Decision Trees

Aquiles Farias

Decision Trees - References

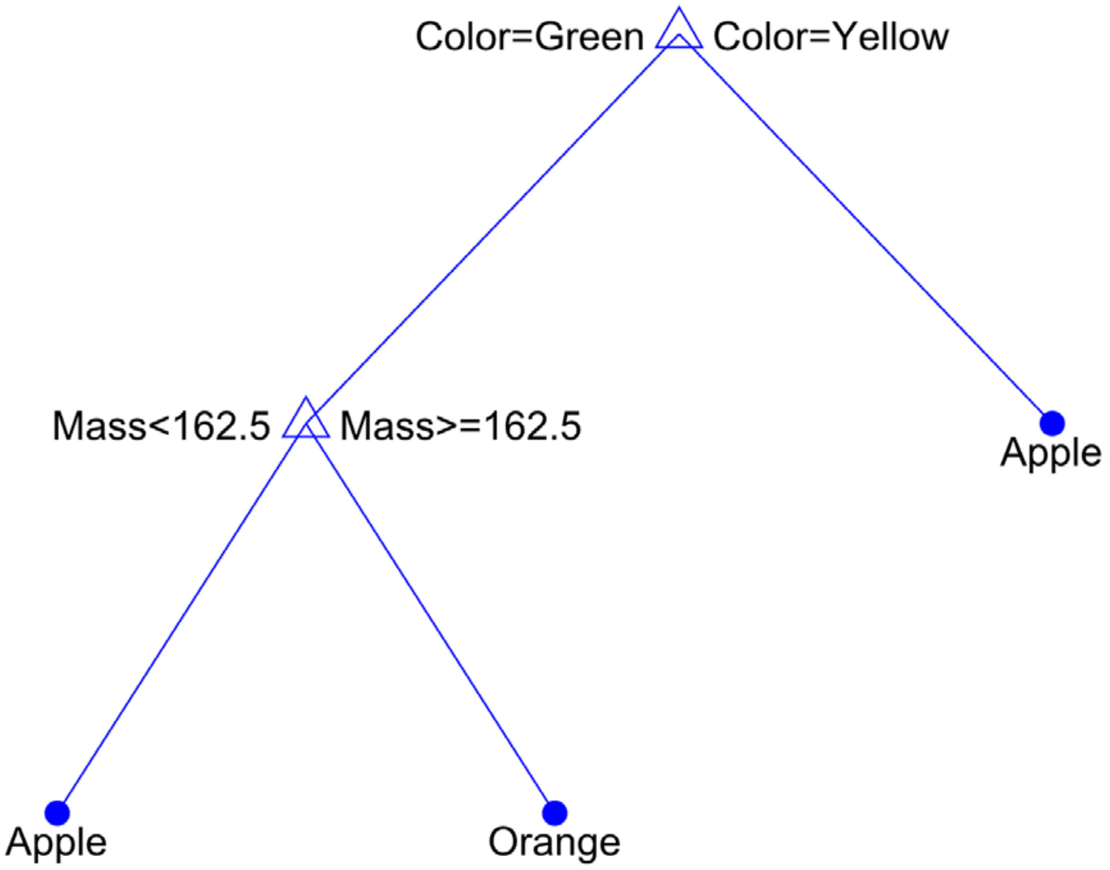
- [1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.
- [2] Coppersmith, D., S. J. Hong, and J. R. M. Hosking. “Partitioning Nominal Attributes in Decision Trees.” *Data Mining and Knowledge Discovery*, Vol. 3, 1999, pp. 197–217.
- [3] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.
- [4] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.
- [5] Quinlan, J. R. “Induction of Decision Trees”. Machine Learning 1: 81-106, Kluwer Academic Publishers, 1986
- [6] Quinlan, J. R. C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

Toy dataset

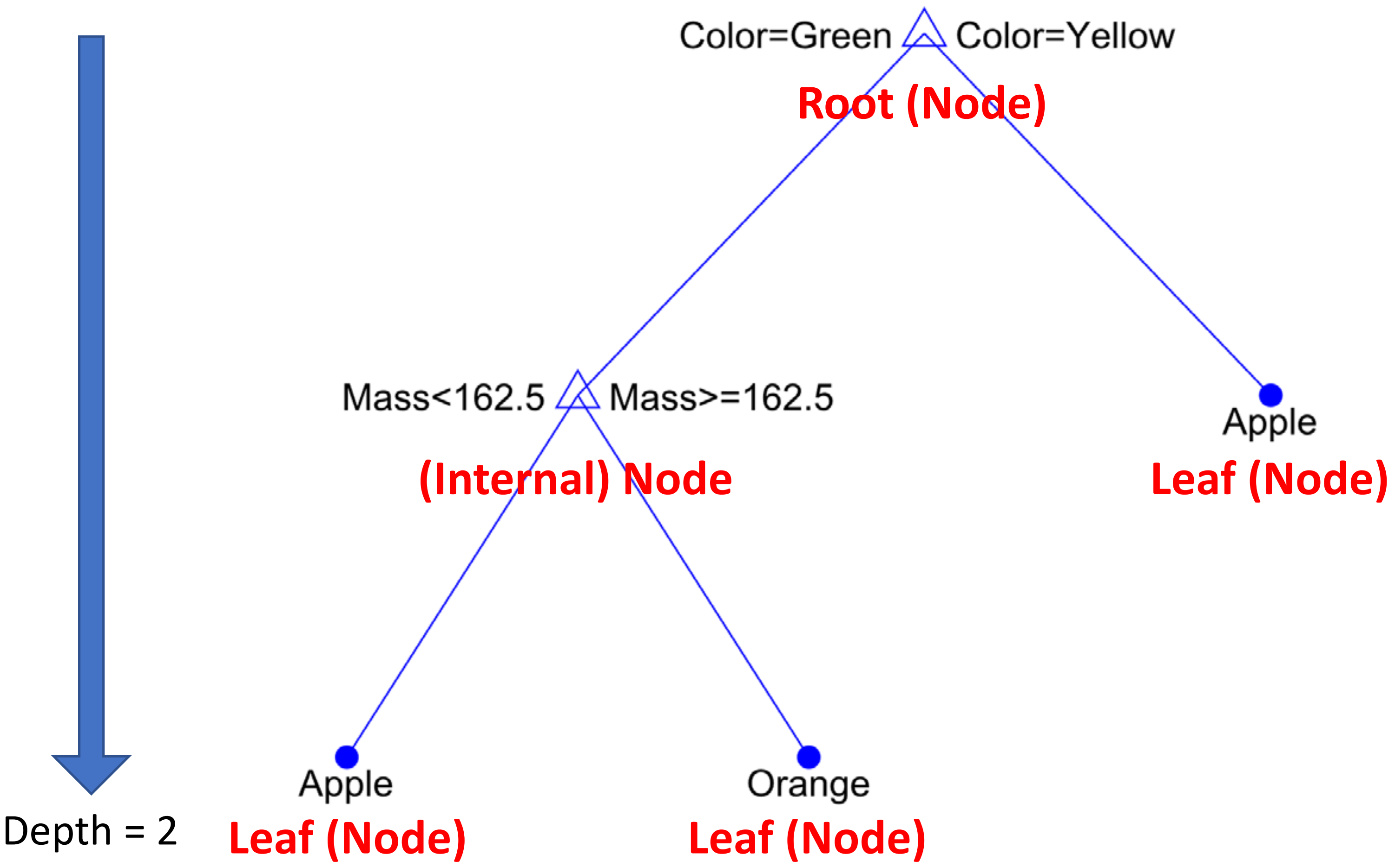
Color	Mass	Fruit
Green	163	Orange
Green	162	Apple
Yellow	164	Apple
Green	164	Orange
Yellow	168	Apple
Green	180	Orange

Decision Trees

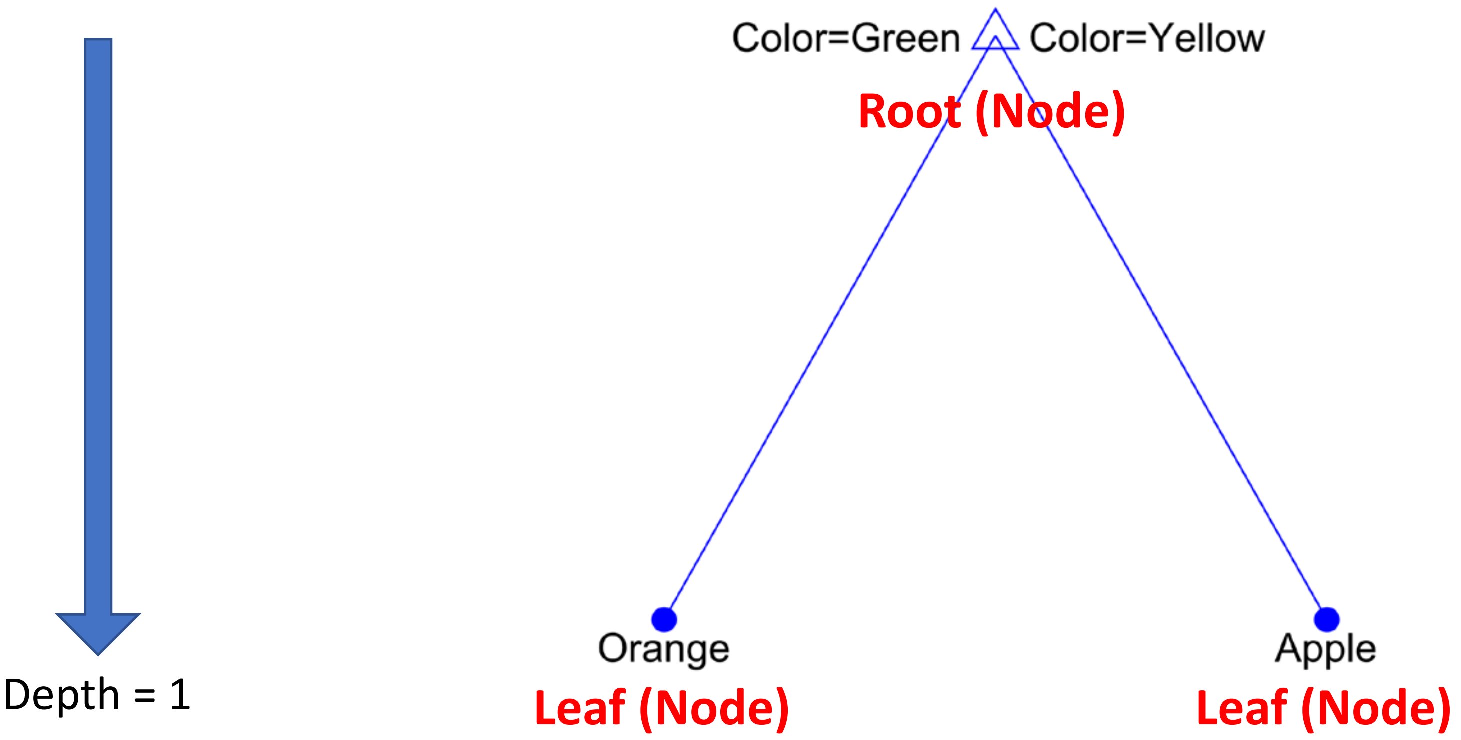
- Supervised learning
- Non-parametric
- Very flexible
- Easy to interpret
- Classification and Regression
- No need to rescale nor center the data



Decision Trees

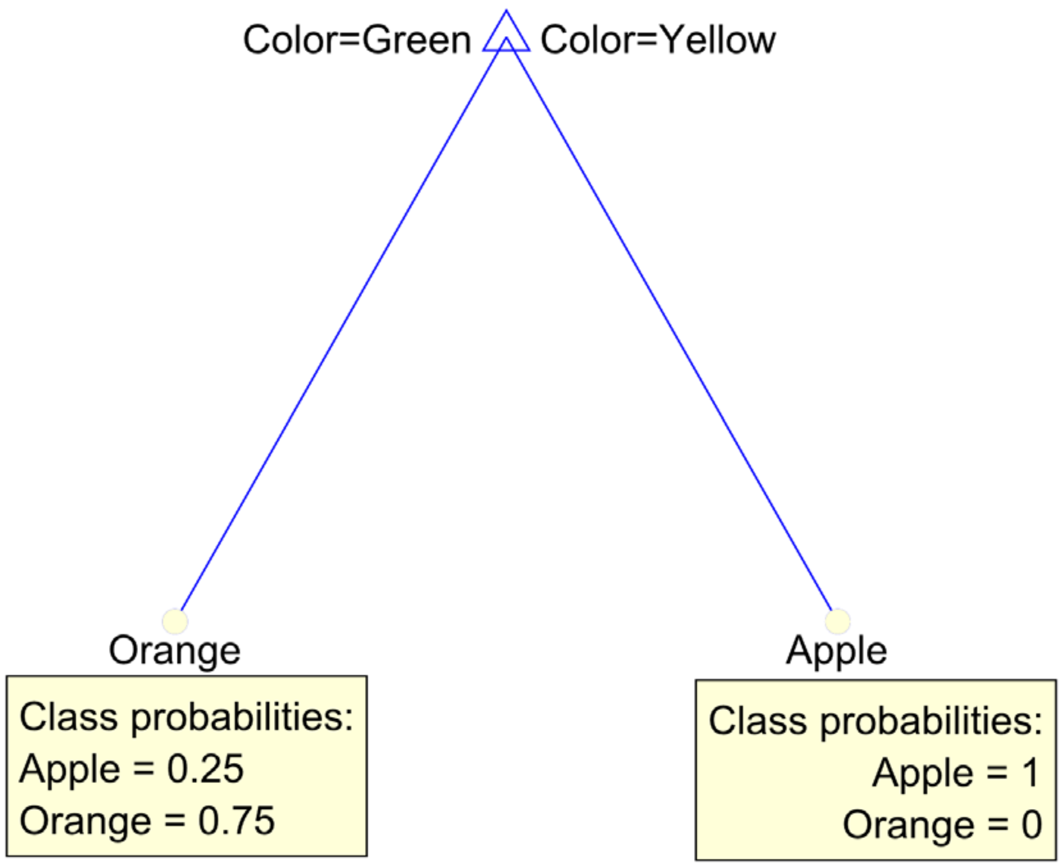


Decision Trees

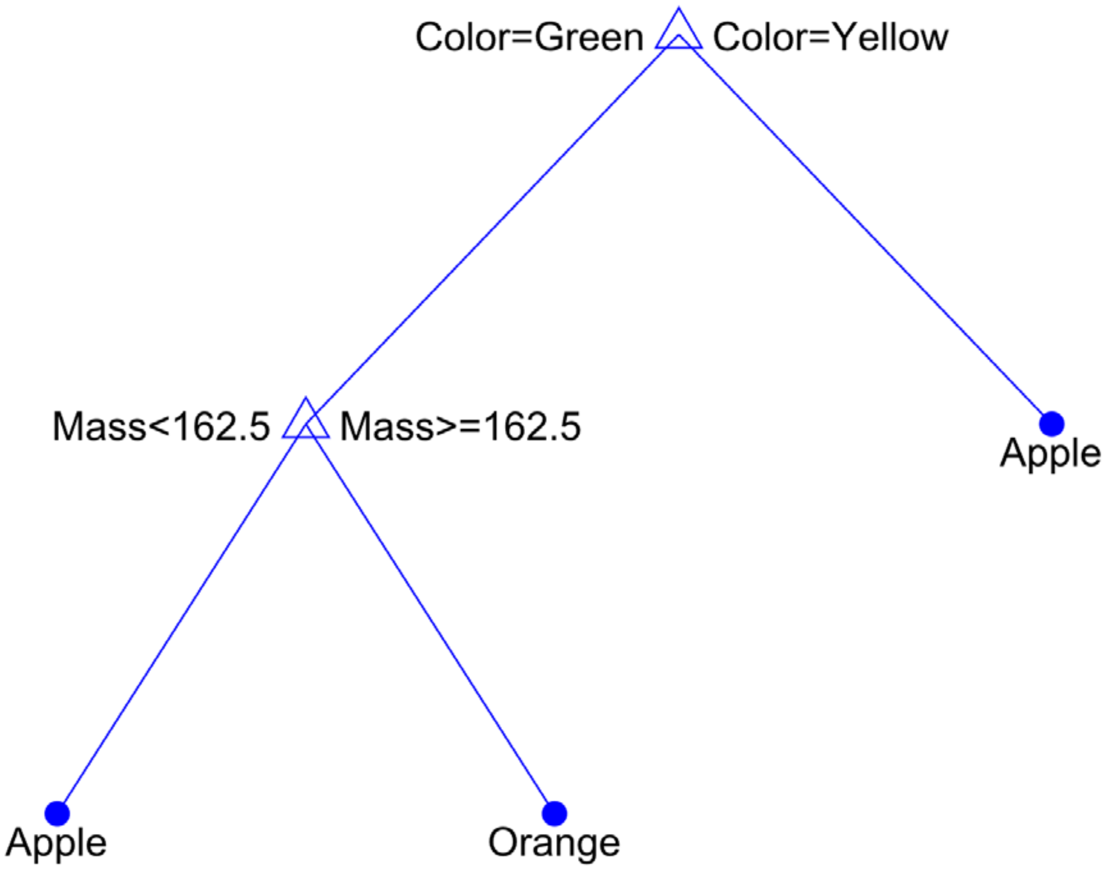


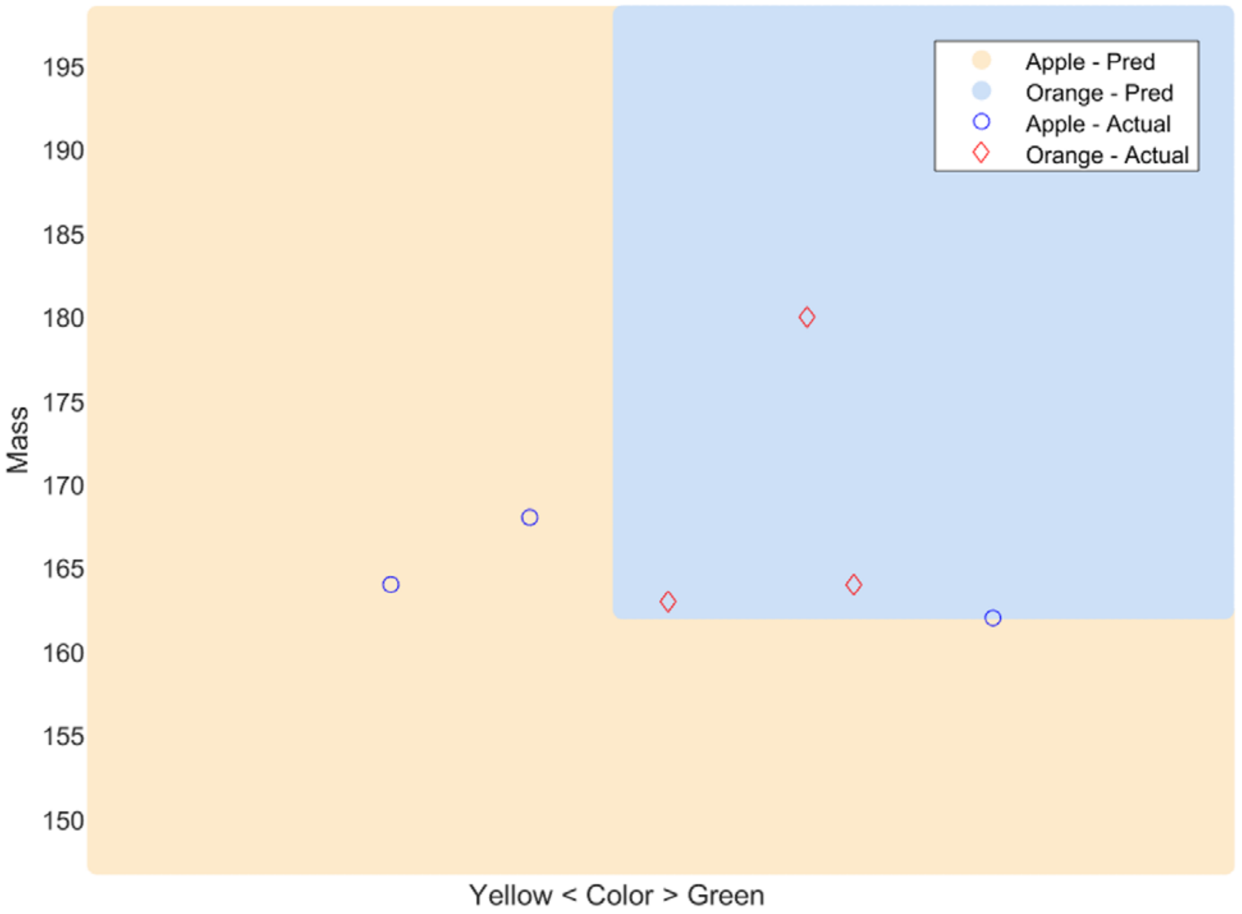
Decision Trees - Interpretability

Color	Mass	Fruit
Green	163	Orange
Green	162	Apple
Yellow	164	Apple
Green	164	Orange
Yellow	168	Apple
Green	180	Orange



Decision Tree – Decision Boundaries





Growing a classification tree

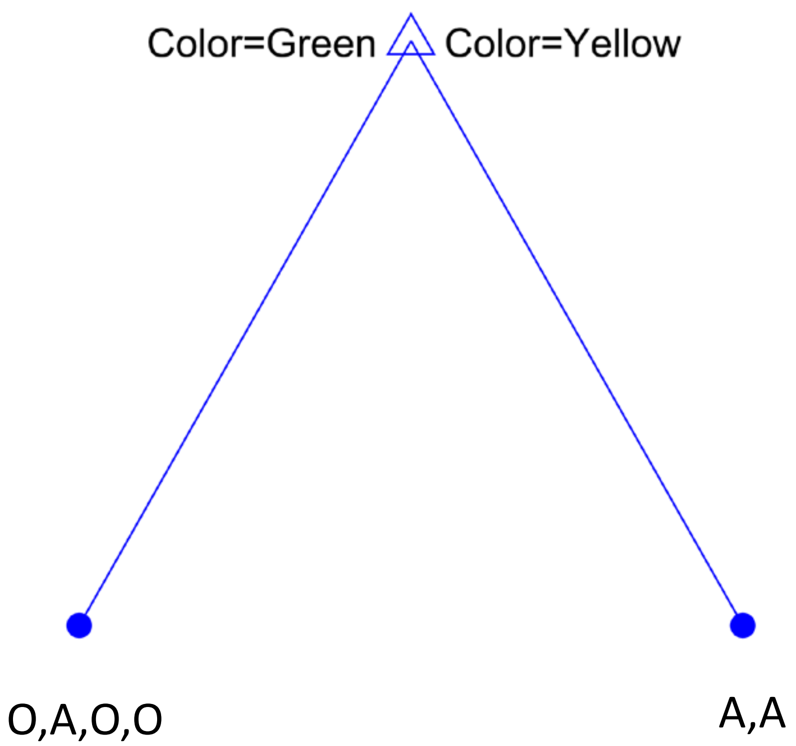
- Pick one of the predictor variables, x_i
- Pick a value of x_i, say s_i, that divides the training data into two (not necessarily equal) portions
- Measure how homogeneous each of the resulting portions are
- Try different values of x_i, and s_i to maximize purity in initial split
- After you get the best split, repeat the process for a second split, and so on until reaching stop criteria

Growing a tree

Color	Mass	Fruit
Green	163	Orange
Green	162	Apple
Yellow	164	Apple
Green	164	Orange
Yellow	168	Apple
Green	180	Orange

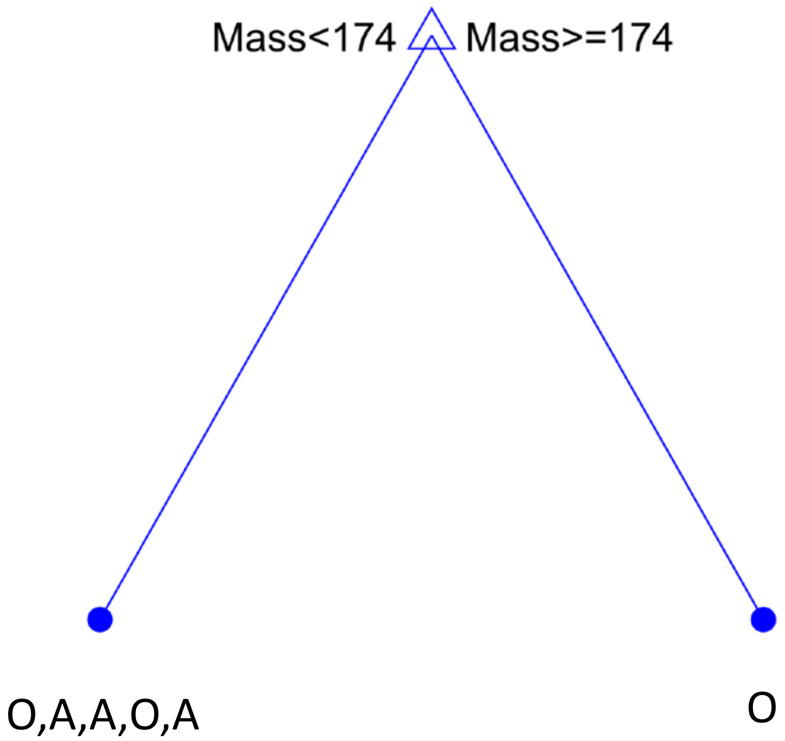
Growing a tree – Splitting using Color

Color	Mass	Fruit
Green	163	Orange
Green	162	Apple
Yellow	164	Apple
Green	164	Orange
Yellow	168	Apple
Green	180	Orange



Growing a tree – Splitting using Mass

Color	Mass	Fruit
Green	163	Orange
Green	162	Apple
Yellow	164	Apple
Green	164	Orange
Yellow	168	Apple
Green	180	Orange



Split criterion

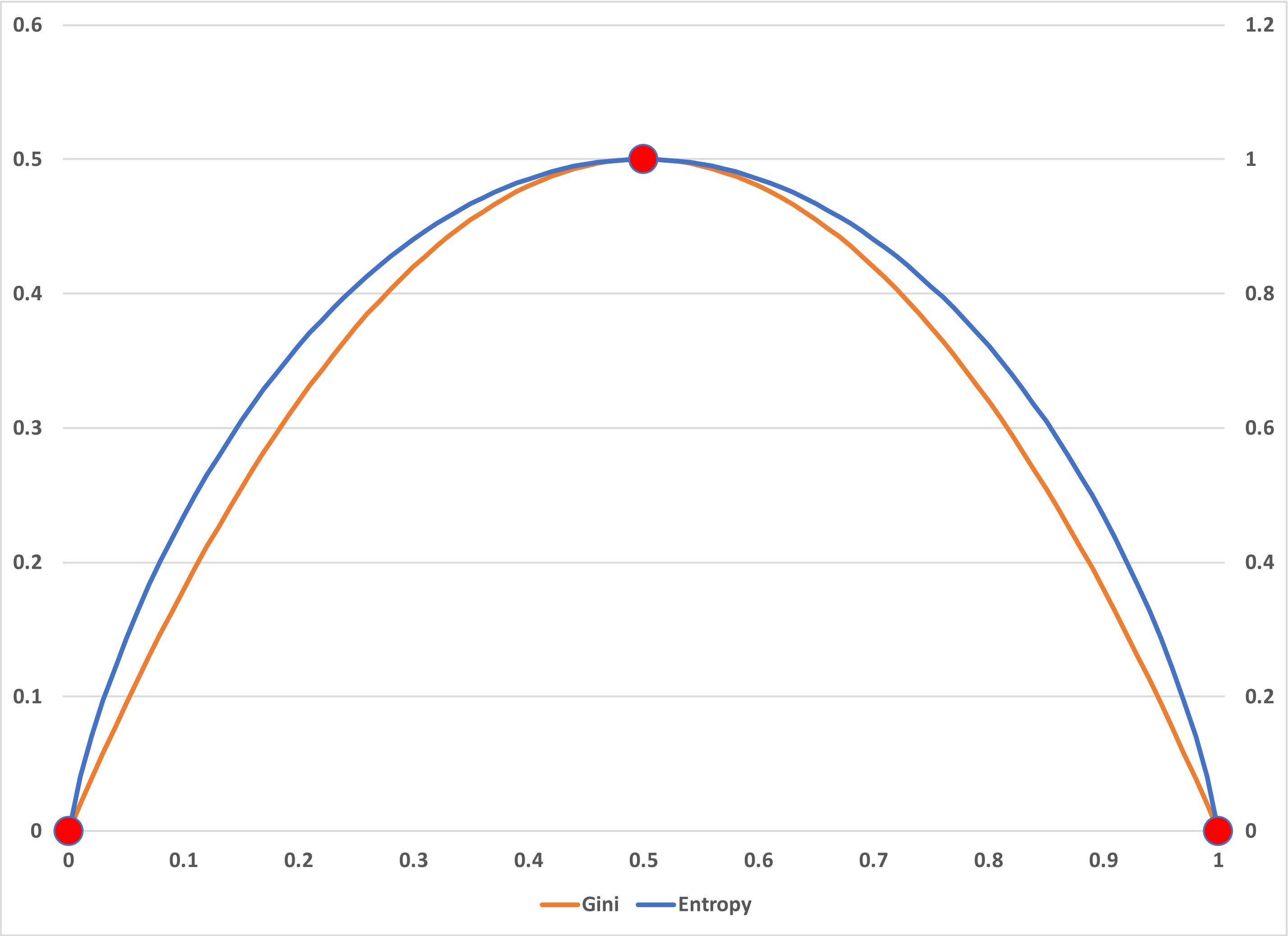
Gini Impurity Index

$$I_G(p) = p_1(1-p_1) + \\ + p_2(1-p_2) = \\ = 2p_1(1-p_1)$$

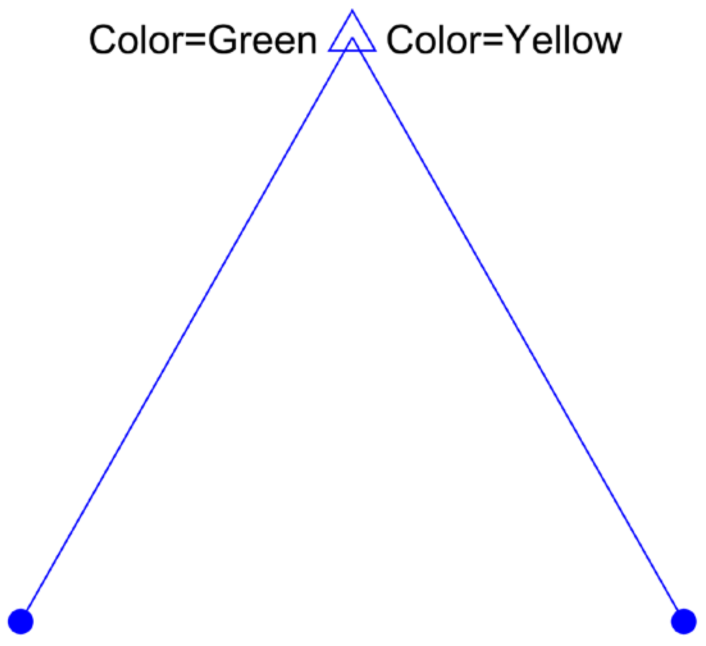
Entropy

$$I_E(p) = -(p_1 \log p_1 + p_2 \log p_2)$$

$p = \{Apples, Oranges\} - set\ of\ items$



Growing a tree – Splitting using Color

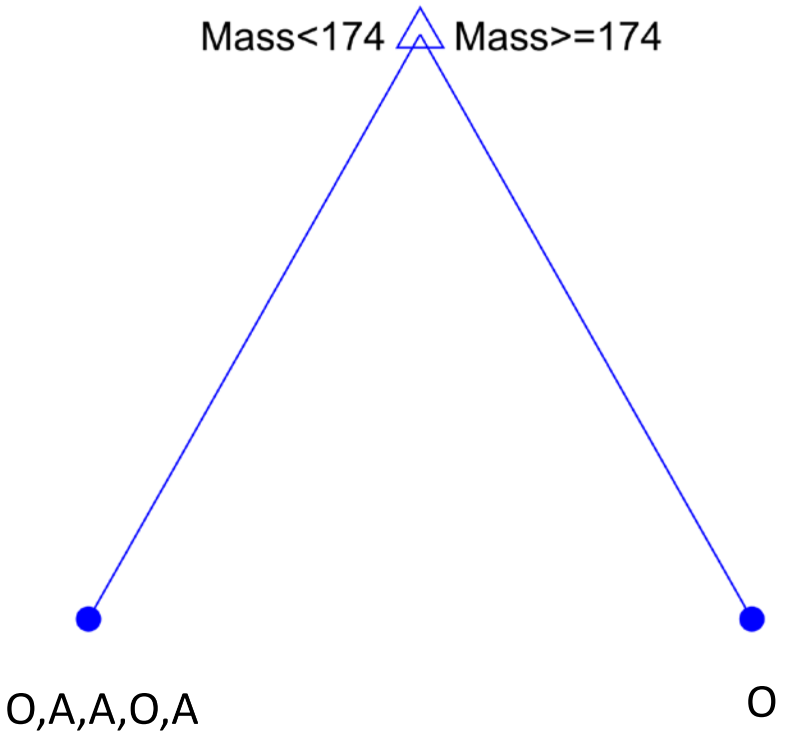


Node	p(A)	p(O)	Gini	Entropy
Root	0.50	0.50	**0.50**	**1**
Splitting Color			**0.25**	**0.41**
Left (Color=G)	0.25	0.75	0.38	0.81
Right (Color=Y)	1.00	0.00	0	0.00
Splitting Mass			**0.40**	**0.49**
Left (Mass<174)	0.60	0.40	0.48	0.97
Right (Mass>=174)	1.00	0.00	0	0.00

$$Information\ Gain(D_P) = I(D_P) - \frac{N_{Left}}{N_P} I(D_{Left}) - \frac{N_{Right}}{N_P} I(D_{Right})$$

$$I = Information\ criterion, D = Dataset, N = \#elements$$

Growing a tree – Splitting using Mass

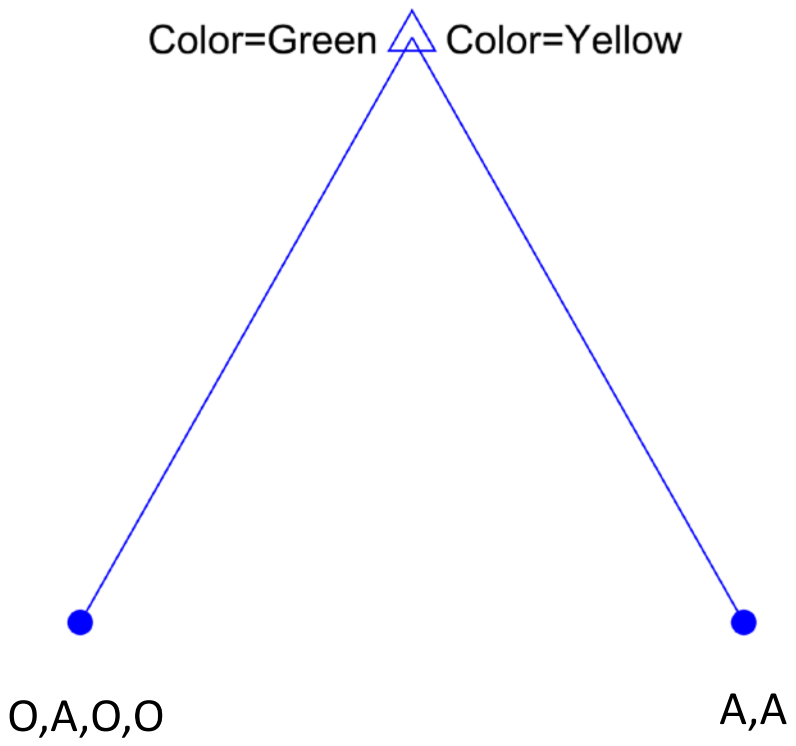


Node	p(A)	p(O)	Gini	Entropy
Root	0.50	0.50	**0.50**	**1**
Splitting Color			**0.25**	**0.41**
Left (Color=G)	0.25	0.75	0.38	0.81
Right (Color=Y)	1.00	0.00	0	0.00
Splitting Mass			**0.40**	**0.49**
Left (Mass<174)	0.60	0.40	0.48	0.97
Right (Mass>=174)	1.00	0.00	0	0.00

$$Information\ Gain(D_P) = I(D_P) - \frac{N_{Left}}{N_P} I(D_{Left}) - \frac{N_{Right}}{N_P} I(D_{Right})$$

$$I = Information\ criterion, D = Dataset, N = \#elements$$

Growing a tree – Color wins

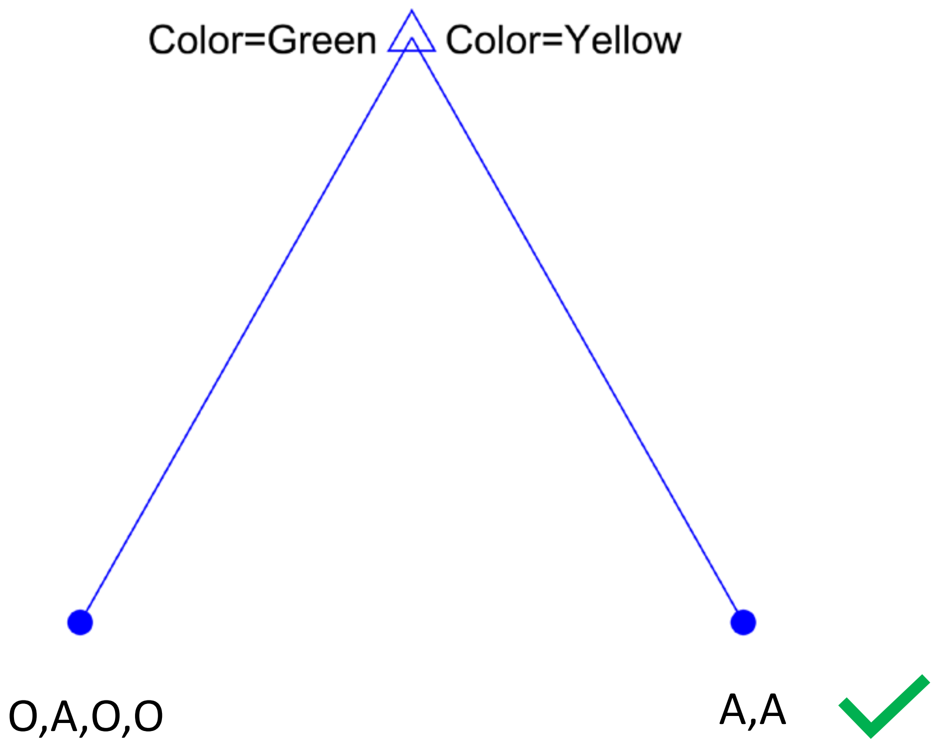


Node	p(A)	p(O)	Gini	Entropy
Root	0.50	0.50	**0.50**	**1**
Splitting Color			**0.25**	**0.41**
Left (Color=G)	0.25	0.75	0.38	0.81
Right (Color=Y)	1.00	0.00	0	0.00
Splitting Mass			**0.40**	**0.49**
Left (Mass<174)	0.60	0.40	0.48	0.97
Right (Mass>=174)	1.00	0.00	0	0.00

$$Information\ Gain(D_P) = I(D_P) - \frac{N_{Left}}{N_P} I(D_{Left}) - \frac{N_{Right}}{N_P} I(D_{Right})$$

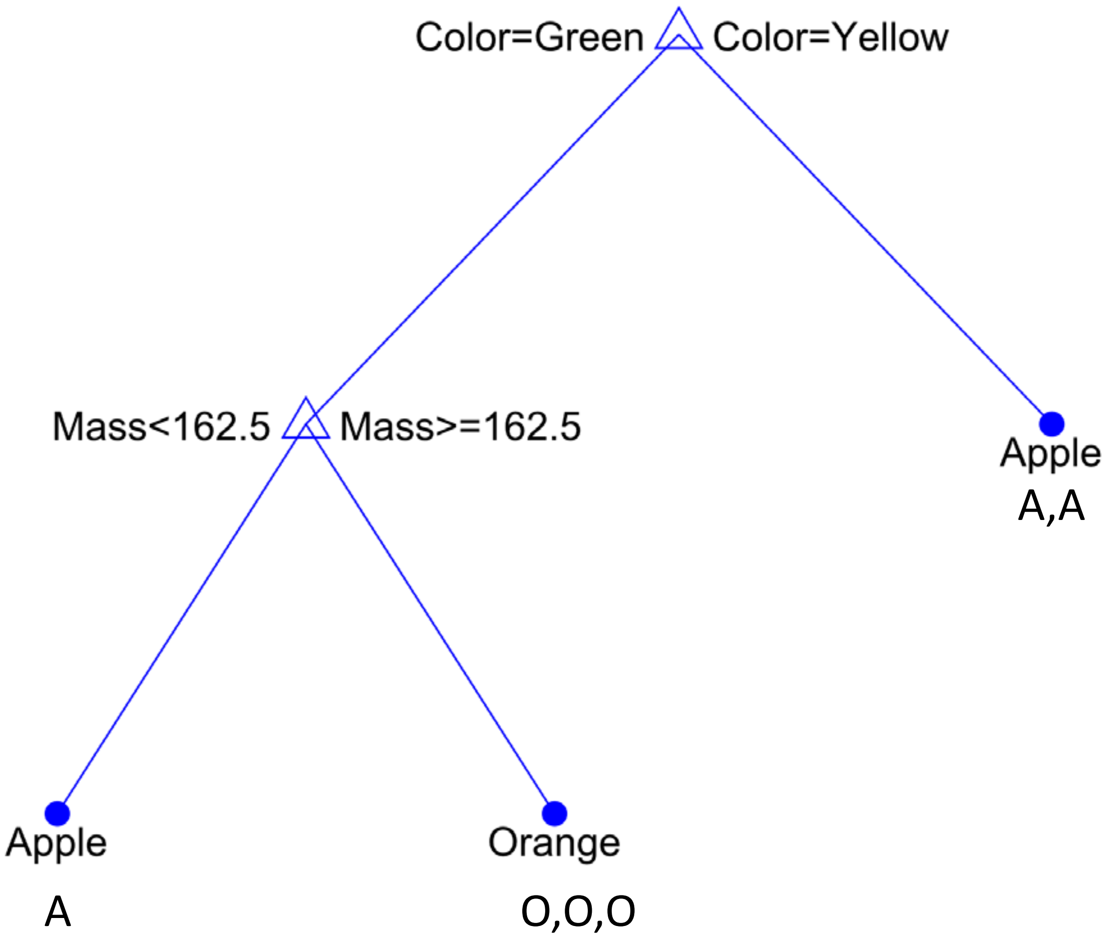
$$I = Information\ criterion, D = Dataset, N = \#elements$$

Growing a bigger tree



Node	p(A)	p(O)	Gini	Entropy
Color = Green	0.25	0.75	**0.38**	**0.811278**
Splitting @ 162.5			**0**	**0.00**
Left (Mass<162.5)	1.00	0.00	0	0.00
Right (Mass>=162.5)	0.00	1.00	0	0.00
Splitting @ 163.5			**0.25**	**0.50**
Left (Mass<163.5)	0.50	0.50	0.5	1.00
Right (Mass>=163.5)	0.00	1.00	0	0.00
Splitting @ 172			**0.33**	**0.46**
Left (Mass<172)	0.33	0.67	0.44	0.92
Right (Mass>=172)	0.00	1.00	0	0.00

Growing a bigger tree



Node	p(A)	p(O)	Gini	Entropy
Color = Green	0.25	0.75	**0.38**	**0.811278**
Splitting @ 162.5			**0**	**0.00**
Left (Mass<162.5)	1.00	0.00	0	0.00
Right (Mass>=162.5)	0.00	1.00	0	0.00
Splitting @ 163.5			**0.25**	**0.50**
Left (Mass<163.5)	0.50	0.50	0.5	1.00
Right (Mass>=163.5)	0.00	1.00	0	0.00
Splitting @ 172			**0.33**	**0.46**
Left (Mass<172)	0.33	0.67	0.44	0.92
Right (Mass>=172)	0.00	1.00	0	0.00

Growing a classification tree

- Pick one of the predictor variables, x_i
- Pick a value of x_i, say s_i, that divid[...] two (not necessarily equal) portion[...]
- Measure how [...]ortions are
- Try [...]ze purity in initial split
- Afte[...]peat the process for a second split, and so o[...]p criteria

Finding the optimal tree is NP-Complete -> The time required to solve the problem using any currently known algorithm increases very quickly as the size of the problem.

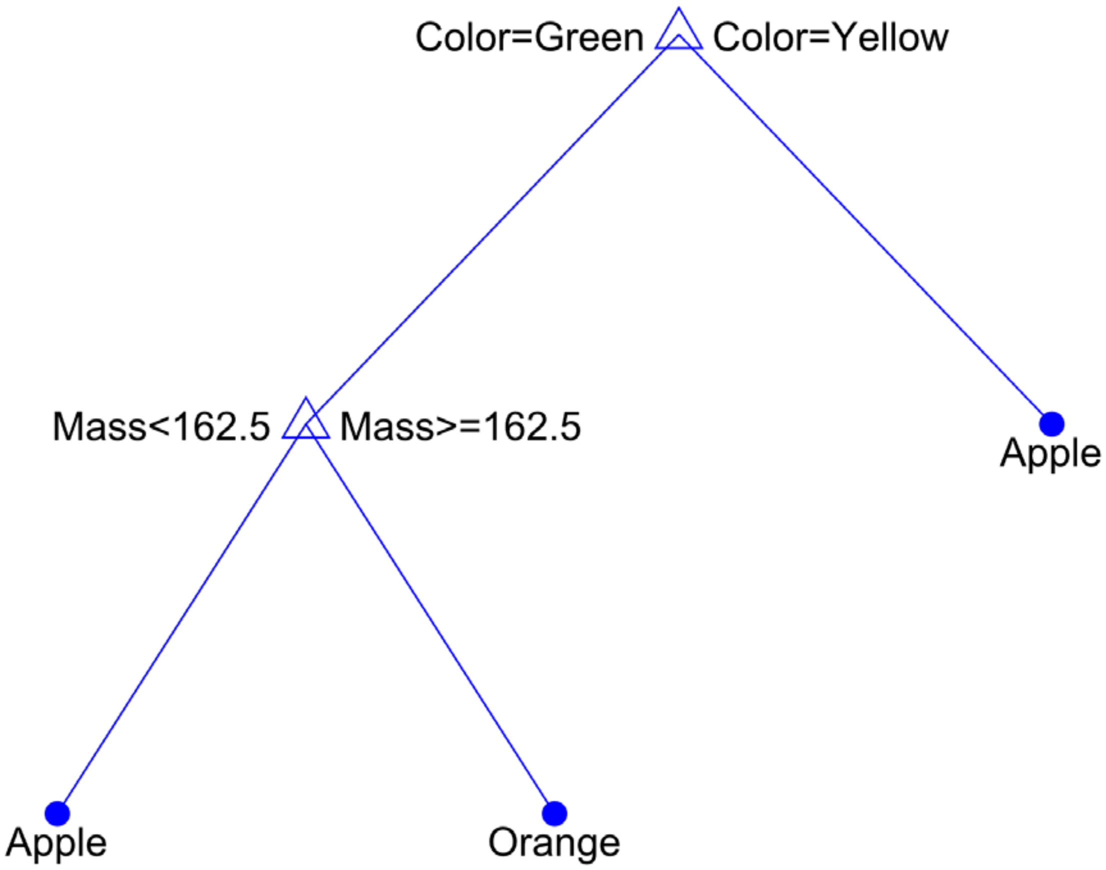
Growing a classification tree

- Stop criteria
  - No more information gain splitting the parent
    - All leaf nodes are homogeneous (perfect classification)
    - Impossible to disentangle data points (same attributes for different targets)
  - Maximum depth
  - Maximum number of decision splits
  - Minimum number of leaf node observations
  - Minimum parent size
  - Other
- Pruning

Pruning

- After growing a tree you may prune it to reduce complexity (and avoid overfitting)
  - Use a validation set for this or use cross-validation
  - For each (internal) node *n* calculate the loss function value (validation set) of the entire subtree (or the tree)
  - Replace the node with majority class label and calculate the new loss (function value)
  - Prune the node with highest loss reduction (alternatively the 1-sd rule can be used)
  - Repeat this procedure until no node is pruned

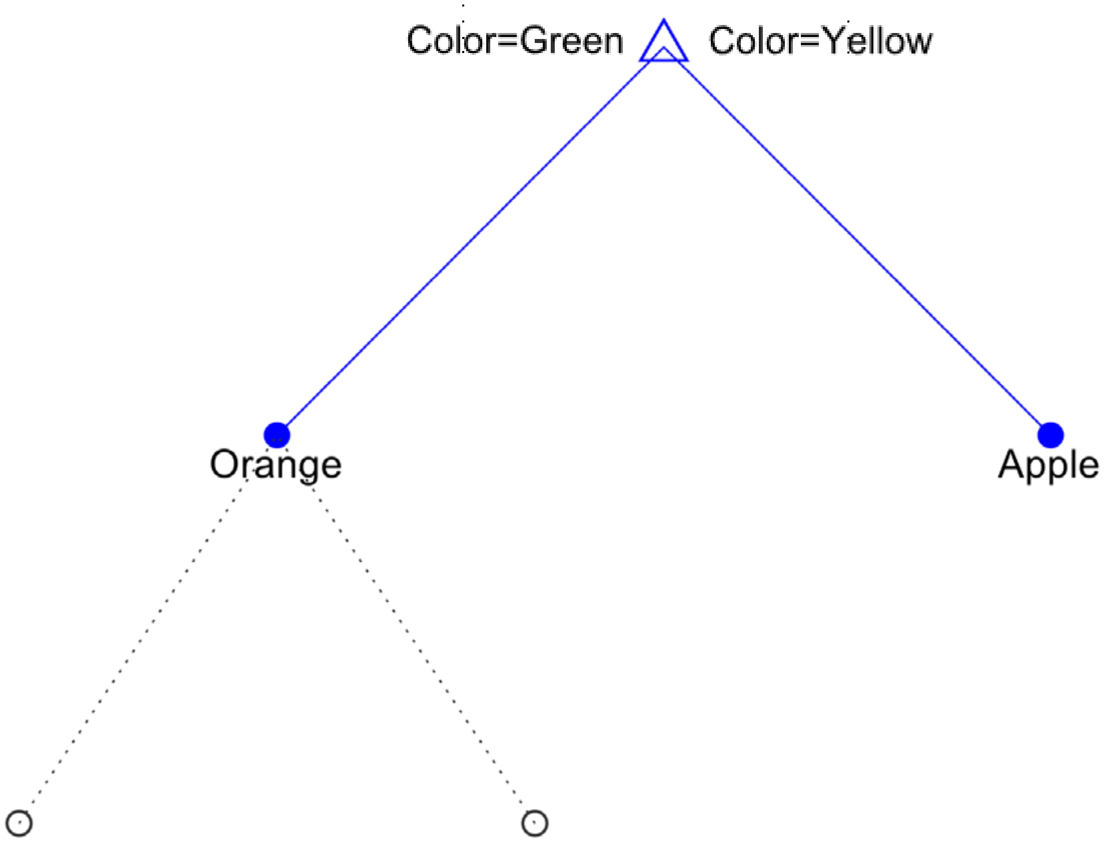
Pruning a tree



Validation Set		
Color	**Mass**	**Fruit**
Yellow	156	Apple
Green	156	Orange
Green	164	Apple
Green	190	Orange

Accuracy = 2/4 = 50%

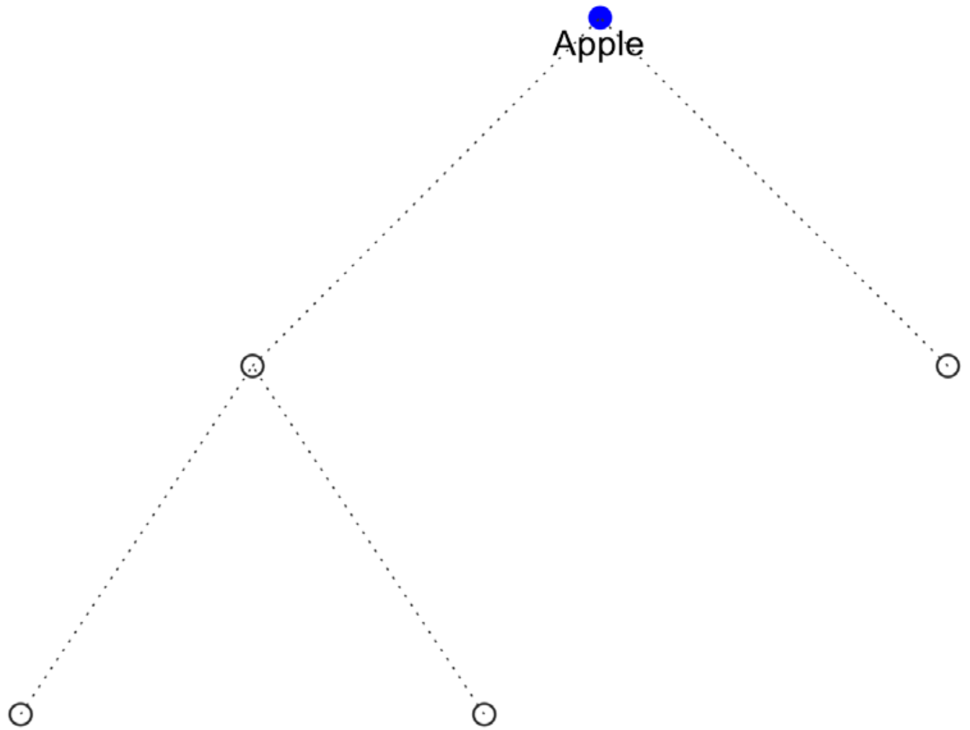
Pruning a tree



Validation Set		
Color	**Mass**	**Fruit**
Yellow	156	Apple
Green	156	Orange
Green	164	Apple
Green	190	Orange

Accuracy = 3/4 = 75%

Pruning a tree – One more?



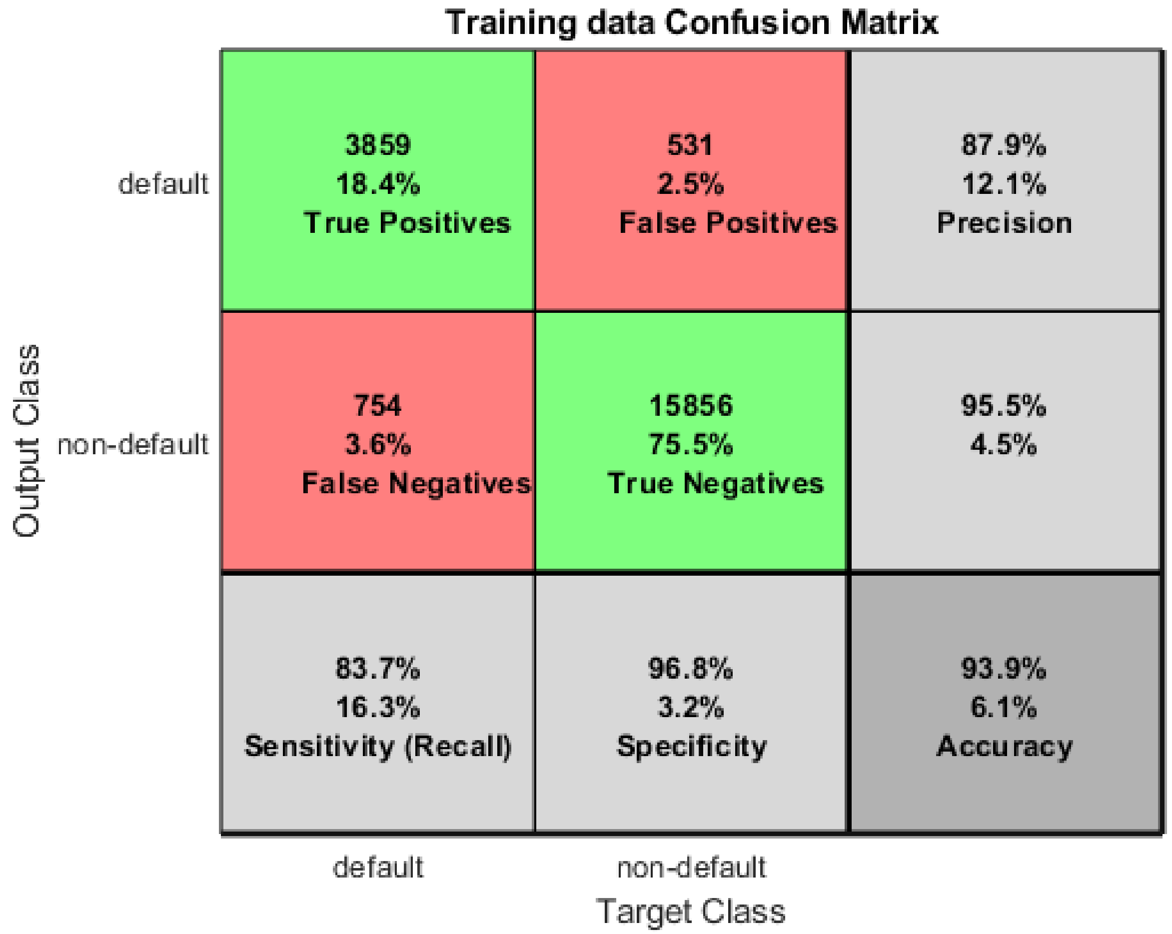
Validation Set		
Color	**Mass**	**Fruit**
Yellow	156	Apple
Green	156	Orange
Green	164	Apple
Green	190	Orange

Accuracy = 2/4 = 50%

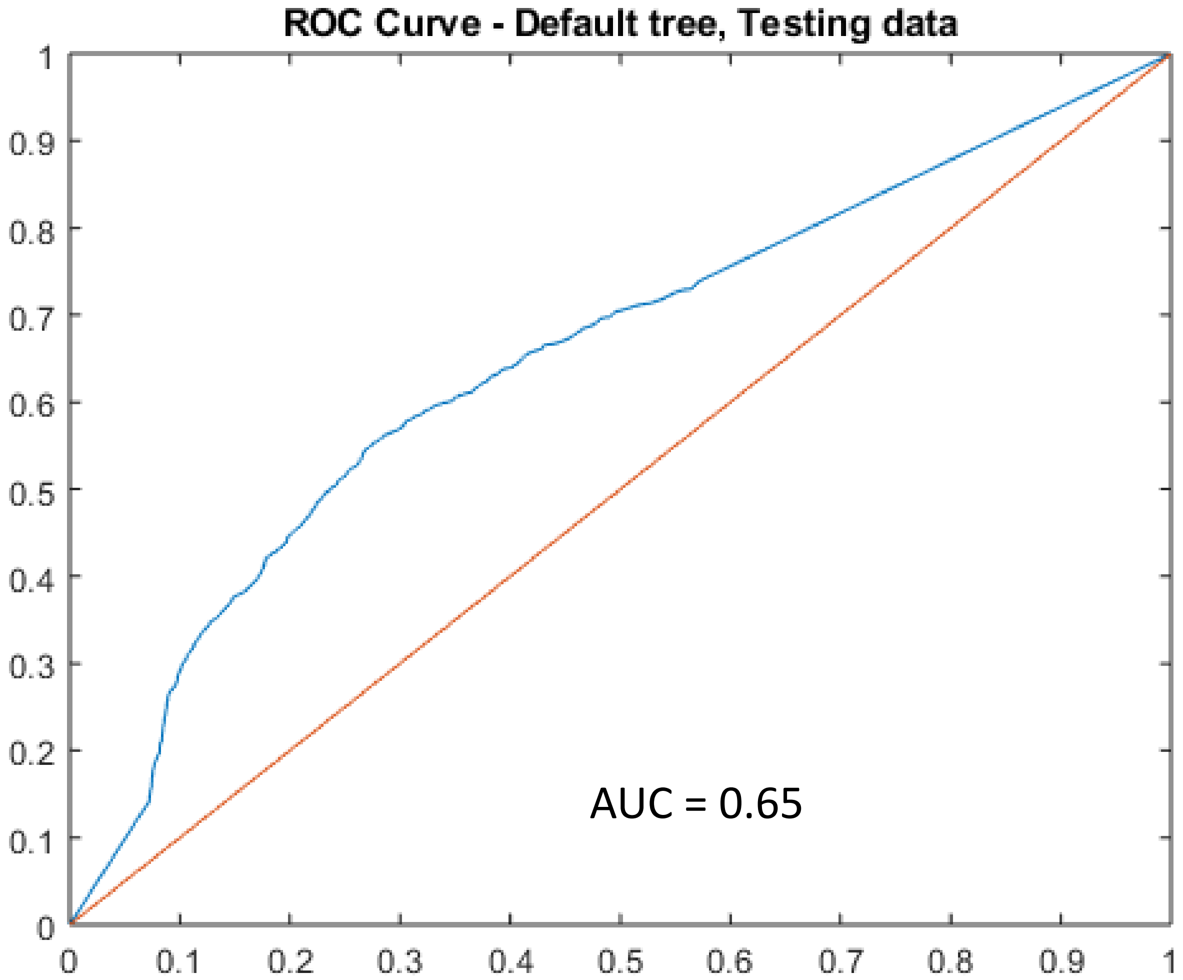
Classification Tree – performance evaluation

- Confusion matrix
- Precision
- Accuracy
- Sensitivity (recall)
- Specificity
- The ROC curve
- Area under the curve

Classification Tree – Confusion Matrix

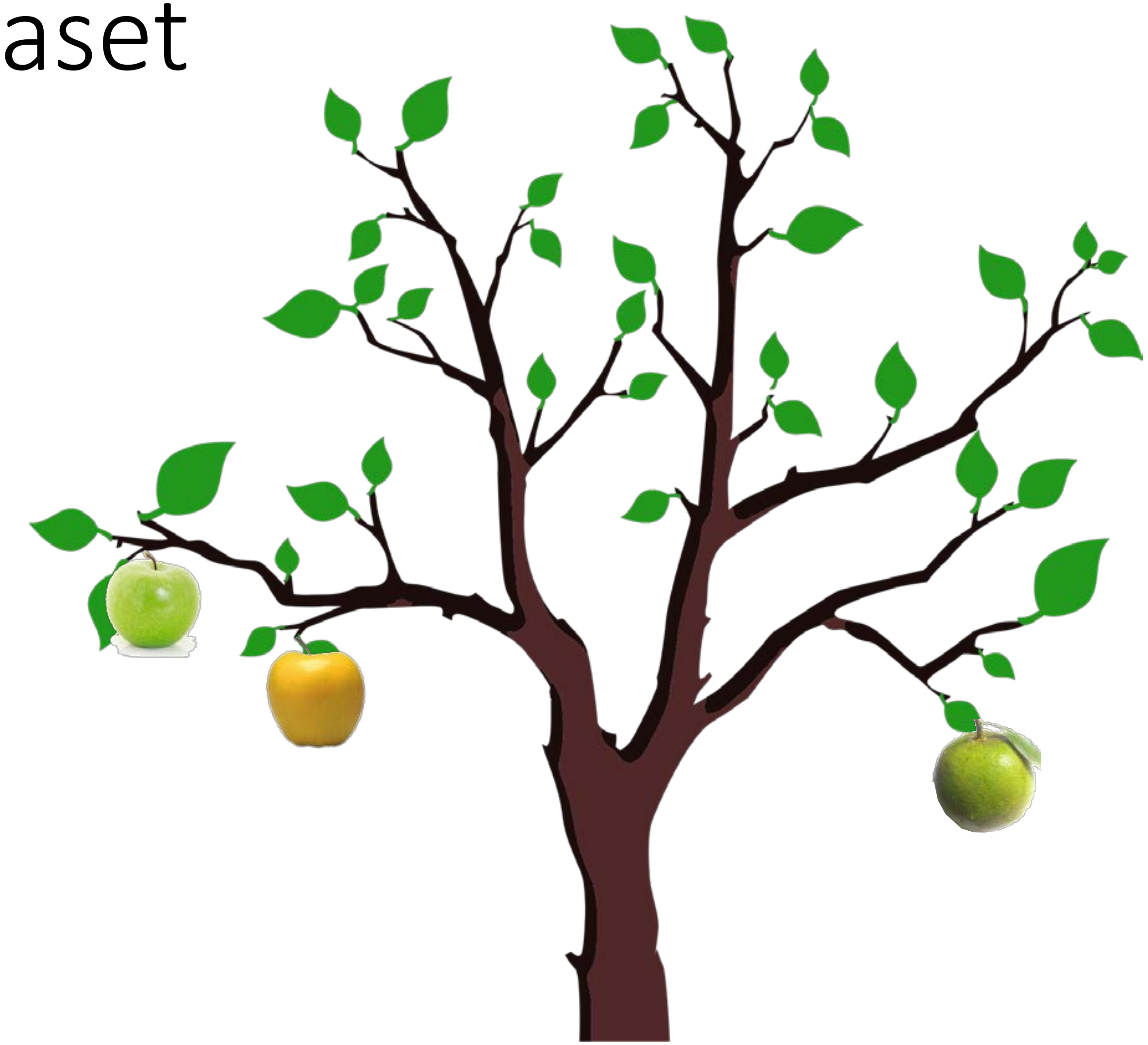


Classification Tree – Confusion Matrix



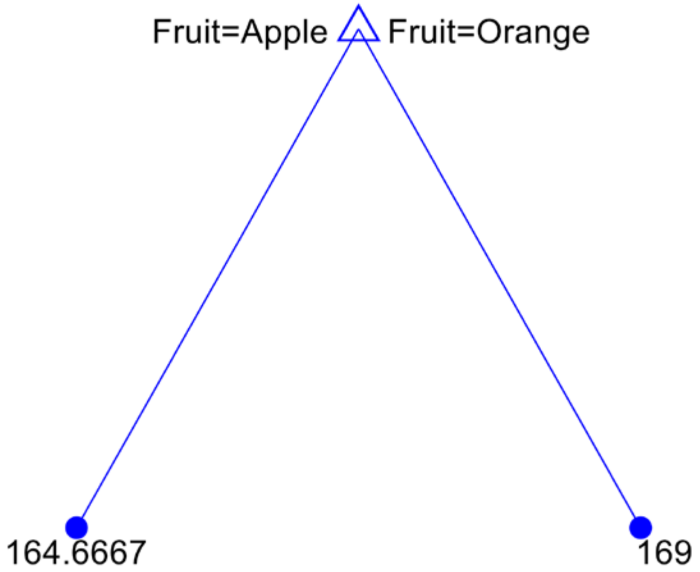
Regression tree - Toy dataset

Fruit	Width	Mass
Apple	7.5	162
Orange	7.1	163
Apple	7.3	164
Orange	7.2	164
Apple	7.5	168
Orange	7.6	180



Growing a regression tree – Splitting using Fruit

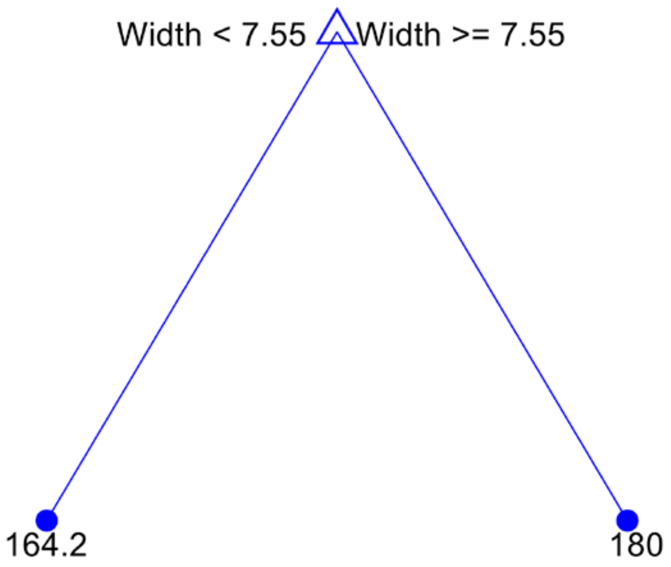
Fruit	Width	Mass
Apple	7.5	162
Orange	7.1	163
Apple	7.3	164
Orange	7.2	164
Apple	7.5	168
Orange	7.6	180



$$Root\ Mean\ Squared\ Error = RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}{N}} = 5.78$$

Growing a regression tree – Splitting using Width

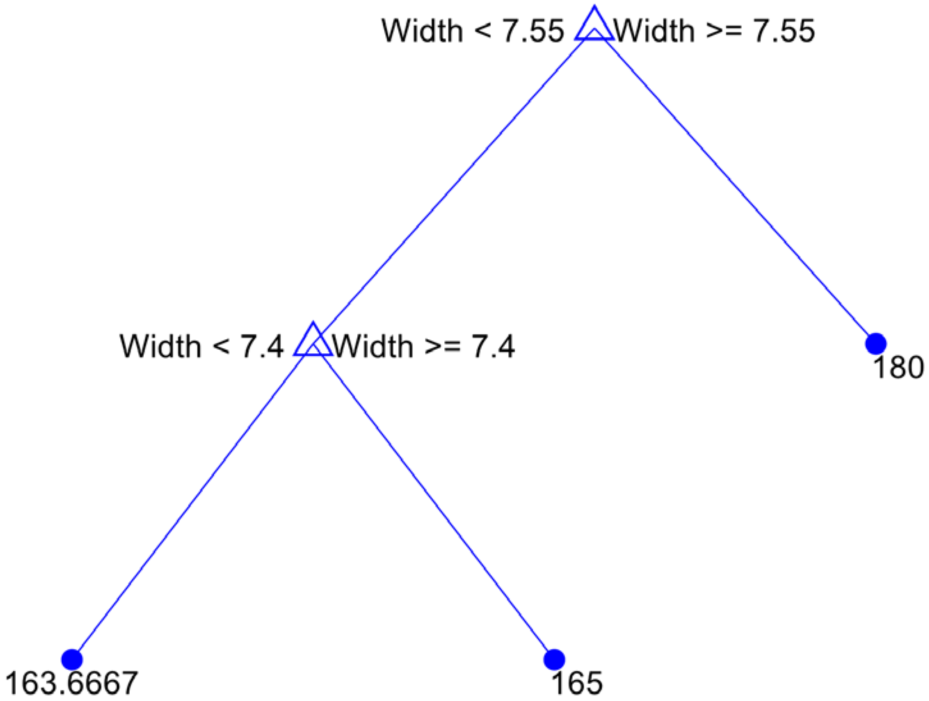
Fruit	Width	Mass
Apple	7.5	162
Orange	7.1	163
Apple	7.3	164
Orange	7.2	164
Apple	7.5	168
Orange	7.6	180



$$Root\ Mean\ Squared\ Error = RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}{N}} = 1.86$$

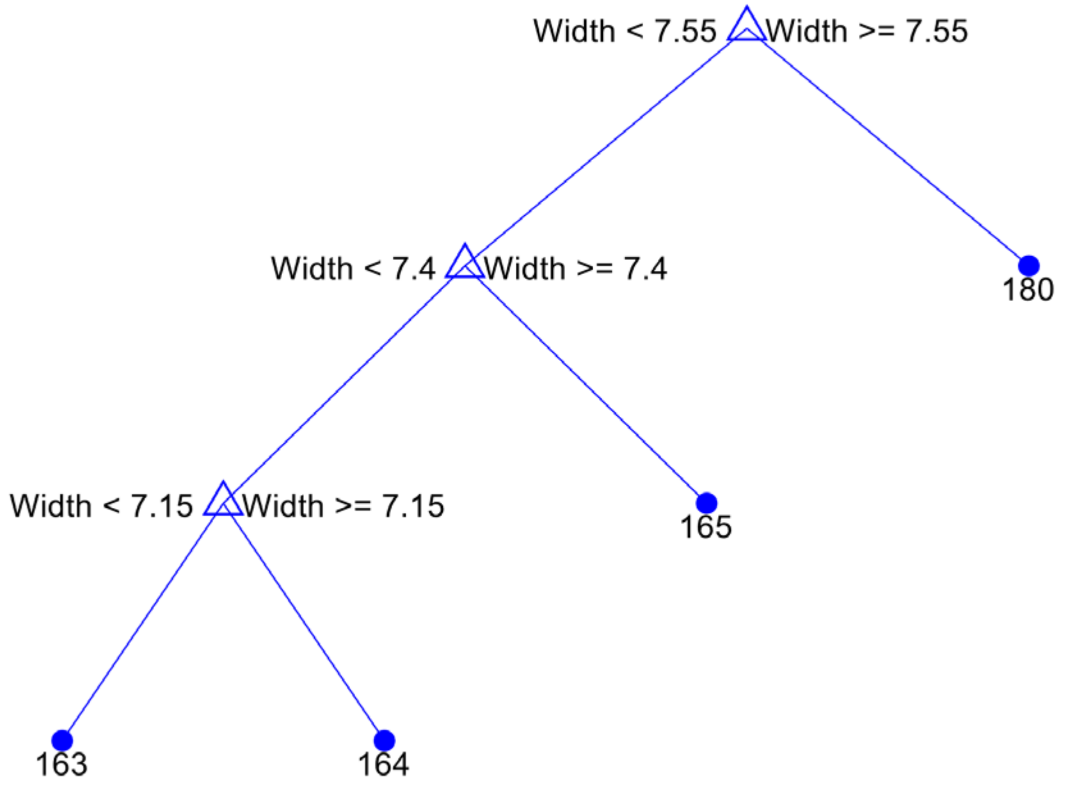
Growing Regression Trees

Fruit	Width	Mass
Apple	7.5	162
Orange	7.1	163
Apple	7.3	164
Orange	7.2	164
Apple	7.5	168
Orange	7.6	180



Growing Regression Trees

Fruit	Width	Mass
Apple	7.5	162
Orange	7.1	163
Apple	7.3	164
Orange	7.2	164
Apple	7.5	168
Orange	7.6	180



Decision Tree - Caveats

- Problems with decision trees
  - Expensive (computational point of view)
  - Instability - Changes easily with new data
  - Data rotation could be a problem
    - Only partitions the data using rectangles

Hands-on