THE UNIVERSITY OF TENNESSEE KNOXVILLE

AICIP RESEARCH

COSC 522 – Machine Learning

Lecture 15 – Decision Tree and Random Forest

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu

Course Website: http://web.eecs.utk.edu/~hqi/cosc522/

Roadmap

- Supervised learning
  - Classification
    - Maximum Posterior Probability (MPP): For a given x, if $P(w_1|x) > P(w_2|x)$, then x belongs to class 1, otherwise 2.
      - Parametric Learning
        - Three cases
        - Estimate Gaussian parameters using MLE
      - Nonparametric Learning
        - Parzen window (fixed window size)
        - K-Nearest Neighbor (variable window size)
    - Neural Network
    - SVM
    - Decision Tree
  - Regression (linear regression with nonlinear basis functions)
    - Neural Network
    - SVM
    - Decision Tree
- Unsupervised learning
  - Non-probabilistic approaches
    - kmeans, wta
  - Hierarchical approaches
    - Agglomerative clustering
  - Neural Network
- Supporting preprocessing techniques
  - Dimensionality Reduction
    - Supervised linear (FLD)
    - Unsupervised linear (PCA)
    - Unsupervised nonlinear (t-SNE)
- Supporting postprocessing techniques
  - Classifier Fusion
    - NB
    - BKS
    - XGBoost
  - Performance Evaluation
- Optimization techniques
  - Gradient Descent (GD)

Questions

- What is nominal data?
- What are root, descendent, and leaf node? What is link?
- How many splits from a node?
- What is impurity?
- What is entropy impurity? Can you understand why the more equally distributed the probabilities, the higher the impurity?
- What is Gini impurity?
- What is the cost function for choosing the right query for a decision tree?
- What is the optimization method?
- How to determine leaf nodes?
- What is MDL?
- What is the rationale behind pruning?
- What is Bagging (intuitively)?
- What is Random Forest (intuitively)?

Nominal Data

- Descriptions that are discrete and without any natural notion of similarity or even ordering

Some Terminologies

- Decision tree
  - Root
  - Link (branch) - directional
  - Leaf
  - Descendent node

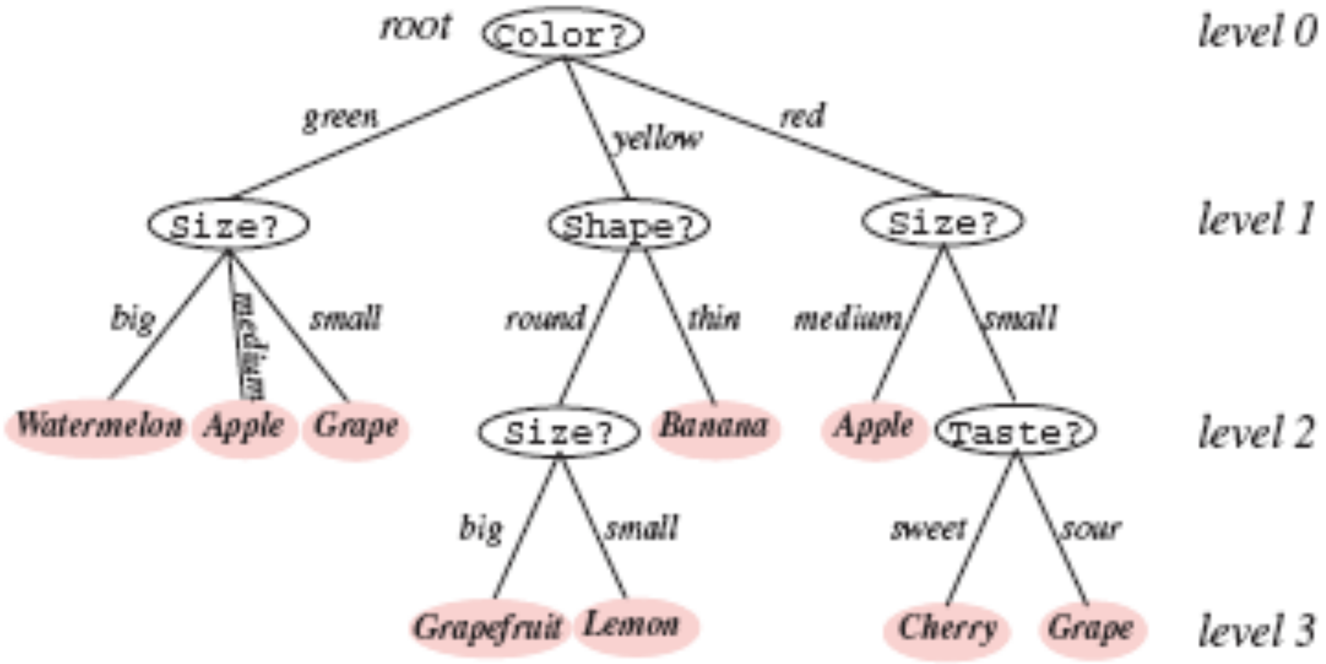


FIGURE 8.1. Classification in a basic decision tree proceeds from top to bottom. The questions asked at each node concern a particular property of the pattern, and the downward links correspond to the possible values. Successive nodes are visited until a terminal or leaf node is reached, where the category label is read. Note that the same question, Size?, appears in different places in the tree and that different questions can have different numbers of branches. Moreover, different leaf nodes, shown in pink, can be labeled by the same category (e.g., Apple). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

CART

- Classification and regression trees
- A generic tree growing methodology
- Issues studied
  - How many splits from a node?
  - Which property to test at each node?
  - When to declare a leaf?
  - How to prune a large, redundant tree?

Number of Splits

- Binary tree
- Expressive power and comparative simplicity in training

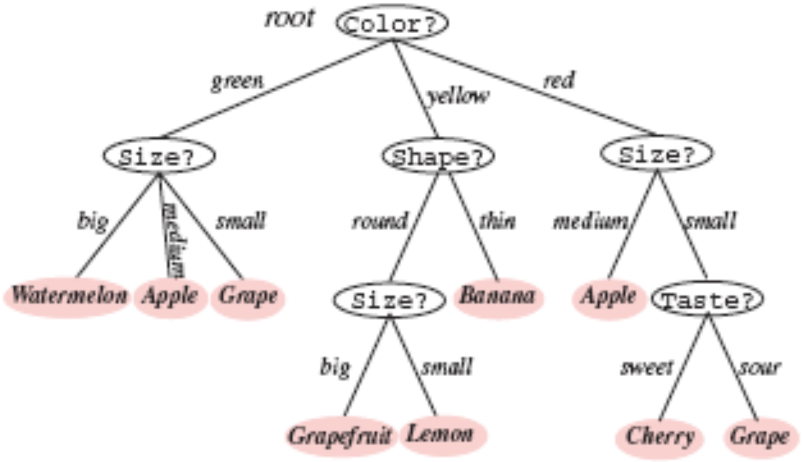


FIGURE 8.1. Classification in a basic decision tree proceeds from top to bottom. The qu
each node concern a particular property of the pattern, and the downward links correspon
values. Successive nodes are visited until a terminal or leaf node is reached, where the categ
Note that the same question, Size?, appears in different places in the tree and that differe
have different numbers of branches. Moreover, different leaf nodes, shown in pink, can b
same category (e.g., Apple). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Patte*
Copyright © 2001 by John Wiley & Sons, Inc.

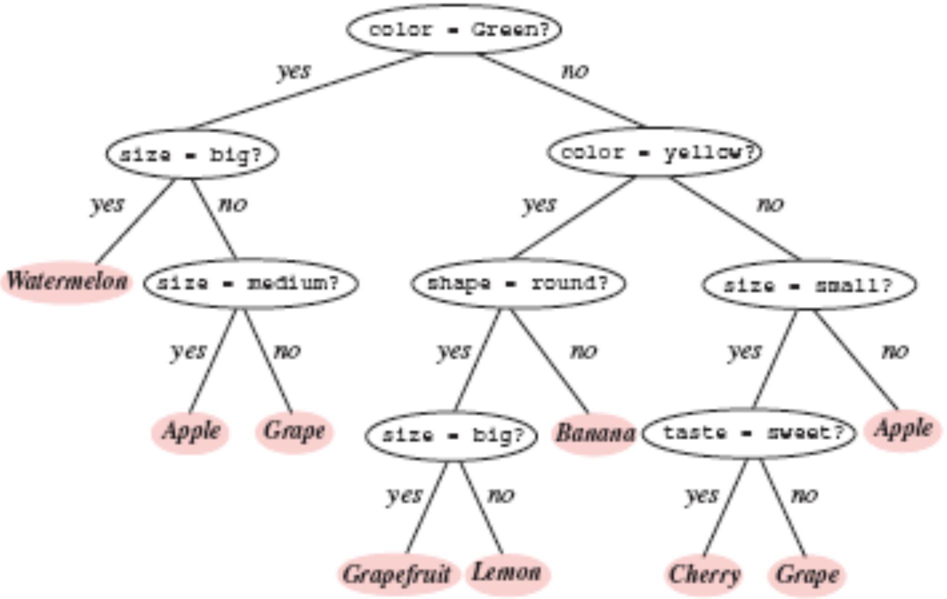


FIGURE 8.2. A tree with arbitrary branching factor at different nodes can always be represented by a functionally equivalent binary tree—that is, one having branching factor $B = 2$ throughout, as shown here. By convention the "yes" branch is on the left, the "no" branch on the right. This binary tree contains the same information and implements the same classification as that in Fig. 8.1. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Node Impurity – Occam' s Razor

- The fundamental principle underlying tree creation is that of simplicity: we prefer simple, compact tree with few nodes
- http://math.ucr.edu/home/baez/physics/General/occam.html
- Occam's (or Ockham's) razor is a principle attributed to the 14th century logician and Franciscan friar; William of Occam. Ockham was the village in the English county of Surrey where he was born.
- The principle states that "Entities should not be multiplied unnecessarily."
- "when you have two competing theories which make exactly the same predictions, the one that is simpler is the better."
- Stephen Hawking explains in A Brief History of Time: "We could still imagine that there is a set of laws that determines events completely for some supernatural being, who could observe the present state of the universe without disturbing it. However, such models of the universe are not of much interest to us mortals. It seems better to employ the principle known as Occam's razor and cut out all the features of the theory which cannot be observed."
- Everything should be made as simple as possible, but not simpler

CART

- Classification and regression trees
- A generic tree growing methodology
- Issues studied
  - How many splits from a node?
  - Which property to test at each node?
  - When to declare a leaf?
  - How to prune a large, redundant tree?

Questions

- What is nominal data?
- What are root, descendent, and leaf node? What is link?
- How many splits from a node?
- What is impurity?
- What is entropy impurity? Can you understand why the more equally distributed the probabilities, the higher the impurity?
- What is Gini impurity?
- What is the cost function for choosing the right query for a decision tree?
- What is the optimization method?
- How to determine leaf nodes?
- What is MDL?
- What is the rationale behind pruning?
- What is Bagging (intuitively)?
- What is Random Forest (intuitively)?

Property Query and Impurity Measurement

- We seek a property query T at each node N that makes the data reach the immediate descendent nodes as pure as possible
- We want *i(N)* to be 0 if all the patterns reach the node bear the same category label
- Entropy impurity (information impurity)

$$i(N) = -\sum_{j} P(\omega_j) \log_2 P(\omega_j)$$

$P(\omega_j)$ is the fraction of patterns at node N that are in category ω_j

Other Impurity Measurements

- Variance impurity (2-category case)

$$i(N) = P(\omega_1)P(\omega_2)$$

- Gini impurity

$$i(N) = \sum_{i \neq j} P(\omega_i)P(\omega_j) = 1 - \sum_{j} P^2(\omega_j)$$

- Misclassification impurity

$$i(N) = 1 - \max_{j} P(\omega_j)$$

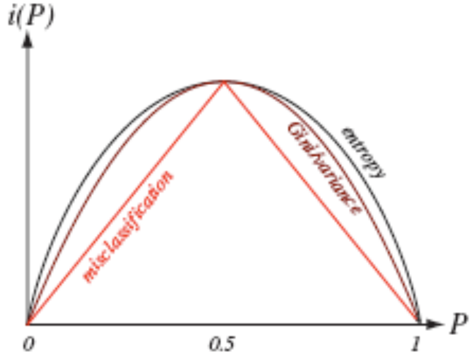


FIGURE 8.4. For the two-category case, the impurity functions peak at equal class frequencies and the variance and the Gini impurity functions are identical. The entropy, variance, Gini, and misclassification impurities (given by Eqs. 1–4, respectively) have been adjusted in scale and offset to facilitate comparison here; such scale and offset do not directly affect learning or classification. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Choose the Property Test?

- Choose the query that decreases the impurity as much as possible

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R)$$

  - N_L, N_R: left and right descendent nodes
  - $i(N_L)$, $i(N_R)$: impurities
  - P_L: fraction of patterns at node N that will go to N_L

- Solve for extrema (local extrema)

Example

- Node N:
  - 90 patterns in ω_1
  - 10 patterns in ω_2
- Split candidate:
  - 70 ω_1 patterns & 0 ω_2 patterns to the right
  - 20 ω_1 patterns & 10 ω_2 patterns to the left

CART

- Classification and regression trees
- A generic tree growing methodology
- Issues studied
  - How many splits from a node?
  - Which property to test at each node?
  - When to declare a leaf?
  - How to prune a large, redundant tree?
  - If the leaf is impure, how to classify?
  - How to handle missing data?

Questions

- What is nominal data?
- What are root, descendent, and leaf node? What is link?
- How many splits from a node?
- What is impurity?
- What is entropy impurity? Can you understand why the more equally distributed the probabilities, the higher the impurity?
- What is Gini impurity?
- What is the cost function for choosing the right query for a decision tree?
- What is the optimization method?
- How to determine leaf nodes?
- What is MDL?
- What is the rationale behind pruning?
- What is Bagging (intuitively)?
- What is Random Forest (intuitively)?

When to Stop Splitting?

- Two extreme scenarios
  - Overfitting (each leaf is one sample)
  - High error rate
- Approaches
  - Validation and cross-validation
    - 90% of the data set as training data
    - 10% of the data set as validation data
  - Use threshold
    - Unbalanced tree
    - Hard to choose threshold
  - Minimum description length (MDL)
    - i(N) measures the uncertainty of the training data
    - Size of the tree measures the complexity of the classifier itself

$$MDL = \alpha \bullet size + \sum_{leaf\ \ nodes} i(N)$$

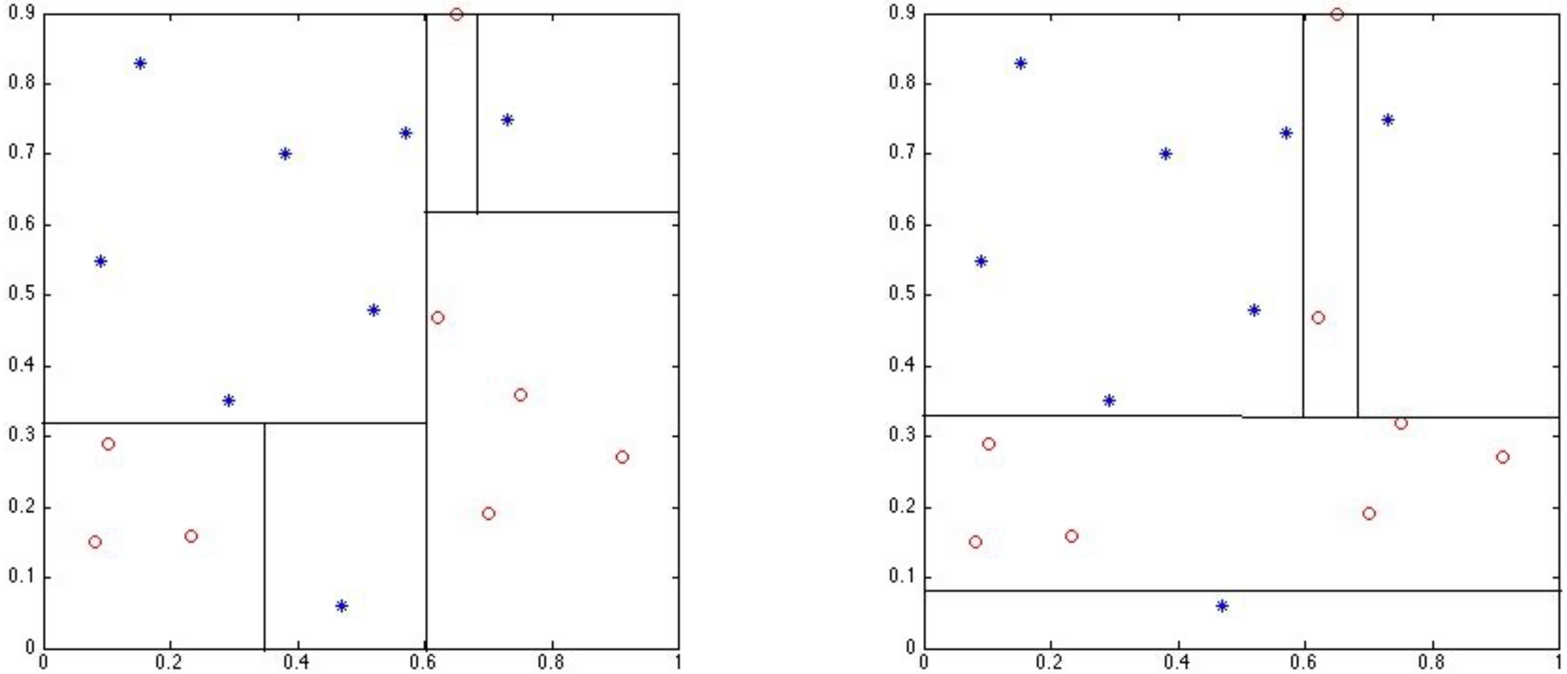
CART

- Classification and regression trees
- A generic tree growing methodology
- Issues studied
  - How many splits from a node?
  - Which property to test at each node?
  - When to declare a leaf?
  - How to prune a large, redundant tree?

Pruning

- Another way to stop splitting
- Horizon effect
  - Lack of sufficient look ahead
- Let the tree fully grow, i.e. beyond any putative horizon, then all pairs of neighboring leaf nodes are considered for elimination

Instability



Questions

- What is nominal data?
- What are root, descendent, and leaf node? What is link?
- How many splits from a node?
- What is impurity?
- What is entropy impurity? Can you understand why the more equally distributed the probabilities, the higher the impurity?
- What is Gini impurity?
- What is the cost function for choosing the right query for a decision tree?
- What is the optimization method?
- How to determine leaf nodes?
- What is MDL?
- What is the rationale behind pruning?
- What is Bagging (intuitively)?
- What is Random Forest (intuitively)?

Random Forest

- Potential issue with decision trees
- Prof. Leo Breiman
- Ensemble learning methods
  - Bagging (**B**ootstrap **agg**regati**ng**): Proposed by Breiman in 1994 to improve the classification by combining classifications of randomly generated training sets
  - Random forest: bagging + random selection of features at each node to determine a split

Reference

- [CART] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
- [Bagging] L. Breiman, “Bagging predictors,” *Machine Learning*, 24(2):123-140, August 1996. (citation: 16,393)
- [RF] L. Breiman, “Random forests,” *Machine Learning*, 45(1):5-32, October 2001.