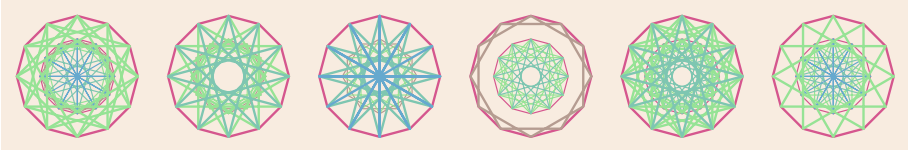
Notes

CSC316 Machine Learning
Professor Leon Tabak

23 February 2022



Notes

- decision trees
  - classification and regression
  - component of Random Forests
  - among the most powerful algorithms
  - DecisionTreeClassifier, DecisionTreeRegressor
  - example with the iris dataset
  - Gini impurity (defined on p. 177) vs. entropy
  - CART (Classification and Regression Tree) training algorithm
  - an example of a white box model
  - (examples of black box models include Random Forests and neural networks)
- Ensemble Learning and Random Forests

  - "weak learner"—little better than random guessing
  - "strong learner"—high accuracy
  - many weak learners working together can match the performance of a strong learner
  - popular ensemble methods include bagging, boosting, and stacking
  - ensemble methods work best when there is diversity, independence
- dimensionality reduction
  - "curse of dimensionality"
  - reducing dimensions speeds up our algorithms
  - reducing dimensions makes it easier to visualize our data
  - warning! reducing dimensions also means losing info
  - projection
  - examine some hard cases
  - PCA—Principal Component Analysis
- unsupervised learning techniques
  - data is not labeled
  - example: clustering—dividing data into sets, in which elements of a set have some common characteristics
  - K-Means algorithm
    * decide how many clusters we think there are
    * guess where the centers of the clusters are
    * assign each item in dataset to cluster with nearest center
    * compute centroid of each cluster (sum all vectors and divide by n)
    * centroids become new centers of clusters
    * reassign items to clusters—again put each item the cluster whose center is closest
  - other applications of unsupervised learning—anomaly detection, density estimation