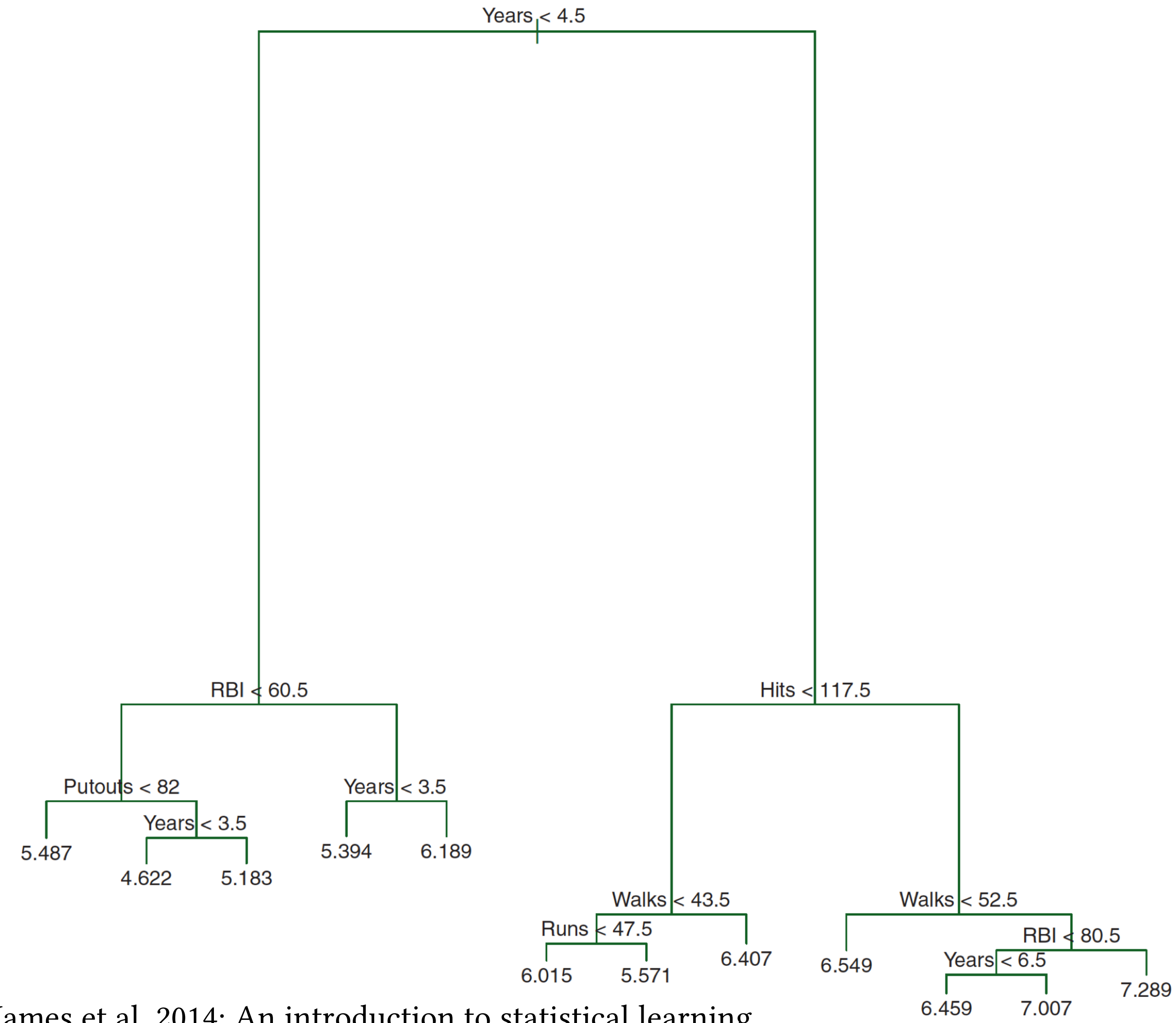
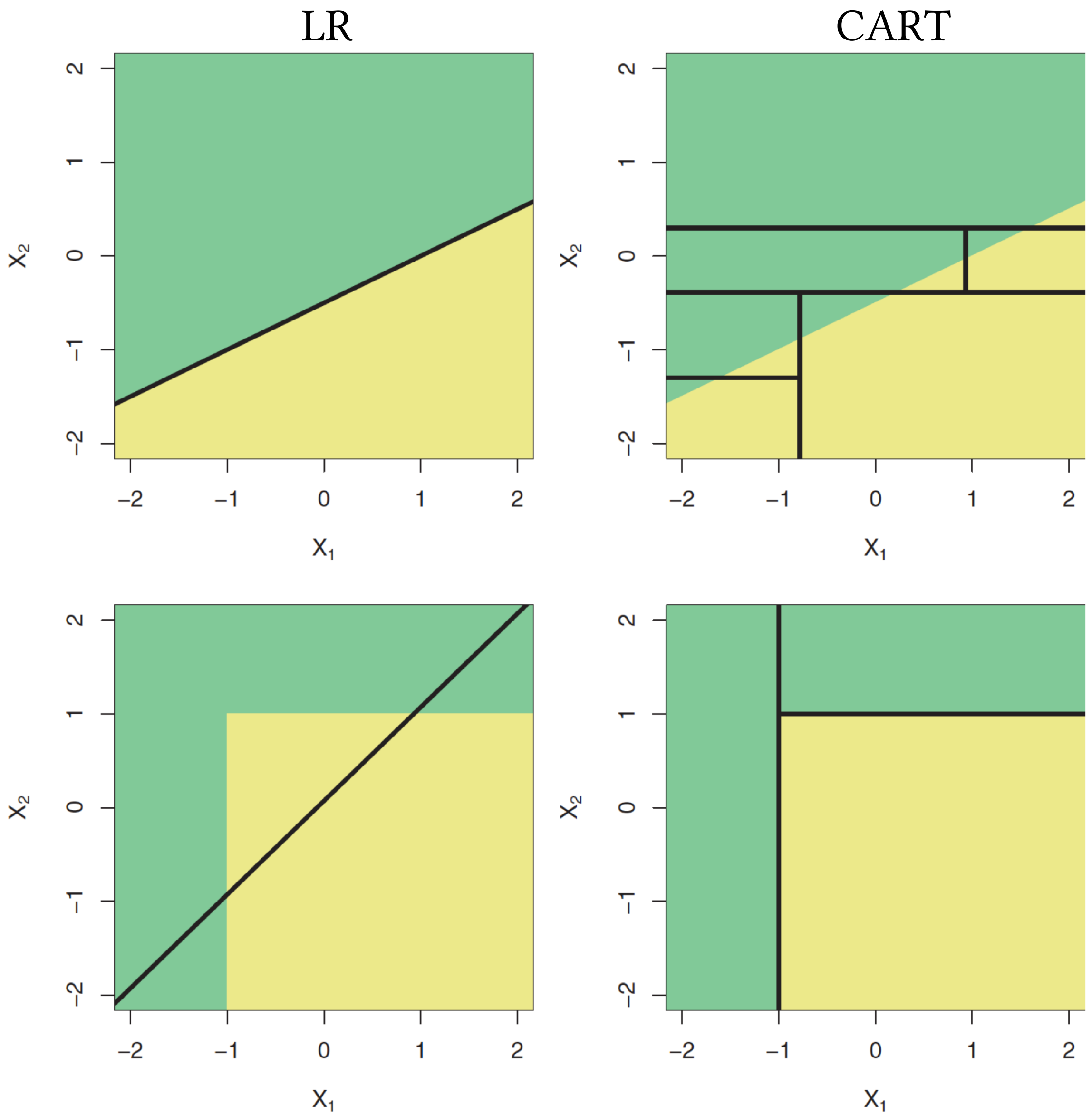
Tree-based methods

classification and regression trees

Motivation





James et al. 2014: An introduction to statistical learning

Training data

Type of fruit	**Diameter**	**Colour**	**...**
Apple	3	Green	...
Apple	3	Yellow	...
Grape	1	Red	...
Grape	1	Red	...
Lemon	3	Yellow	...
...	...	...	...

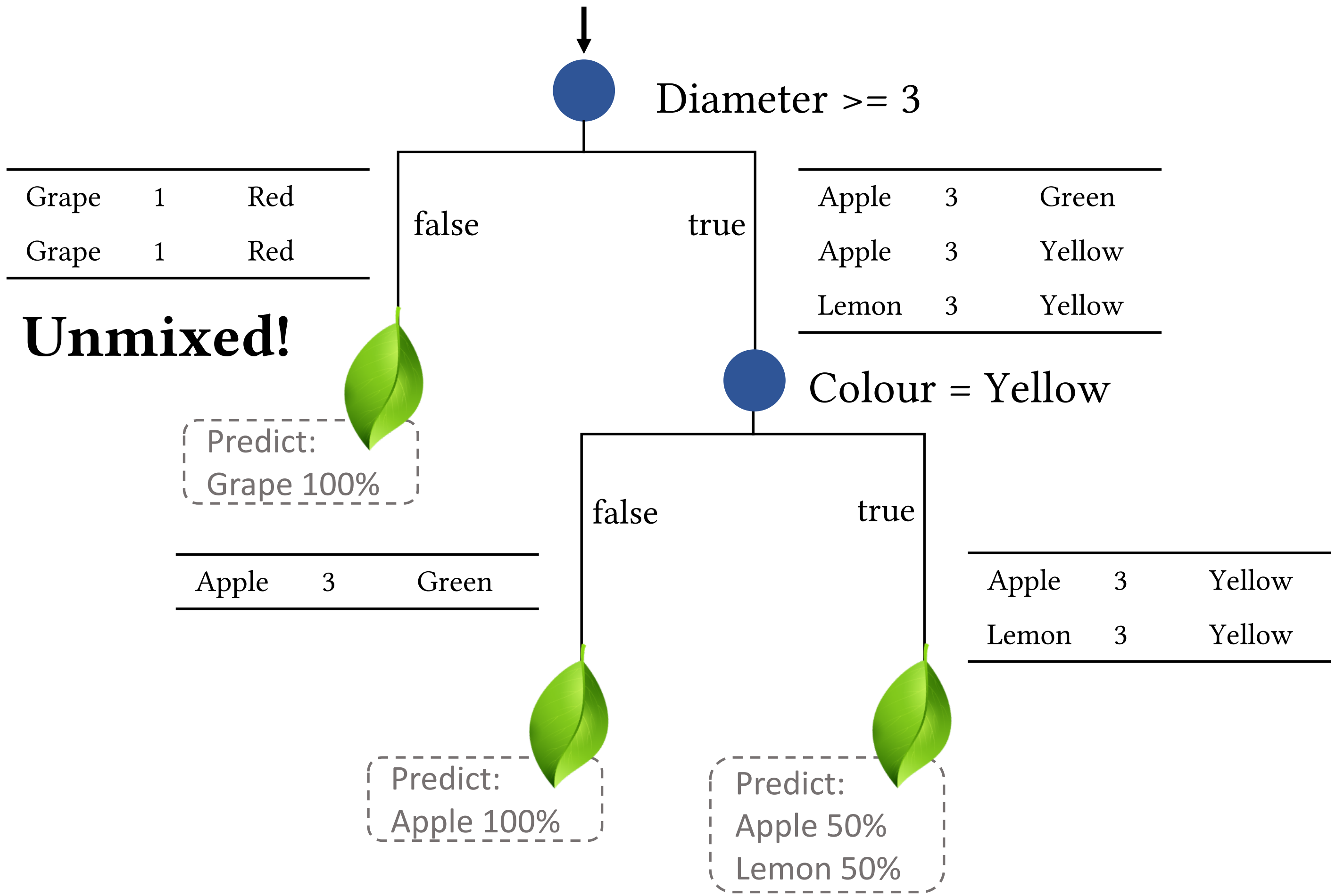
Not perfectly separable!

Decision tree learning

- Family of algorithms
  - ID3, C4.5, C5.0, **CART**
- Common concept: When to ask which questions?

Classification and Regression Tree (CART)

Fruit	Diam.	Colour
Apple	3	Green
Apple	3	Yellow
…	…	…



Root node

Terminal node/ leaf

Internal node

Recursive binary splitting

and when to ask which question

- Computationally costly/infeasible to consider all possible splits
- Top-down and greedy
  - Top-down: starting with root node, all observations belong to one region
  - Greedy: at current node we make the best split, we don't look ahead
- Which questions to ask?

Fruit	Diam.	Colour
Apple	3	Green
Apple	3	Yellow
…	…	…

Possible questions

diameter >= 3; colour = Green

diameter >= 3; colour = Yellow

…

Recursive binary splitting

and when to ask which question

- Regression trees (continuous response)
  - Prediction: mean of each region
- select predictor X_j and splitting criteria/cutpoint s
- For any j and s define half-planes:

  $R_1(j, s) = \{X \mid X_j < s\}$ and $R_2(j, s) = \{X \mid X_j >= s\}$
- Minimize RSS

$$\sum_{i:\, xi \in R_1(j,s)} (yi - \hat{y}_{R1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (yi - \hat{y}_{R2})^2$$

James et al. 2014: An introduction to statistical learning
Hastie et al. 2009: Elements of statistical learning

Recursive binary splitting

and when to ask which question

- Classification trees (qualitative, categorical response)
  - Prediction: mode of each region
- select predictor X_j and splitting criteria/cutpoint s
- For any j and s define half-planes:

$R_1(j, s) = \{X \mid X_j < s\}$ and $R_2(j, s) = \{X \mid X_j >= s\}$

James et al. 2014: An introduction to statistical learning
Hastie et al. 2009: Elements of statistical learning

Recursive binary splitting

and when to ask which question

- Classification trees (qualitative, categorical response)
  - Prediction: mode of each region
- Minimise Gini impurity:

$$G = 1 - \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

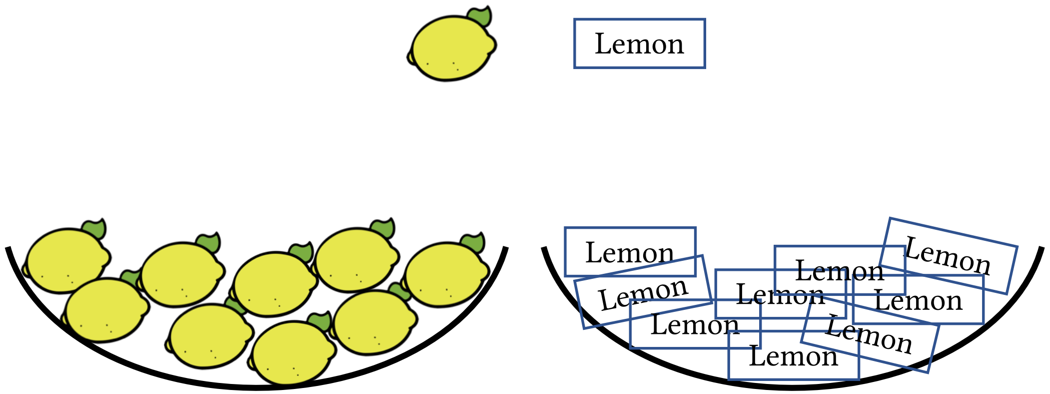
- $\hat{p}_{mk}$ is the proportion of observations in the mth region of class k
- Gini index small if proportion is close to 0 or 1
- Gini index: measure of node impurity

James et al. 2014: An introduction to statistical learning
Hastie et al. 2009: Elements of statistical learning

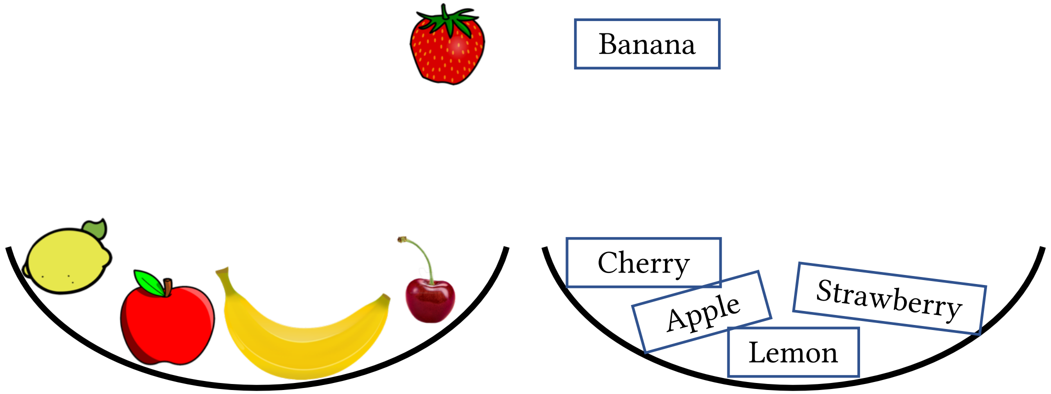
Gini impurity

Chance of being incorrect if you randomly assign a class label to an observation in the same set

Gini index/ impurity = 0



Gini index/ impurity = 0.8



https://github.com/random-forests/tutorials/blob/master/decision_tree.ipynb

Information gain

- Calculate impurity for root node (e.g. $n_0 = 5$, $G_0 = 0.8$)
- Try each possible question and calculate impurity for resulting child nodes (e.g. $n_1 = 4$, $G_1 - 0.63$ and $n_2 = 1$, $G_2 = 0.1$)
- Take weighted average:
    - higher weight on large set
    - E.g. $4/5 * 0.63 + 1/5 * 0.1 = 0.52$
- And subtract from root node impurity:

Information Gain = $0.8 - 0.52 = 0.28$

- Ask question with highest information gain at current node

R

https://github.com/bryantravissmith/FromScratch/blob/master/SupervisedLearning/DecisionTrees/Implementing%20Decision%20Trees%20From%20Scratch%20Using%20R.ipynb

What’s next?

- Splitting until information gain is zero leads to overfitting
  - Pruning
  - Cross-validation
- Single decisions trees are easy to interpret, but usually result in poor predictive power
- Random forests (next session)