MILWAUKEE MACHINE LEARNING

Agenda

- 5:30 - Networking
- 6:00 - Presentation
- 6:30 - Discussion and networking

Bathroom

- Take bathroom key from whiteboard tray
- Bathrooms are down one floor at the end of the hallway

Sponsors:

RKKN

RANDOM FORESTS

Milwaukee Machine Learning Meetup

Mitchell Henke / @MitchellHenke

ABOUT ME

- Software/Machine Learning @ RokkinCat
- Specialize in data, databases, APIs
- Self-taught ML

CAPABILITIES

- Regression
  - How much is this house worth based on location, size, … ?
- Classification
  - What kind of animal is this based on age, weight, … ?

ADVANTAGES

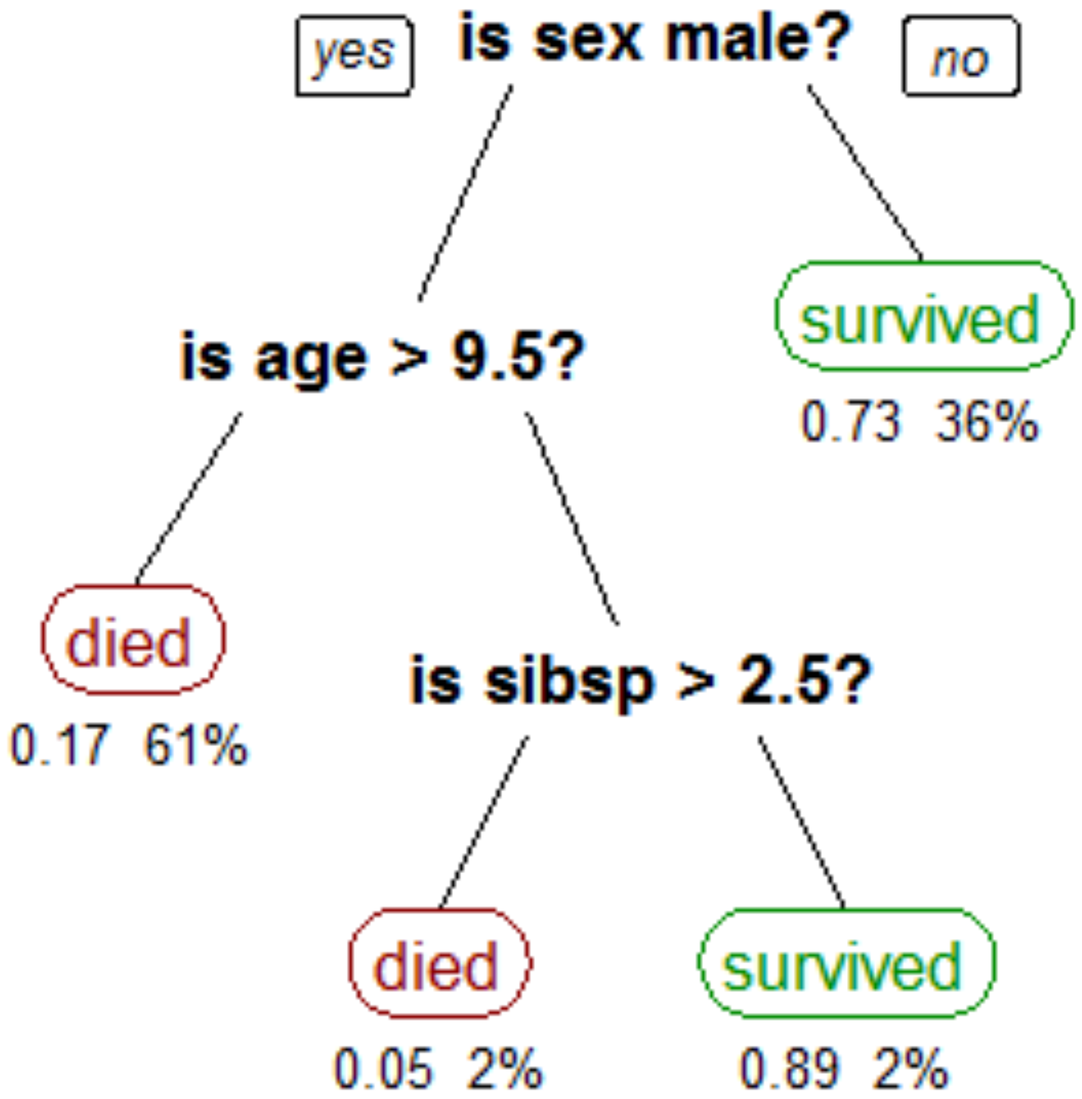
- Relatively simple to build and test
- Minimal hyperparameters
- Variable selection
- Intuitive

RANDOM FORESTS

- Binary Decision Trees
- Ensemble Learning
- Bootstrap Aggregation

BINARY DECISION TREES

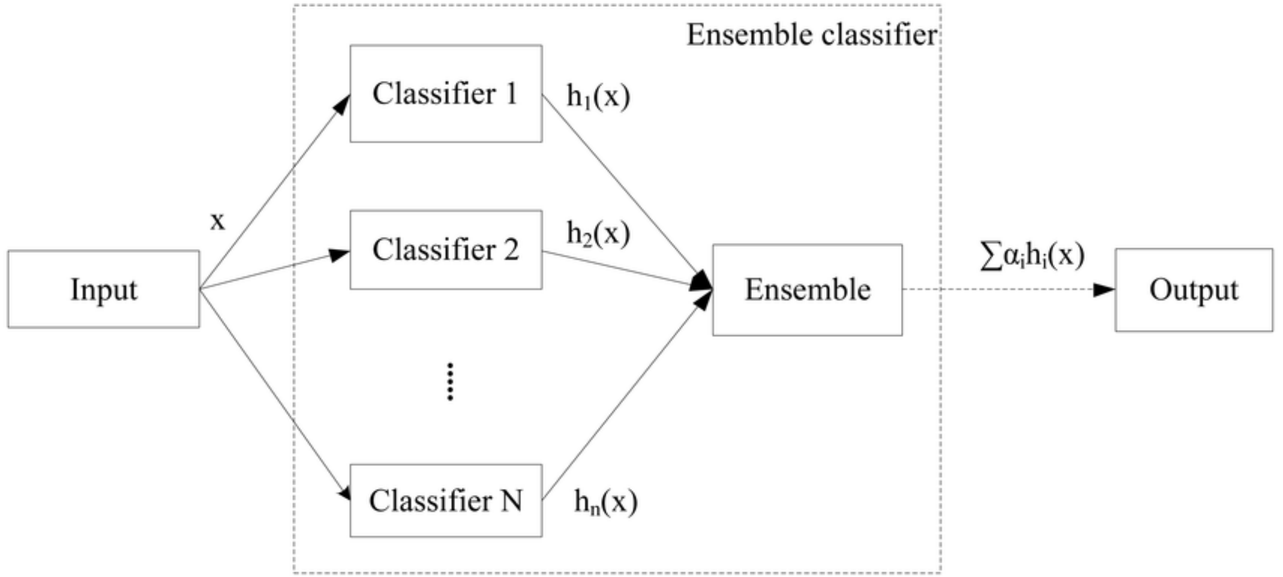
- Given a set of data, find the best split
  - Iterate over every column and value
- Repeat splitting until:
  - Max depth
  - Minimum samples
  - One data point



Retrieved from https://en.wikipedia.org/wiki/Decision_tree_learning#/media/File:CART_tree_titanic_survivors.png

ENSEMBLE LEARNING

- Build many “weak” learners on different subsets of data
- Reduce variance of high-variance algorithms
- In regression, predictions are average of all the trees
- In classification, each tree “votes” and the prediction is the class with the most votes



Retrieved from https://www.researchgate.net/publication/276549421/figure/fig1/AS:339851649011717@1458038355897/Ensemble-learning.png

BOOTSTRAP AGGREGATION (BAGGING)

- Build a random dataset by sampling with replacement
- Random features
- Each tree ends up with a slightly different set of data
- Less concern about overfitting

ROBUST

- Simple Hyperparameters
- Overfitting
- Wide Datasets

SIMPLE HYPERPARAMETERS

- Number of trees in the forest
- Maximum features
- Maximum tree depth
- Minimum samples in leaf node

OVERFITTING

- Beyond a given number, adding trees has little effect on overfitting (and performance)
- Increase minimum samples in leaf nodes
- Reduce maximum tree depth

WIDE DATASETS

- The decision tree algorithm is greedy, and will ignore or make minimal use of unimportant features
- Still valuable to remove unimportant or correlated features

INTERPRETABLE

- Out of Bag Error
- Feature Importance
- Improve Understanding of Data and Model

OUT OF BAG ERROR

- Rough approximation for validation
- Due to bagging, not all data points appear in each tree
- Predict each data point on all of the trees that **don't** contain that data point
  - Calculate resulting error

FEATURE IMPORTANCE

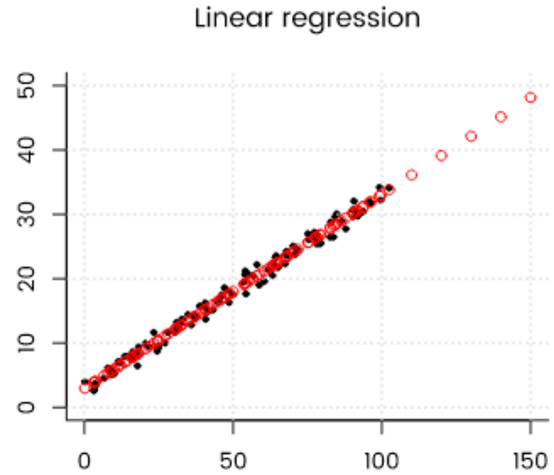
- Shuffle a single column's values and calculate the difference in error
- This is a fantastic technique to guide feature understanding and pruning

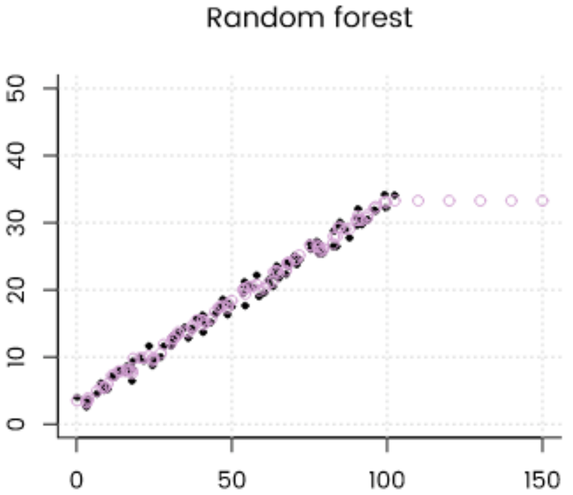
IMPROVE UNDERSTANDING OF DATA AND MODEL

- Visualize decision trees
- Calculate how each attribute in a row changes the final prediction
- Examples!

LIMITATIONS

- More likely to struggle with imbalanced data
- Cannot extrapolate
- Unstructured data
  - Images, text, etc.





Retrieved from http://freerangestats.info/blog/2016/12/10/extrapolation

THANKS

Resources

- https://course.fast.ai/ml.html
- https://scikit-learn.org/stable/modules/ensemble.html#forest
- https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/