PyfA L9

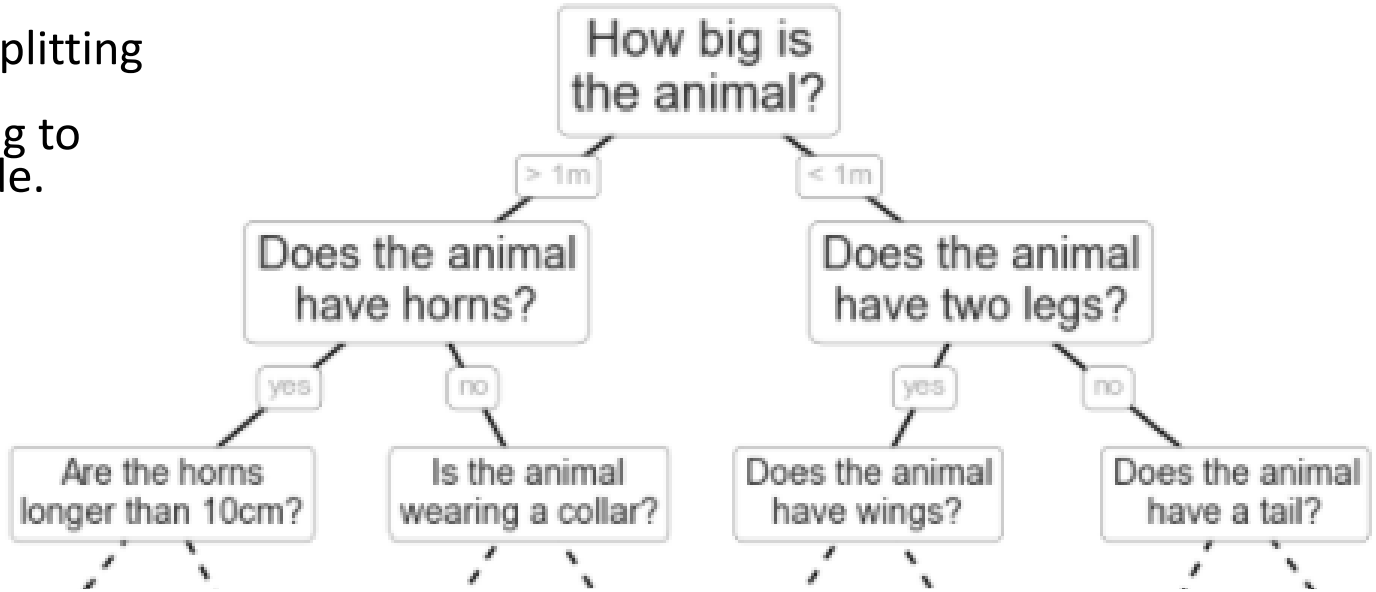
Gizem Aydin, Ph.D.

Recap

- **Regression**: Target variable is **continuous**, and model is used to **predict** its value (e.g. predicting housing prices)
  - Linear Regression (sklearn.linearmodels)
  - Regression Tree / Random Forest Regressor
- **Classification**: Target variable is categorical, **identify the "class"** target variable may belong to (e.g. predicting the Iris flower class)
  - Logistic regression (like Linear Regression but the target is a 0 or 1)
  - Decision Tree Classifier / Random Forest Classifier (sklearn.tree.DecisionTreeClassifier and sklearn.ensemble.RandomForestClassifier)
  - k nearest neighbors (k-NN) model (sklearn.neighbor.KNeighborsClassifier)

Decision Trees

- **Decision Trees (DTs)** are a non-parametric supervised learning method used for **classification** and **regression**. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features -- to find the sequence of questions that has the best accuracy at classifying the data in the fewest steps.
- Decision Tree algorithm selects features such that the uncertainty is reduced the most, this is accomplished by computing the importance of each feature, which lends itself to the right question to ask.
- The decision tree learns by recursively splitting the dataset from the root onwards (in a greedy, node by node manner) according to the splitting metric at each decision node.



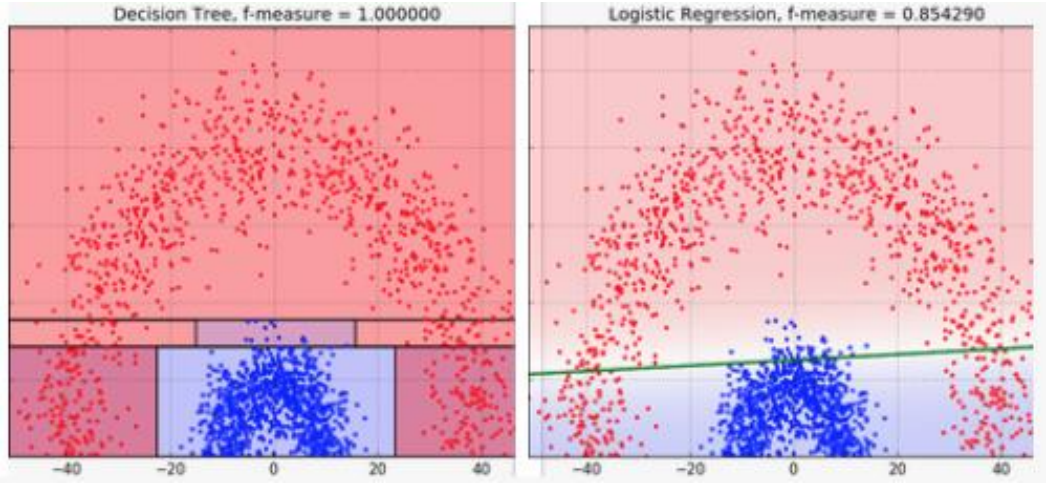
Example decision tree to classify animals

Decision Trees vs. Regression

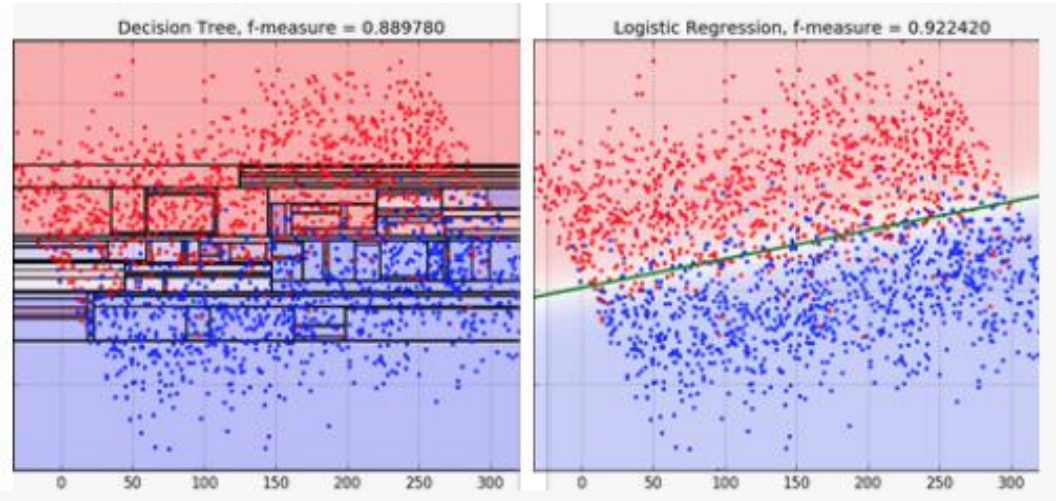
Decision Boundaries and Interpretability

Logistic Regression and **trees** differ in the way that they generate decision boundaries i.e. the lines that are drawn to separate different classes.

Decision Trees bisect the space into smaller and smaller regions, whereas **Logistic Regression** fits a single line to divide the space exactly into two.



Dataset #1

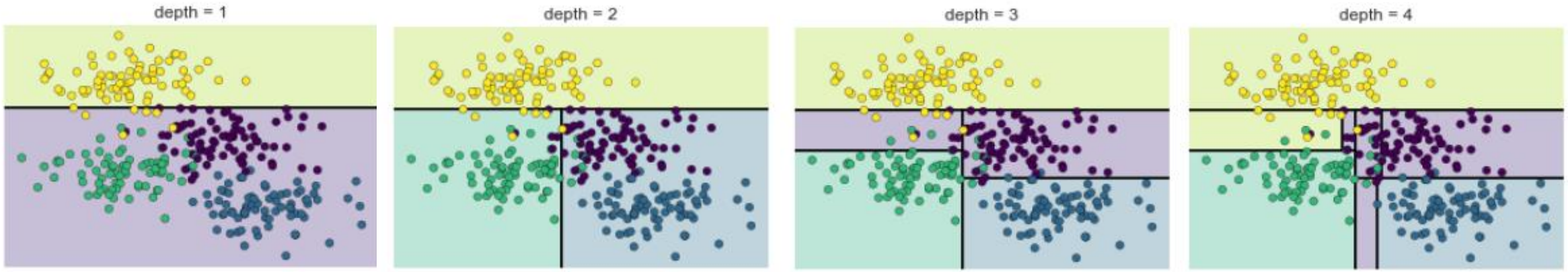


Dataset #2

Source

Advantages and Disadvantages of Decision Tree Models

+ easy to understand and interpret, visualize. numerical and categorical data
- over-fitting, class-domination might occur (use balanced datasets)



How does a decision tree split the data? Once the first feature that reduces the uncertainty the most is selected (this is the depth = 1), there is no need to further divide this branch.

Except for nodes that contain all of one color, at each level every region is again split along the best feature (depth 2, 3, 4 and so on).
As the depth increases, shape of the classification regions get more interesting (not in a good way), this is **overfitting**. We can use pre-pruning to proactively control overfitting, use random forest ensemble models.

Wisdom of the crowd

"it is possible that the many, though not individually good men, yet when they come together may be better, not individually but collectively, than those who are so, just as public dinners to which many contribute are better than those supplied at one man's cost“.

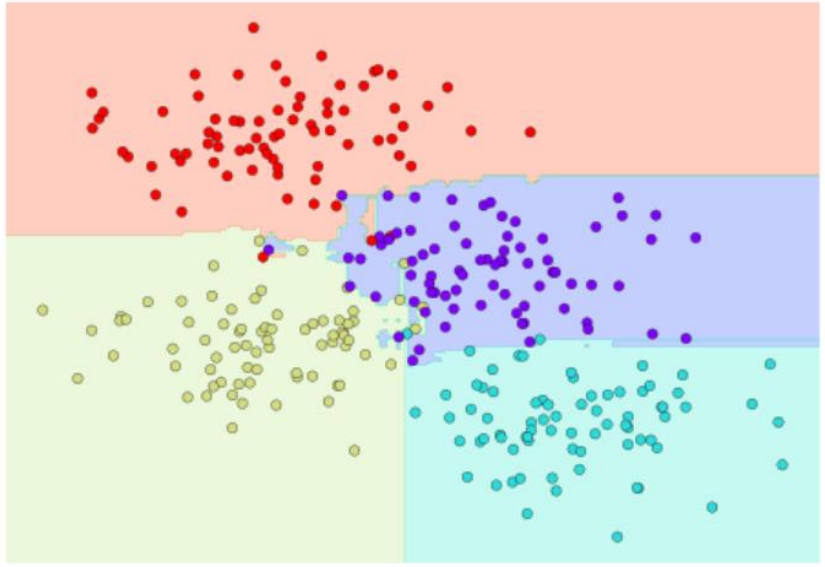
Aristotle, in *Politics*

At a 1906 country fair in Plymouth, 800 people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician Francis Galton observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds. This has contributed to the insight in cognitive science that **a crowd's individual judgments can be modeled as a probability distribution of responses with the median centered near the true value of the quantity to be estimated.**

>> Quora, Reddit, Wikipedia…

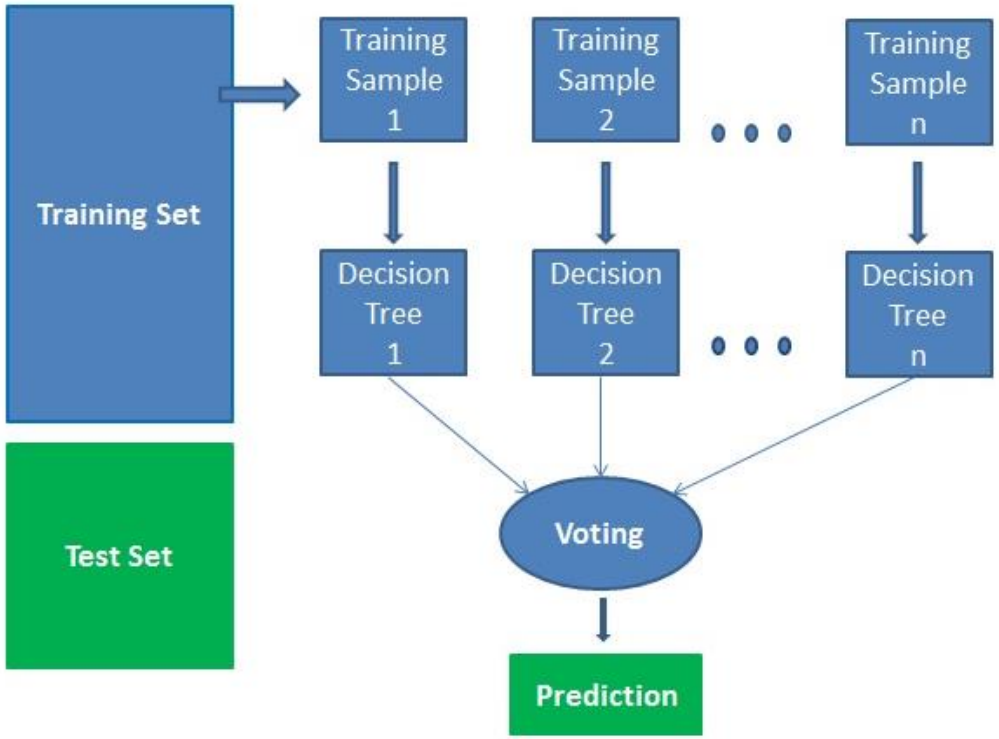
Random Forests

Multiple overfitting estimators can be combined to reduce the effect of overfitting. By using an ensemble of parallel estimators and averaging the results, we can find a better classification model. An ensemble of randomized decision trees is referred as a *random forest*.



Random Forest Algorithm

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the highest average class probability as the final prediction.



Next up in videos:

- Random Forest Classifier – Predicting Iris classification
  - In-class student exercise: Random Forest for Classifying Digits

- Random Forest Regressor - Predicting temperature
  - Data exploration and preparing for modeling
  - Building a random forest regressor model
  - Performance metrics and interpreting the model results
  - Prediction
  - Cross validation (cv) and hyperparameter tuning with RandomizedSearchCV and GridSearchCV