Machine Learning

Decision Trees

IIIT Sri City

Nominal Data

- So far we consider patterns to be represented by feature vectors of real or integer values.
- Easy to come up with a distance (similarity) measure by using a variety of mathematical norms.
- What happens if features are not numbers?
- May not have a numerical representation
- Distance measures might not make sense

Credits: Machine Learning, UCSB

Nominal Data: Examples

- Consider the use of information about teeth in the classification of fish and sea mammals.

E.g.:

  - Some teeth are small and fine (as in baleen whales) for straining tiny prey from the sea. Others (as in sharks) coming in multiple rows.

- Consider describing a piece of fruit by the four properties of color, texture, taste and smell.
  - color = red, texture = shiny,
  - taste = sweet and size = small
- Another common approach is to describe the pattern by a variable length string of nominal attributes, such as sequence of base pairs string in a segment of DNA,
  - E.g.: “AGCTTCAGATTCCA.”

Credits: Pattern Classification, Duda Hart

How to use this data for classification/regression?

- How can we best use such nominal data for classification?
- Most importantly, how can we efficiently learn categories using such non-metric data?
- If there is structure in strings, how can it be represented?

Credits: Machine Learning, UCSB

How to use this data for classification/regression?

- Visualizing using n-dimensional space might be difficult how to map, say, smell, onto an axis?

- There might only be few discrete values (an article is highly interesting, somewhat interesting, not interesting, etc.)

- Even though that helps, do remember you cannot take distance measure in that space

(e.g., Euclidean distance in r-g-b color space does not correspond to human perception of color)

Credits: Machine Learning, UCSB

Decision Trees

- A classification based on a sequence of questions on
  - A particular feature (E.g., is the fruit sweet or not?) or
  - A particular set of features (E.g., is this article relevant and interesting?)
- Answer can be either
  - Yes/no
  - Choice (relevant & interesting, interesting but not relevant, relevant but not interesting, etc.)
  - Usual a finite number of discrete values

Credits: Machine Learning, UCSB

Decision Trees

- It is natural and intuitive to classify a pattern through a sequence of questions, in which the next question asked depends on the answer to the current question.

- This approach is particularly useful for non-metric data, since all of the questions can be asked in a “yes/no” or “true/false” or “value(property) ∈ set of values” style that does not require any notion of metric.

- ***Such sequence of questions is displayed in a directed decision tree or simply tree.***

Credits: Pattern Classification, Duda hart

Decision Trees

- ***Such sequence of questions is displayed in a directed decision tree or simply tree.***

- Where by convention **root node** the first or root node is displayed at the top.

- Root is connected by **successive (directional) links or branches** to other nodes.

- These are similarly connected until we reach **terminal or leaf nodes**, which have no further links.

Credits: Pattern Classification, Duda hart

Decision Tree

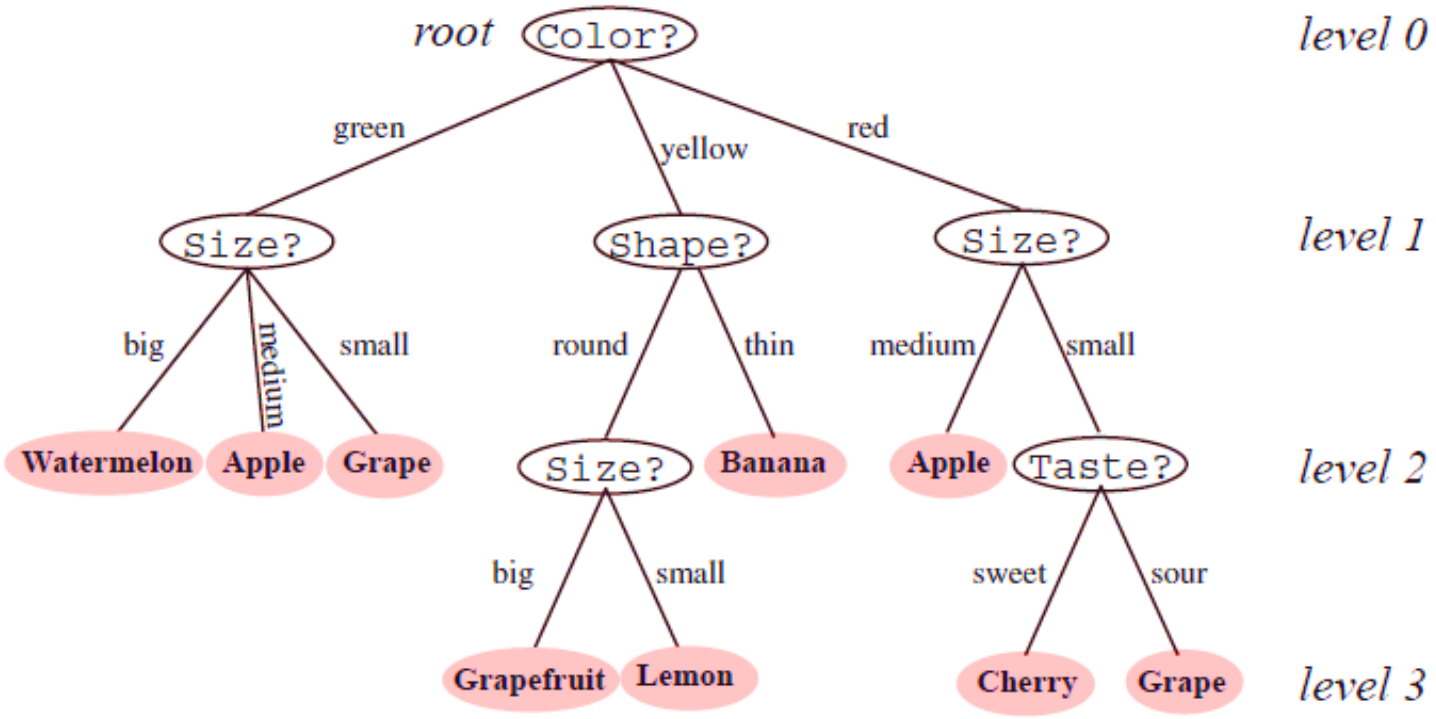


Figure 8.1: Classification in a basic decision tree proceeds from top to bottom. The questions asked at each node concern a particular property of the pattern, and the downward links correspond to the possible values. Successive nodes are visited until a terminal or leaf node is reached, where the category 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**).

Credits: Pattern Classification, Duda hart

DT: How they are used for classification?

- The classification of a particular pattern begins at the root node, which asks for the value of a particular property of the pattern.
- The different links from the root node correspond to the different possible values.
- Based on the answer we follow the appropriate link to a subsequent or descendent node.
- In the trees, the links must be mutually distinct and exhaustive, i.e., one and only one link will be followed.

DT: How they are used for classification?

- The next step is to make the decision at the sub-tree appropriate subsequent node, which can be considered the root of a sub-tree.

- We continue this way until we reach a leaf node, which has no further question.

- Each leaf node bears a category label and the test pattern is assigned the category of the leaf node reached.

Creation of a Decision Tree

- Use supervised learning
  - Samples with tagged label (just like before)
- Process
  - Number of splits
  - Query selection
  - Rule for stopping splitting and pruning
  - Rule for labelling the leaves
  - Variable combination and missing data

Credits: Machine Learning, UCSB

Number of splits

- Each decision outcome at a node is called a *split*, since it corresponds to splitting a subset of the training data.

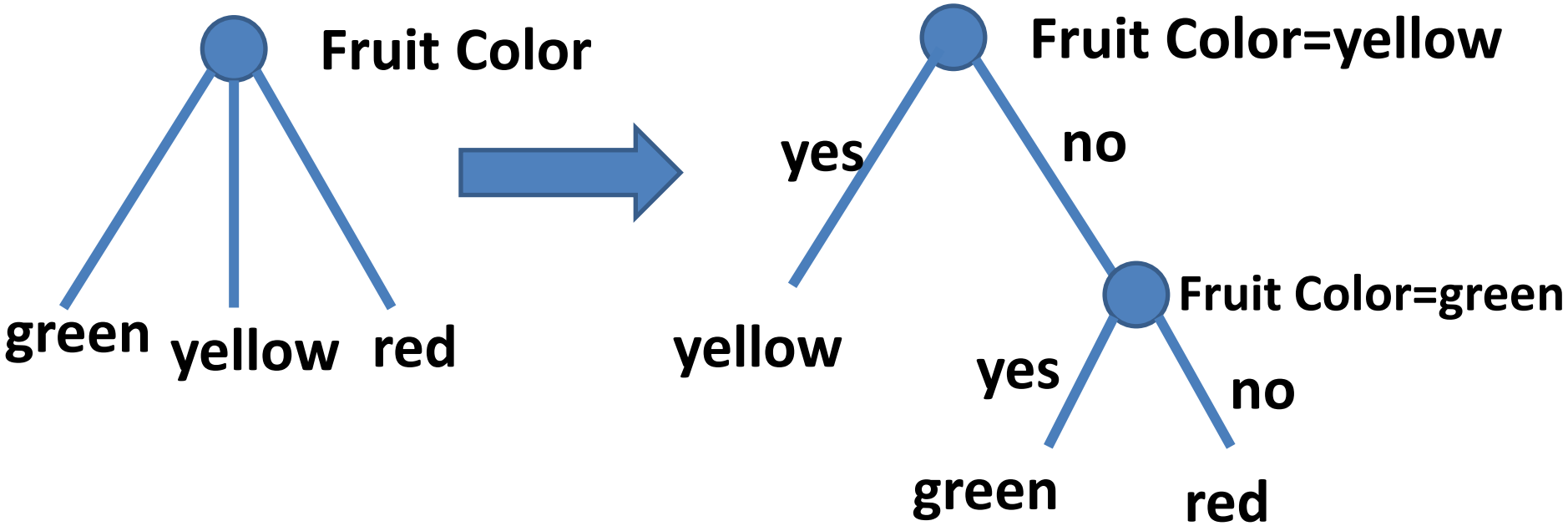
- The root node splits the full training set; each successive decision splits a proper subset of the data.

- The number of splits at a node is closely related to which property need to be tested and specifying *which* particular split will be made at a node.

- The number of links descending from a node is sometimes called branching the node's ***branching factor*** **or** ***branching ratio***, denoted *B*.

Binary vs Multi-way Splits

- Binary vs. Multi-way
  - Can always make a multi-way split into binary splits



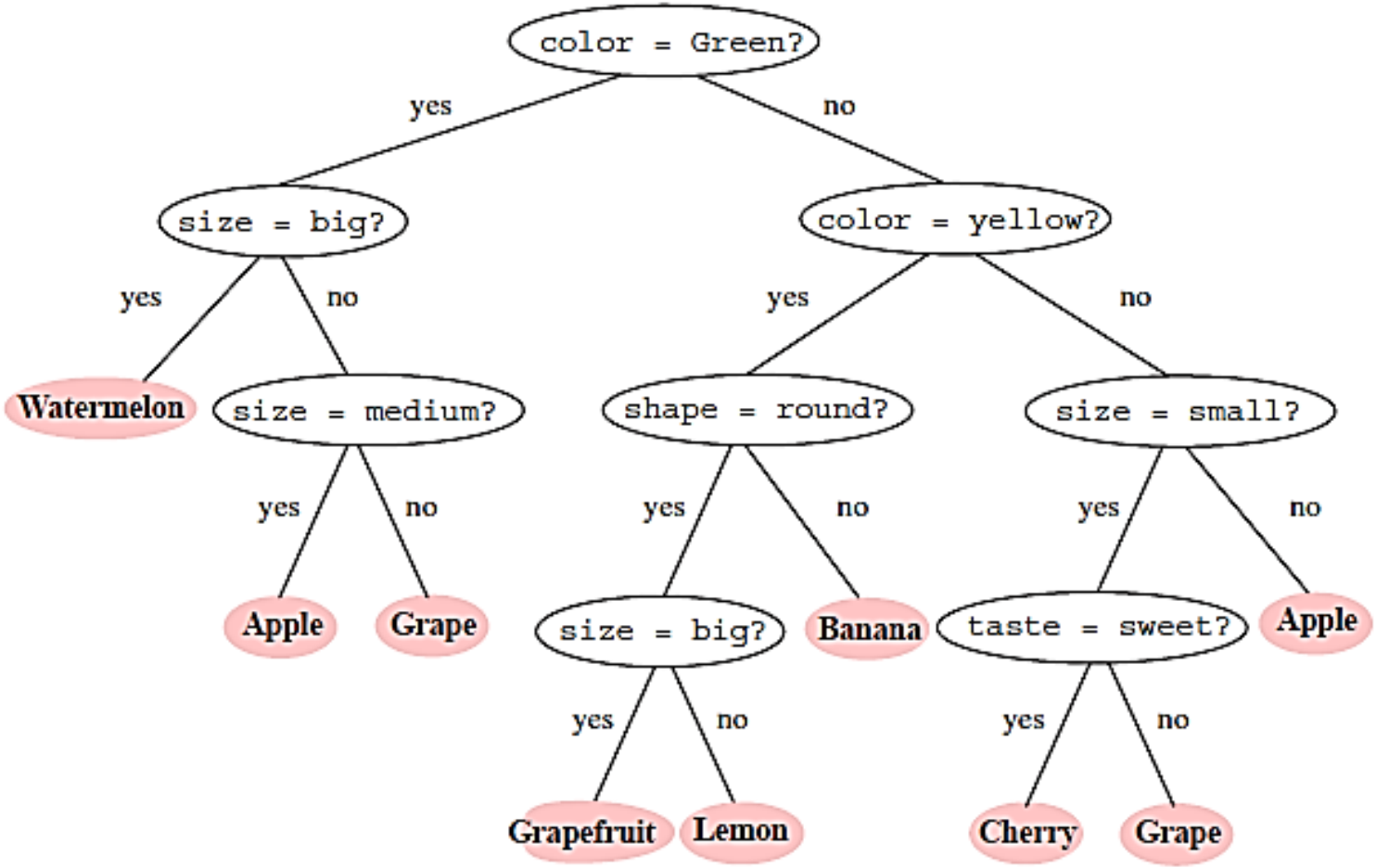


Figure 8.2: A tree with arbitrary branching factor at different nodes can always be represented by a functionally equivalent binary tree, i.e., one having branching factor $B = 2$ throughout. By convention the "yes" branch is on the left, the "no" branch on the right. This binary tree contains the same information and implements the same classification as that in Fig. 8.1.

Credits: Pattern Classification, Duda hart

Test selection

- If a feature is an ordered variable,
  - we might ask is x>c, for some c.

- If a feature is a category, we might ask is x in a particular category
  - *Yes* sends samples to left and *no* sends samples to right

- Simple rectangular partitions of the feature space
  - More complicated ones: is x>0.5 & y<0.3

Credits: Machine Learning, UCSB

Test selection

- The fundamental principle underlying tree creation is that of simplicity: we prefer decisions that lead to a simple, compact tree with few nodes.

- *This is a version of Occam's razor, that the simplest model that explains data is the one to be preferred.*

  "Occam's razor is the problem-solving principle that "entities should not be multiplied without necessity", or more simply, the simplest explanation is usually the right one."

- *To this end, we seek a property test T at each node N that makes the purity data reaching the immediate descendent nodes as "**pure**" as possible.*

Decision boundary

- For example, suppose that the test at each node has the form "is $x_i \leq x_i s$?" This leads to hyperplane decision boundaries that are perpendicular to the coordinate axes, and to decision regions of the form illustrated in Fig.

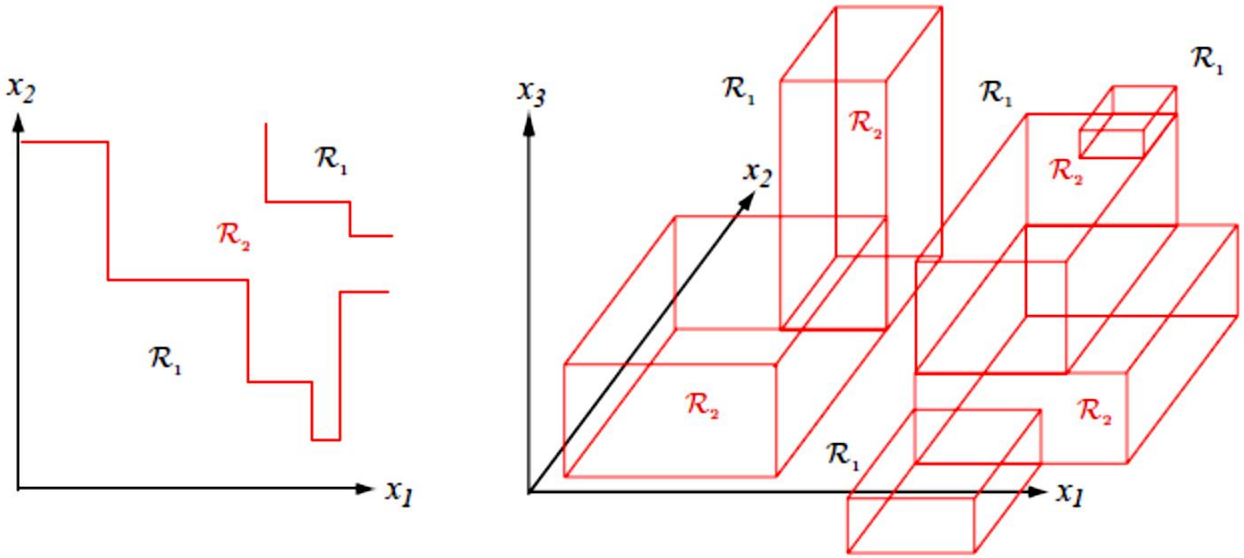


Figure 8.3: Monothetic decision trees create decision boundaries with portions perpendicular to the feature axes. The decision regions are marked $\mathcal{R}_1$ and $\mathcal{R}_2$ in these two-dimensional and three-dimensional two-category examples. With a sufficiently large tree, any decision boundary can be approximated arbitrarily well.

Credits: Pattern Classification, Duda hart

Criteria for Splitting

- Intuitively, to make the populations of the samples in the two children nodes purer than the parent node
- What do you mean by pure?
- In formalizing this notion, it turns out to be more convenient to define the *im*purity, rather than purity.
- General formulation
  - At node n, with k classes
  - Impurity depends on probabilities of samples at that node being in a certain class

$$P(w_i \mid n) \quad i = 1, \cdots, k$$
$$i(n) = f(P(w_1 \mid n), P(w_2 \mid n), \cdots, P(w_k \mid n))$$

Credits: Machine Learning, UCSB

Impurity

- Several different mathematical measures of impurity have been proposed, all of which have basically the same behaviour.
  - Entropy Impurity
  - Variance Impurity
  - Gini Impurity
  - Misclassification impurity