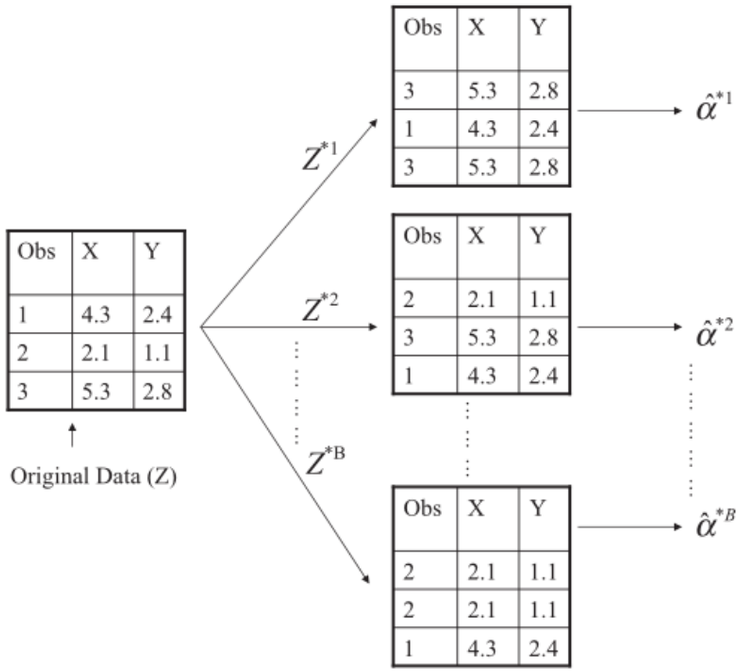
- Random forest models are one of the most widespread classifiers used because they are relatively simple to use and avoid overfitting.

- They do this by **ensembling** or aggregating the results of several individual decision trees.



RUNNING THROUGH THE RANDOM FORESTS

- Random forests generates many decision trees using another resampling method – **bootstrapping**.



- Bootstrapping differs from cross-validation in two major ways - it creates large samples and allows replacement.

RUNNING THROUGH THE RANDOM FORESTS

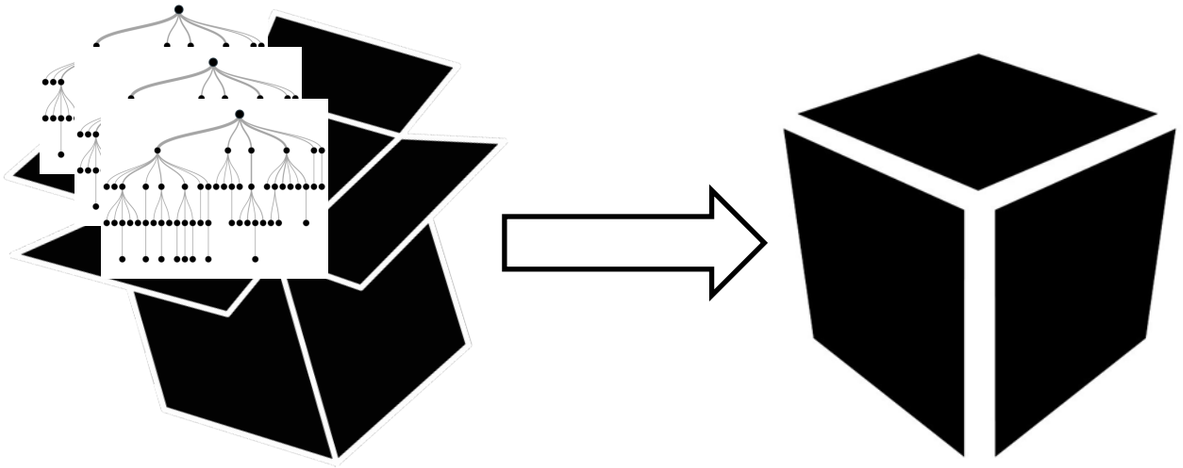
- For every bootstrapped sample, a decision tree is built and then the results are aggregated to form a random forest.

- The idea is that individual trees are likely to overfit, but a set of trees generated from random samples of the original data are unlikely to overfit because each sample will be different.

- *Only the most significant decision rules will be the same across different trees in the same forest.*

RUNNING THROUGH THE RANDOM FORESTS

- This comes with a major tradeoff – random forests are a *black box model* so we lose the interpretability and visualization of decision trees.



- Even though random forests are a black box model, we still have access to all of the same tuning parameters as a regular decision tree plus one more – the number of trees to build before ensembling.

DEMO

RANDOM FORESTS

RANDOM FORESTS

- Open up starter-code-12 and we'll take a look at how to build a random forests model with the sklearn library.

INDEPENDENT PRACTICE

APPLIED FORESTRY

ACTIVITY: APPLIED FORESTRY

EXERCISE

DIRECTIONS (40 minutes):

Use the provided code to fit a random forest model to the California real estate dataset. The goal is to predict the median house value based on the other variables in the dataset.

1. Is this a regression or classification problem?

2. Build a decision tree for this data and then examine the results (e.g. visuals, variable importance).

3. Build a random forest for this data and then examine the results .

DELIVERABLE

Answers to the above questions

CONCLUSION

TOPIC REVIEW

REVIEW Q&A

- What are decision trees?
- What are some common problems with decision trees?
- What are random forests?
- What are some common problems with random forests?

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Unit Project 4!

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

LESSON 12

DECISION TREES AND RANDOM FORESTS

THANKS FOR THE FOLLOWING

CITATIONS

- *Decision Tree Visualization:* https://littleml.files.wordpress.com/2012/01/screen-shot-2012-01-23-at-10-00-17-am1.png

- *90's Flowchart*, Munroe, Randall: https://xkcd.com/210/

- *Questions on some data-mining algorithms*: https://stackoverflow.com/questions/4084668/questions-on-some-data-mining-algorithms

THANKS FOR THE FOLLOWING

CITATIONS

- *An Introduction to Statistical Learning*, James, G et al (2013): http://www-bcf.usc.edu/~gareth/ISL/getbook.html
- *The Lorax (Character)*, Seuss Wikia: http://seuss.wikia.com/wiki/The_Lorax_(Character)
- *Classification and Regression Trees*, Cosma Shalizi: http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf