https://ttpoll.com/p/643115

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Lecture 12: Decision Trees

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Class Announcements

Homework:

Previous homework keys online
Don’t expect TA support during weekends.

Course Project:

Teaming issues.

Lectures:

On October 1st, no attendance record due to the Engineering Expo

Exams:

Exam #1: Thursday, 10/03

- Online
- Window 11 am to 1 pm
- 75 mins
- ***SDS accommodations set in Canvas***.

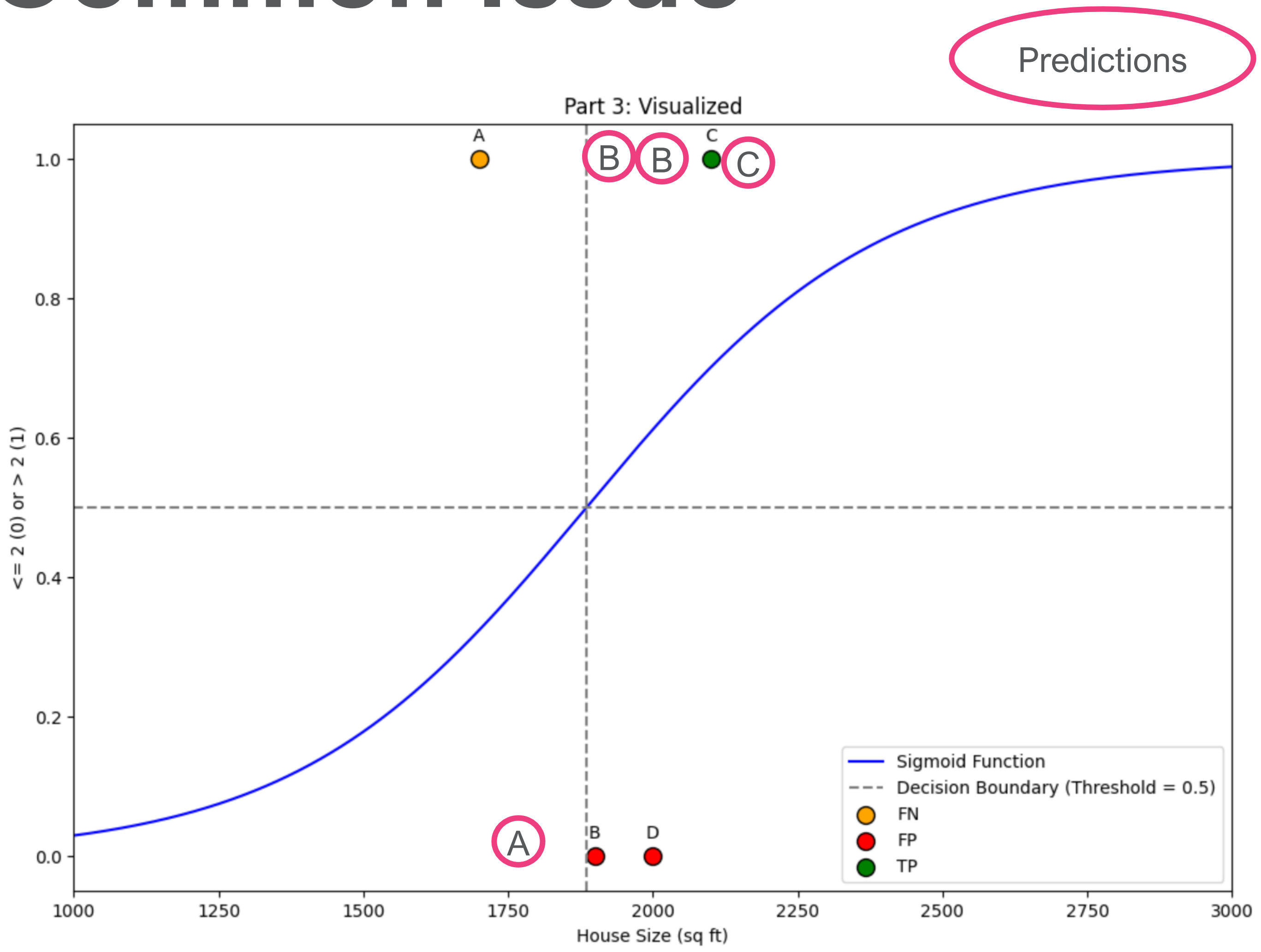
Homework 2 Most Common Issue

True Values: A=1, B=0, C=1, D=0

•**A:** is a False Negative because the model predicted 0 but the actual class is 1.

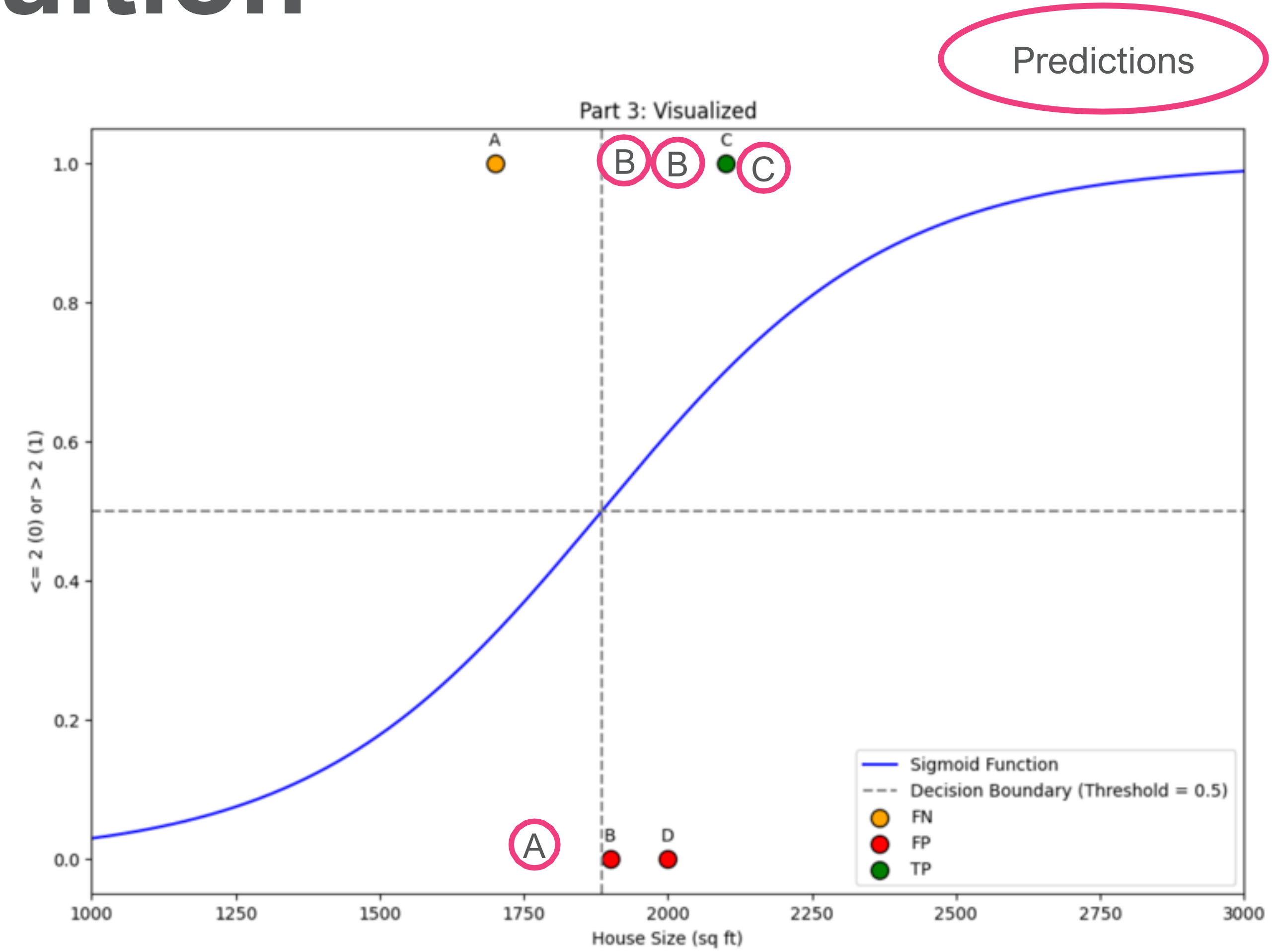
•**B** and **D:** are False Positives because the model predicted 1, but the actual class is 0.

•**C:** is a True Positive because the model correctly predicted 1, which matches the actual class.



Confusion Matrix Intuition

- Options: TP, TN, FP, FN
- Did the model make a mistake?
  - First letter: F
  - Otherwise, the first letter is T.
- Is the prediction Positive?
  - Second letter P.
  - Otherwise, the second letter is N.



Review

- Regularization techniques
  - L2 (ridge) – Penalize large weights and reward weights smaller than one.
  - L1 (Lasso) – Good for feature reduction.
  - ElasticNet – Combines L2 and L1
  - Early stop – Our last resource
- Decision Trees
  - Top-down iterative process
    - Select “best” feature
    - Ask a question on the feature to split data
    - Repeat steps on splits until purity or no new information for new splits.

Pop Quiz

When considering the course material so far, which statement resonates with you the most?

A. I understand the ML concepts so far and can explain them in my own words.

B. I'm not completely sure about the ML concepts so far and doubt I could explain them.

C. I don't yet understand the ML concepts and cannot explain them.

Today’s Topics

Decision Trees

Splitting Criteria

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

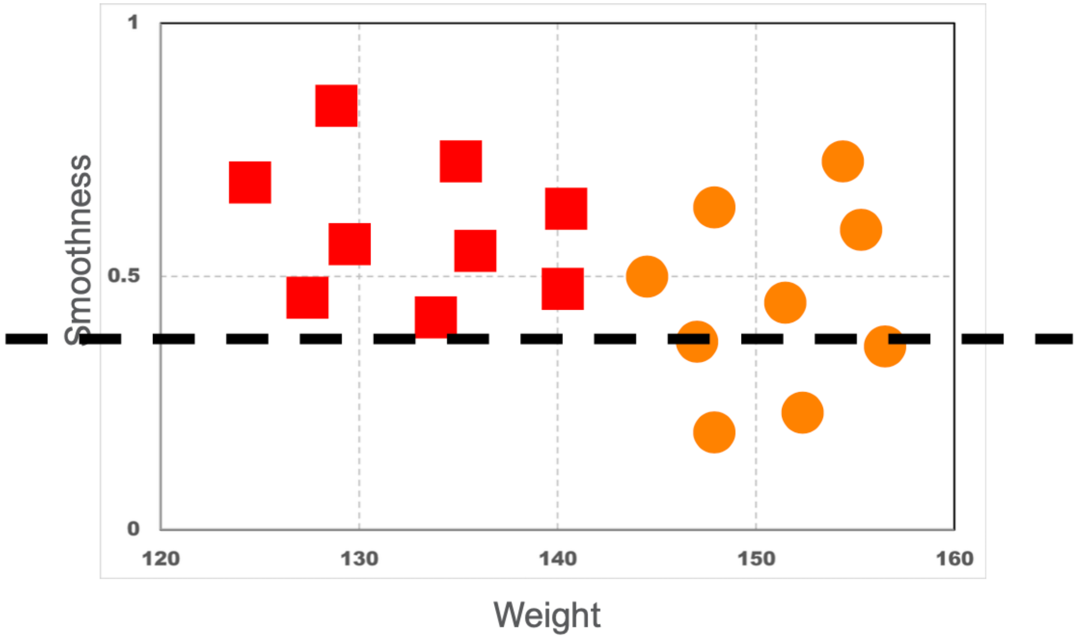
Choosing the *"best"* attribute

- **Key problem:** choosing which attribute to split a given set of examples
- Some possibilities are:
  - Random: Select any attribute at random
  - Least-Values: Choose the attribute with the smallest number of possible values
  - Most-Values: Choose the attribute with the largest number of possible values
  - ***Max-Gain: Choose the attribute that has the largest expected information gain***
    - ***i.e., the attribute that results in the smallest expected size of the subtrees rooted at its child nodes***

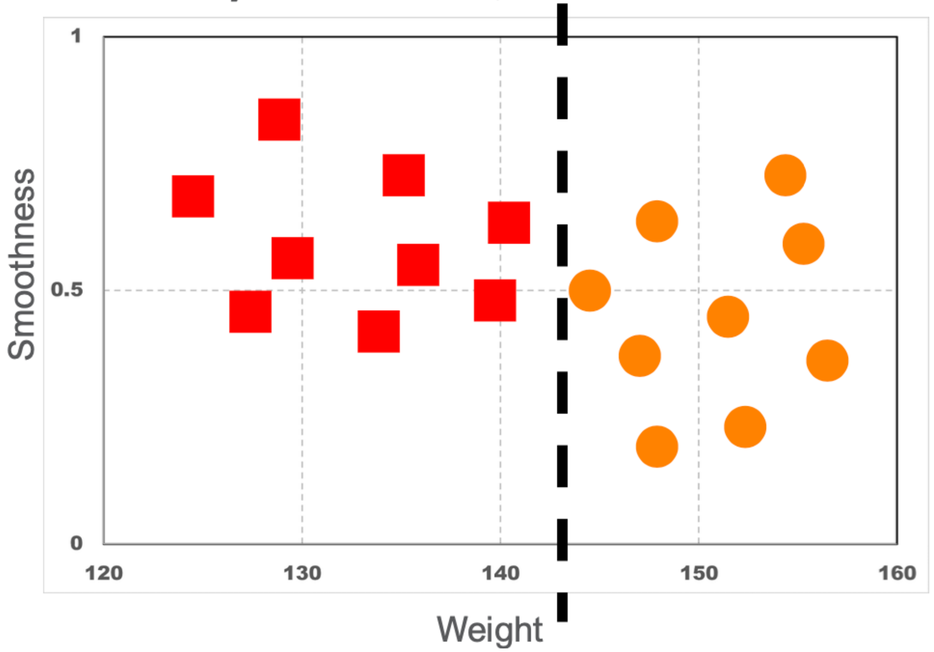
Slide Credit: Dr. Schumman

Information Gain

A) Split on smoothness of 0.4



B) Split on weight of 143



Slide Credit: Dr. Schumman

Information Gain

$$IG(D_p, V) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

V: Feature to split

D_p: dataset of parent node

D_j: dataset of child node j

I: Impurity measurement

N_p: Number of training examples for parent node

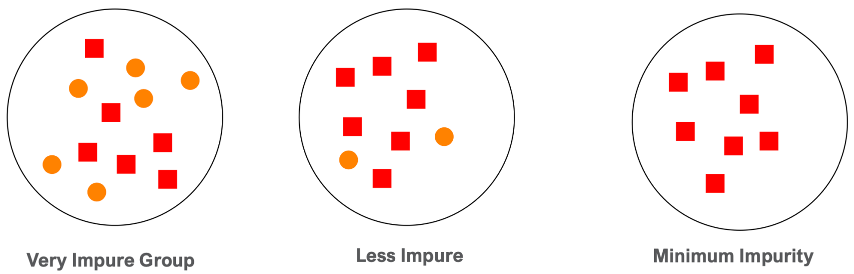
N_j: Number of training examples for child node j

m: Number of child nodes

Information Gain

If we have a ***delta*** between the parent node impurity and the child nodes cumulative impurity, we ***gain information***.

$$IG(D_p, V) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$



V: Feature to split

D_p: dataset of parent node

D_j: dataset of child node j

I: Impurity measurement

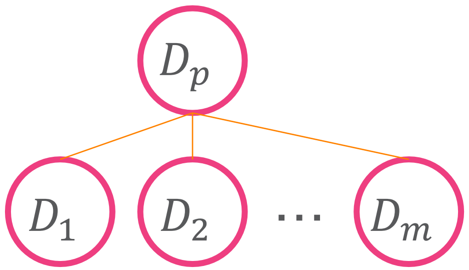
N_p: Number of training examples for parent node

N_j: Number of training examples for child node j

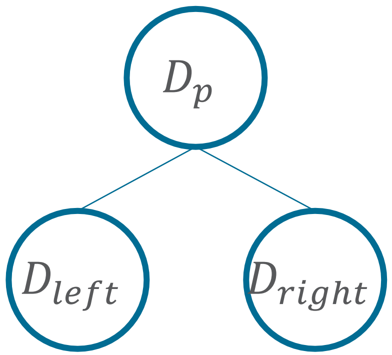
m: Number of child nodes

Information Gain: Binary Tree

$$IG(D_p, V) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

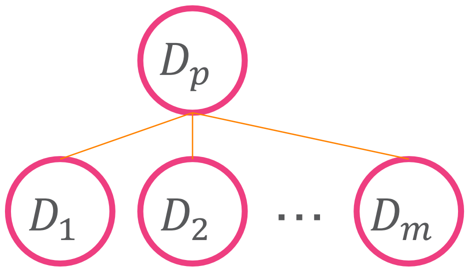


$$IG(D_p, V) = I(D_p) - \frac{N_{Left}}{N_p} I(D_{Left}) - \frac{N_{Right}}{N_p} I(D_{Right})$$



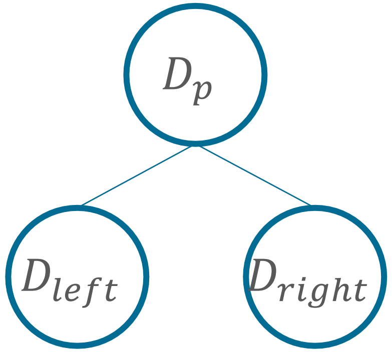
Information Gain: Binary Tree

$$IG(D_p, V) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$



$$IG(D_p, V) = I(D_p) - \frac{N_{Left}}{N_p} I(D_{Left}) - \frac{N_{Right}}{N_p} I(D_{Right})$$

Let's define the impurity metric $I(\cdot)$ to obtain some intuition about Information Gain.

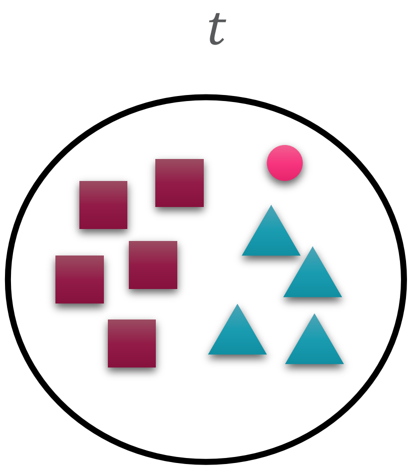


Impurity Metrics

- Entropy (I_H):
  - Attempts to maximize mutual information.
  - How much knowledge about y we gain from knowing split D_j?
- Gini (I_G):
  - Minimizes the probability of misclassification
  - Produces very similar results to Entropy.
- Classification Error (I_E):
  - Less sensitive to changes in the node class distribution
  - Useful when pruning the tree

$p(D = i|t)$

It is the proportion of the samples D in node t that belong to class i.



$$p(D = Square|t) = \frac{5}{10} = 0.5$$

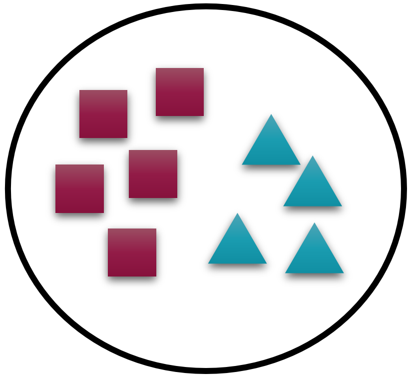
$$p(D = Triangle|t) = \frac{4}{10} = 0.4$$

$$p(D = Circle|t) = \frac{1}{10} = 0.1$$

$p(D = i|t)$ It is the proportion of the samples D in node t that belong to class i.

Binary Node:

t



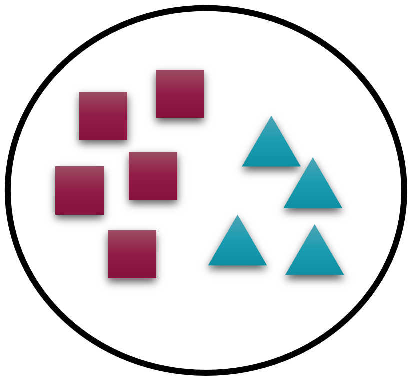
$$p(D = Square|t) = \frac{5}{9} = 0.56 \quad = 1 - p(D = Triangle|t)$$

$$p(D = Triangle|t) = \frac{4}{9} = 0.44 \quad = 1 - p(D = Square|t)$$

$p(D = i|t)$ It is the proportion of the samples D in node t that belong to class i.

Binary Node:

t



$$p(D = \overset{1}{\cancel{Square}}|t) = \frac{5}{9} = 0.56 \quad = 1 - p(D = \overset{0}{\cancel{Triangle}}|t)$$

$$p(D = \overset{0}{\cancel{Triangle}}|t) = \frac{4}{9} = 0.44 \quad = 1 - p(D = \overset{1}{\cancel{Square}}|t)$$

$$p = p(D = 1|t) \Rightarrow p(D = 0|t) = 1 - p$$

Entropy (I_H) - Shannon

- From information theory—the higher the entropy the more information.

$$I_H(D,t) = -\sum_{i=1}^{c} p(D=i|t)\log_2\big(p(D=i|t)\big)$$

$p(D=i|t)$: Proportion of the samples D in node t that belong to class i.

Binary node:

$$I_H(D,t) = -p(D=1|t)\log_2\big(p(D=1|t)\big) - p(D=0|t)\log_2\big(p(D=0|t)\big)$$

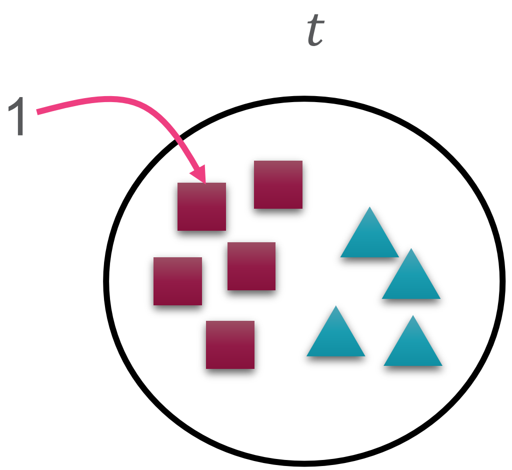
$$= -p\log_2(p) - (1-p)\log_2(1-p)$$

Entropy (I_H) - Shannon

- From information theory—the higher the entropy the more information.

$$I_H = -p\log_2(p) - (1-p)\log_2(1-p)$$

p: Proportion of the samples that belong to the positive (1) class.



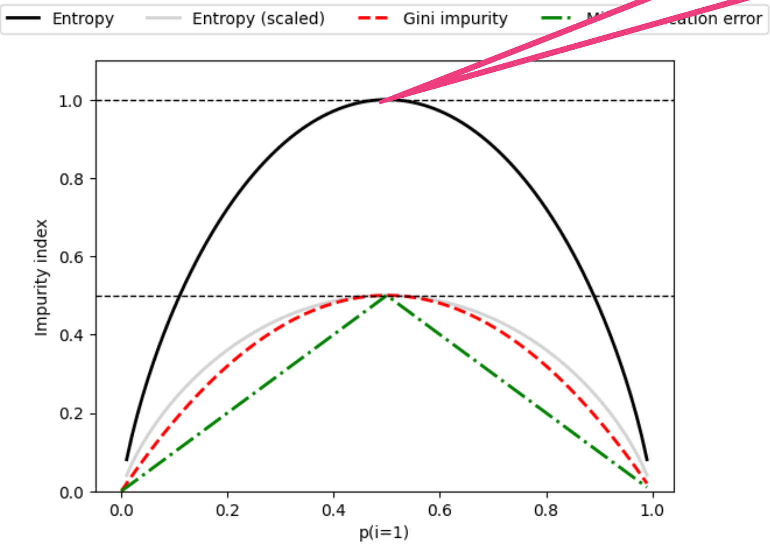
$$p = \frac{5}{9} = 0.56$$

$$I_H = -(0.56)\log_2(0.56) - (1-0.56)\log_2(1-0.56)$$
$$= 0.468 + 0.521$$
$$= 0.99$$

Entropy (I_H) - Shannon

- From information theory—the higher the entropy the more information.

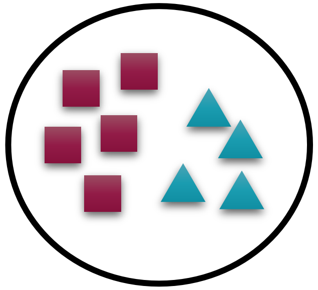
Very close to highest entropy value.



$$I_H = -p\log_2(p) - (1-p)\log_2(1-p)$$

p: Proportion of the samples in node t that belong to the positive (1) class.

t



$$p = \frac{5}{9} = 0.56$$

$$I_H = -(0.56)\log_2(0.56) - (1-0.56)\log_2(1-0.56)$$
$$= 0.468 + 0.521$$
$$= \mathbf{0.99}$$

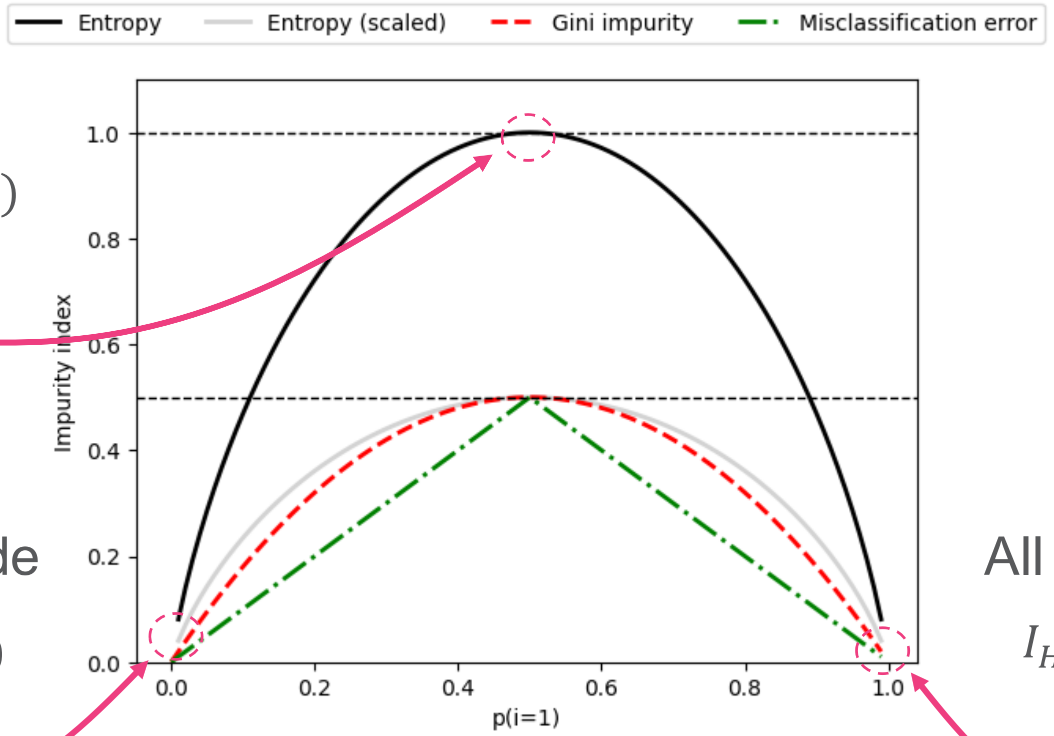
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Entropy (I_H) - Shannon

$$I_H = -p\log_2(p) - (1-p)\log_2(1-p)$$

An equal number of samples for each category

$$I_H = -0.5\cdot(-1) - (1-0.5)\cdot(-1) = 1$$



All negative samples in the node

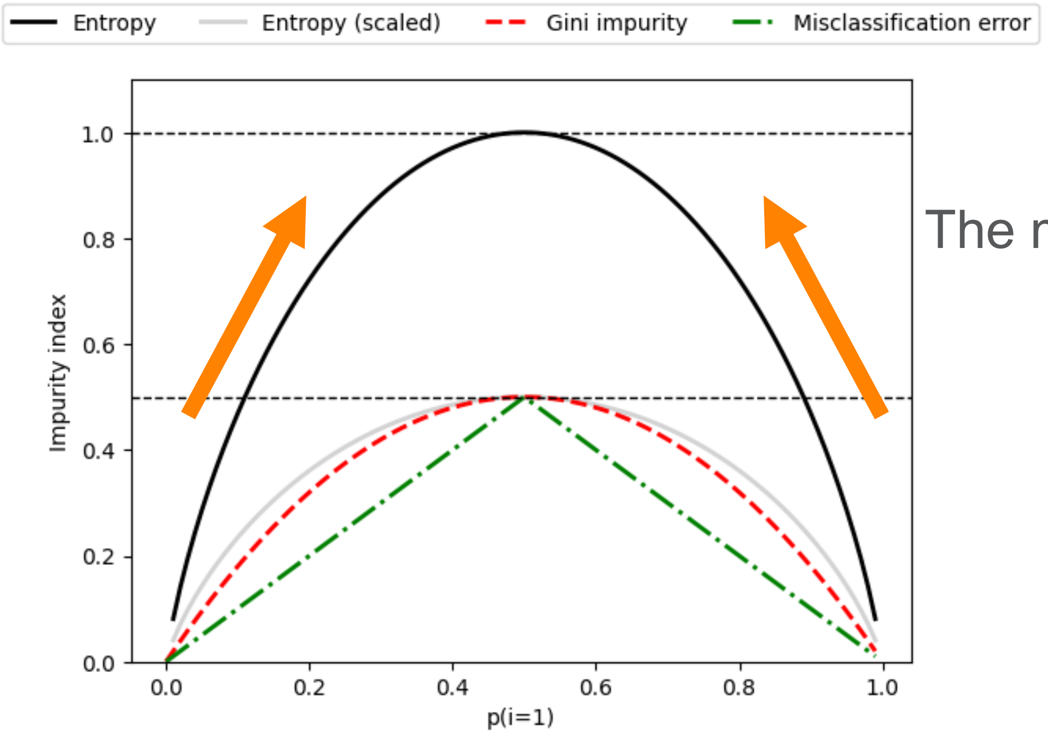
$$I_H = -0\cdot(\infty) - (1-0)\cdot(0) = 0$$

All positive samples in the node

$$I_H = -1\cdot(0) - (1-1)\cdot(\infty) = 0$$

Entropy (I_H) - Shannon

$$I_H = -p\log_2(p) - (1-p)\log_2(1-p)$$



The more mixed the data in a node

Impurity Increases

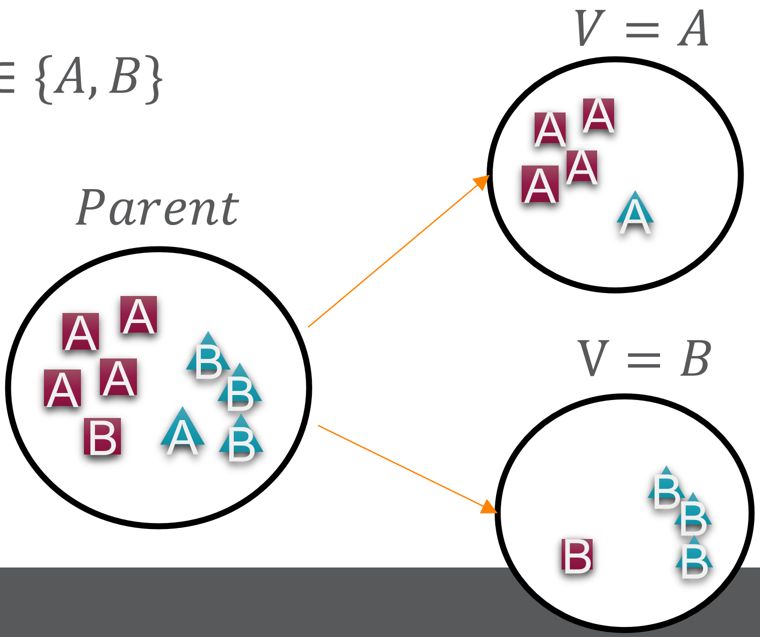
Entropy Increases

Specific Conditional Entropy

Now $t = (V = v)$: The samples in the node meet certain criteria. E.g., $x_2 \leq 2.5$

$$I_H(D|V = v) = -\sum_{i=1}^{m} p(D = i|V = v)\log_2(p(D = i|V = v)) = -p_v\log_2(p_v) - (1 - p_v)\log_2(1 - p_v)$$

$V = x_j \in \{A, B\}$



$$I_H(D|V = A) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) = 0.722$$

$$I_H(D|V = B) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) = 0.81$$

Conditional Entropy

$$I_H(D|V) = \sum_{v \in V} p(V = v) I_H(D|V = v)$$

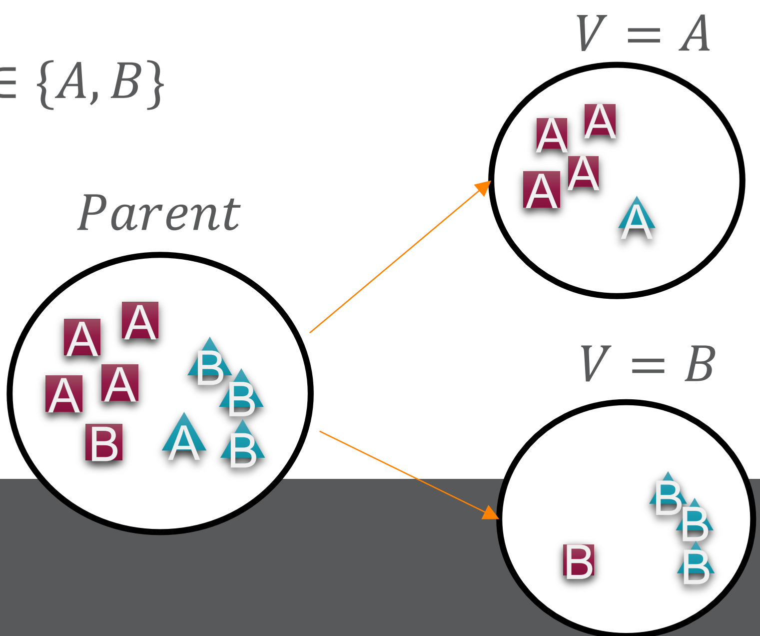
$\frac{N_j}{Np}$

Computed in the previous slide.

$$I_H(D|V = A) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) = 0.72$$

$$I_H(D|V = B) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) = 0.81$$

$V = x_j \in \{A, B\}$



Computed on the parent node.

$p(V = A) = 5/9$

$p(V = B) = 4/9$

$$I_H(D|V) = \frac{5}{9}(0.72) + \frac{4}{9}(0.81)$$
$$= \mathbf{0.76}$$

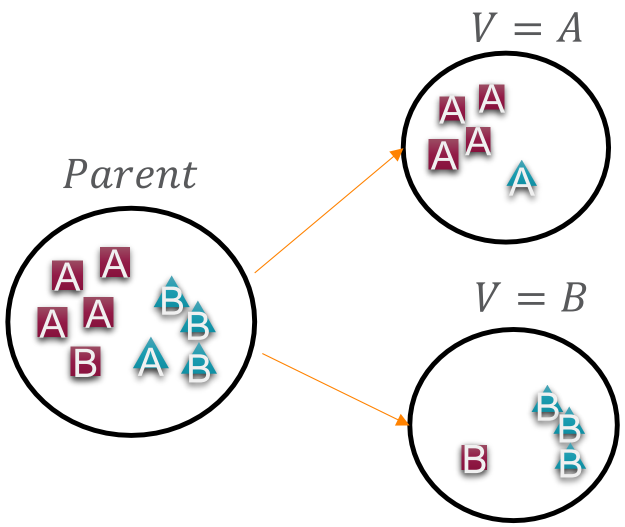
Mutual Information

Mutual Information (I) is the amount of information that one random variable Y contains about another random variable X.

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$H(X,Y) = H(Y) + H(X|Y) \Rightarrow$$

$$\boldsymbol{I(X,Y) = H(X) - H(X|Y)}$$



$$I_H(D,V) = I_H(D) - I_H(D|V) = I_H(D) - \sum_{v \in V} p(V=v) I_H(D|V=v) = 0.99 - 0.76 = 0.23$$

This is our information gain.

Information Gain

Info Gain

Feature to split

$p(V = v)$

Entropy for split j

$$IG(D_p, V) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

Dataset Parent Node

Entropy at parent node

Conditional Entropy $I_H(D_p, V)$

D … taset of parent node

D_j: dataset of child node j

I: Impurity measurement

N_p: Number of training examples for parent node

N_j: Number of training examples for child node j

m: Number of child nodes

THE UNIVERSITY OF TENNESSEE KNOXVILLE

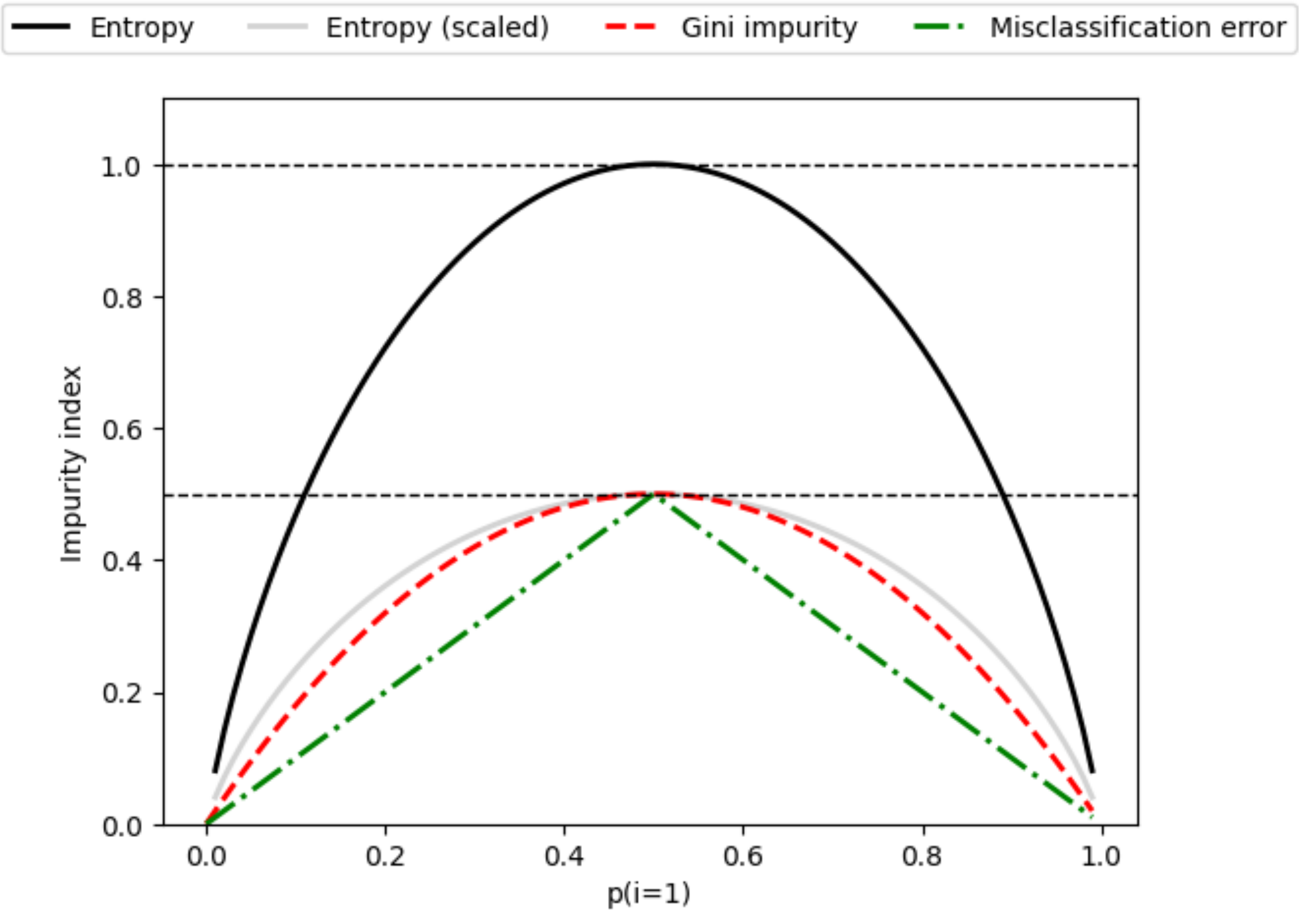
Other Impurity Metrics

- Entropy (I_H)
- **Gini (I_G)**
- Classification Error (I_E)

$$I_G(t) = \sum_{i=1}^{c} p(i|t)\big(1 - p(i|t)\big) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

Binary node:

$$I_G(t) = 1 - p(1|t)^2 - p(0|t)^2$$
$$= -2(p^2 - p)$$



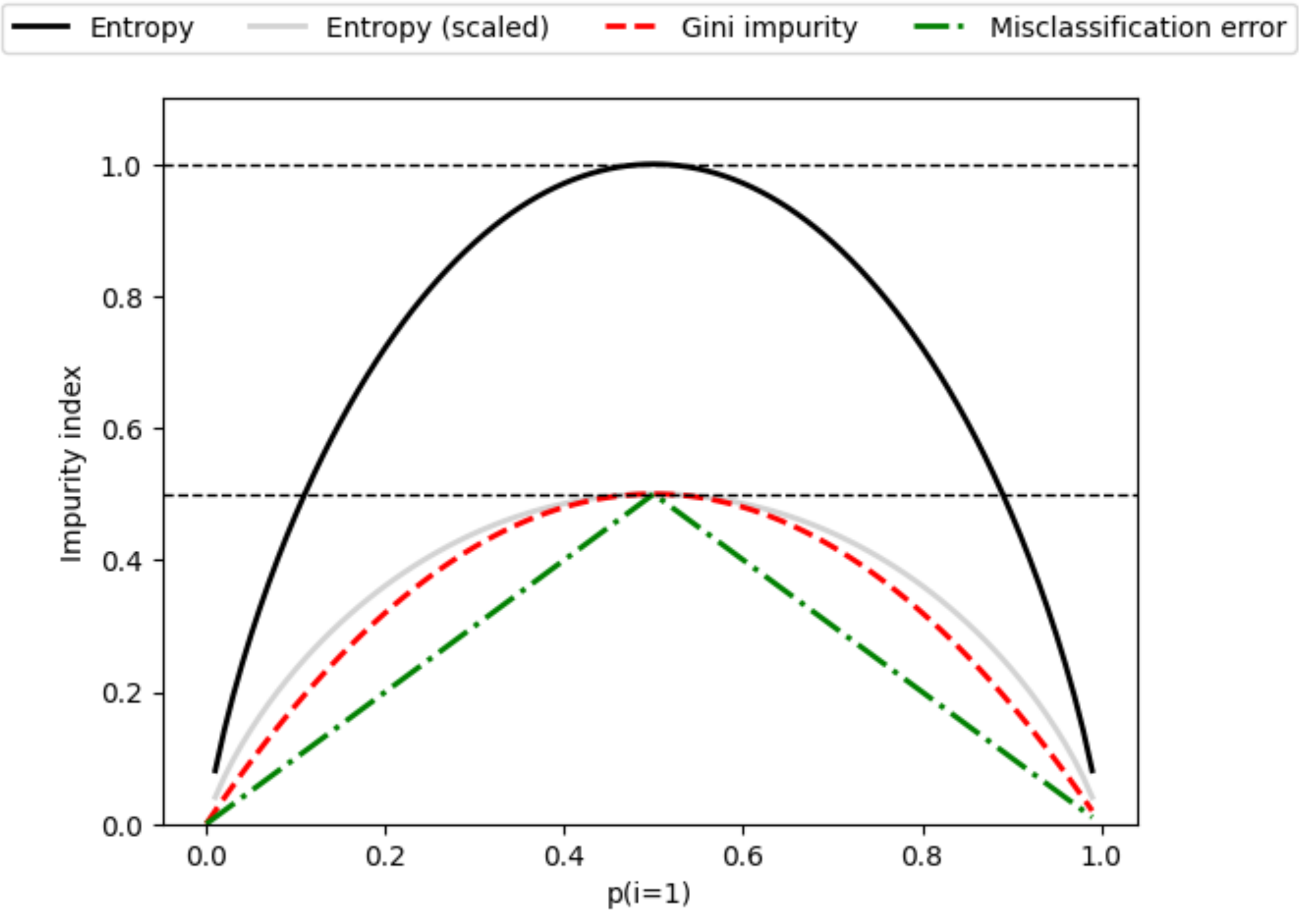
Other Impurity Metrics

- Entropy (I_H)
- Gini (I_G)
- **Classification Error (I_E)**

$$I_E = 1 - \max_{i \in c}\{p(i|t)\}$$

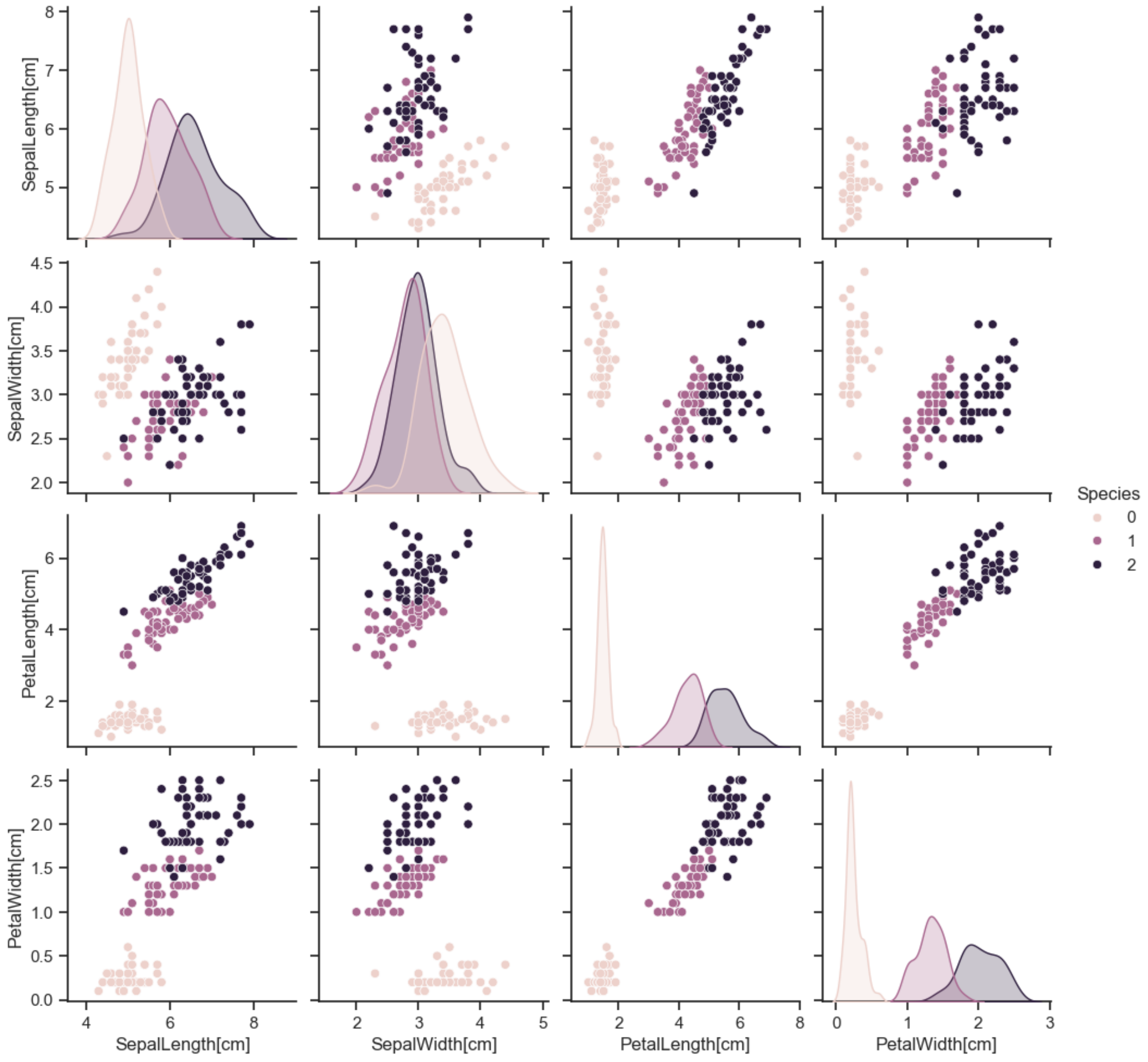
Binary node:

$$I_E = 1 - \max\{p, 1 - p\}$$

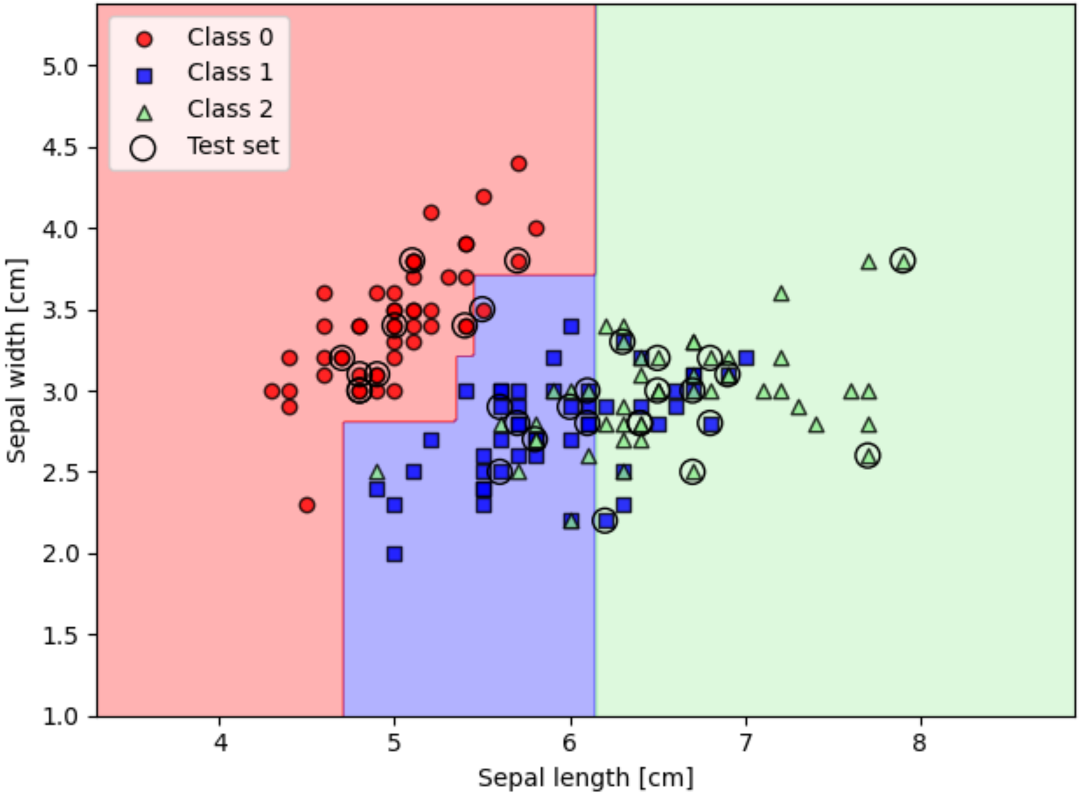


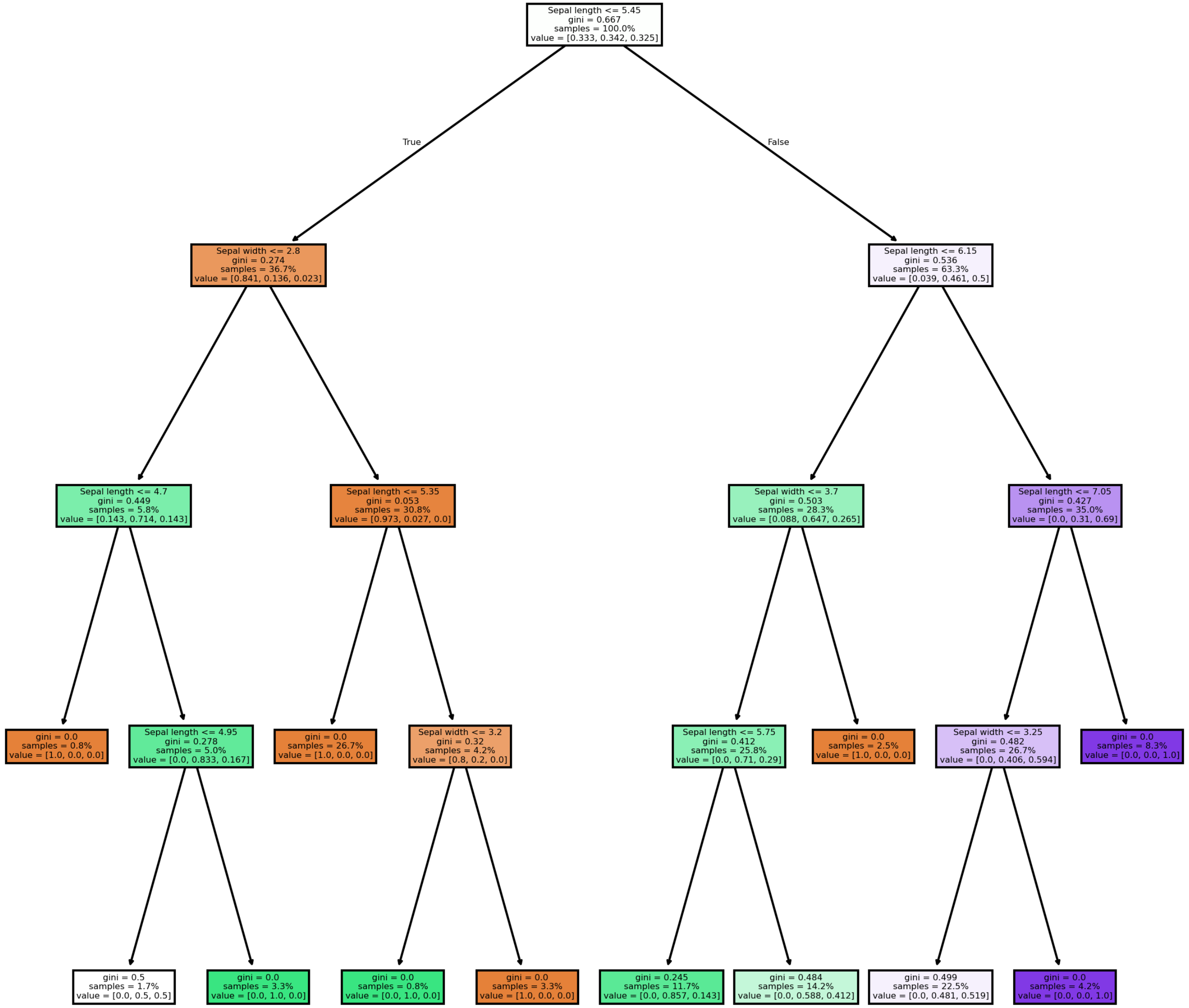
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Demo with Iris dataset

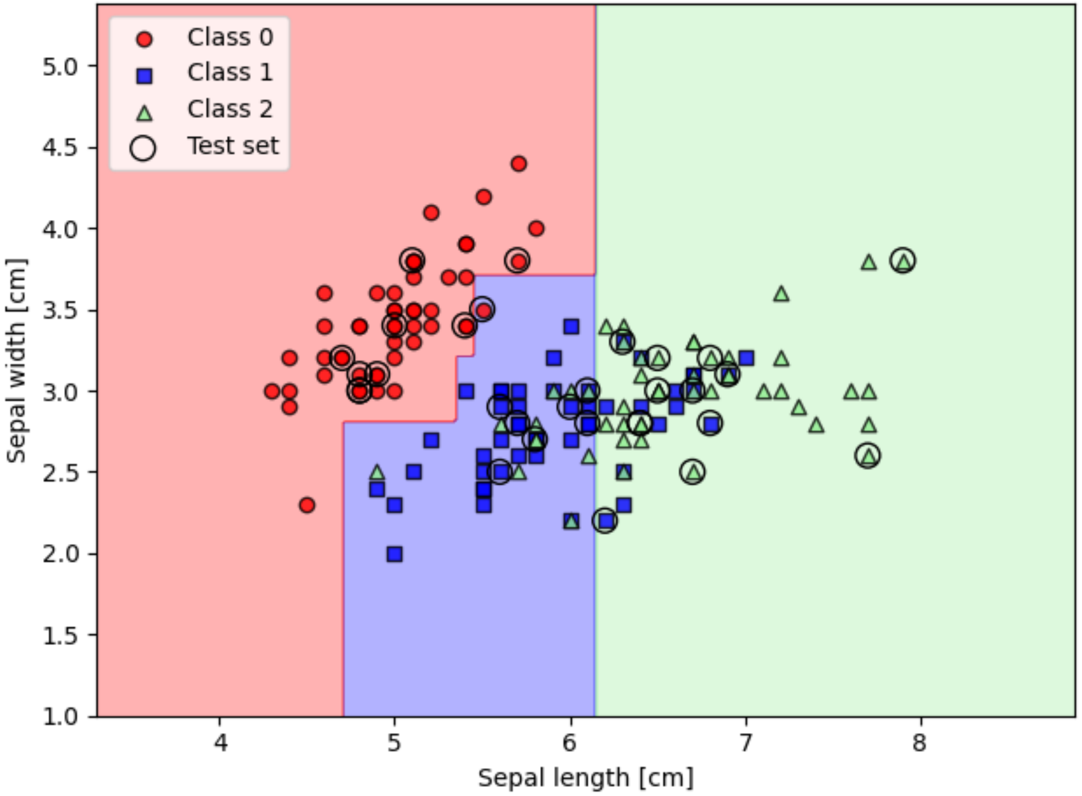


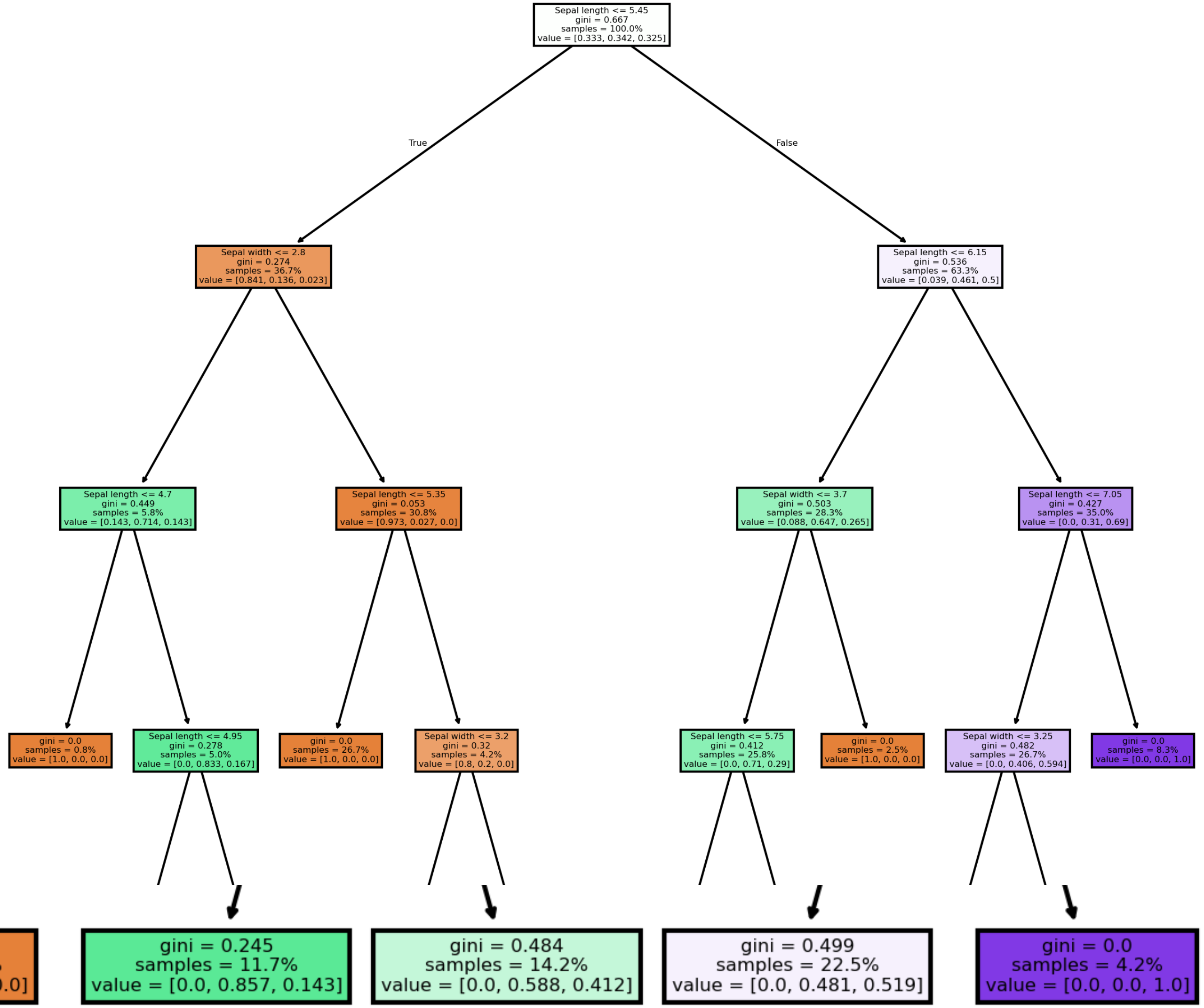
Demo with Iris dataset: Sepal Width and Length



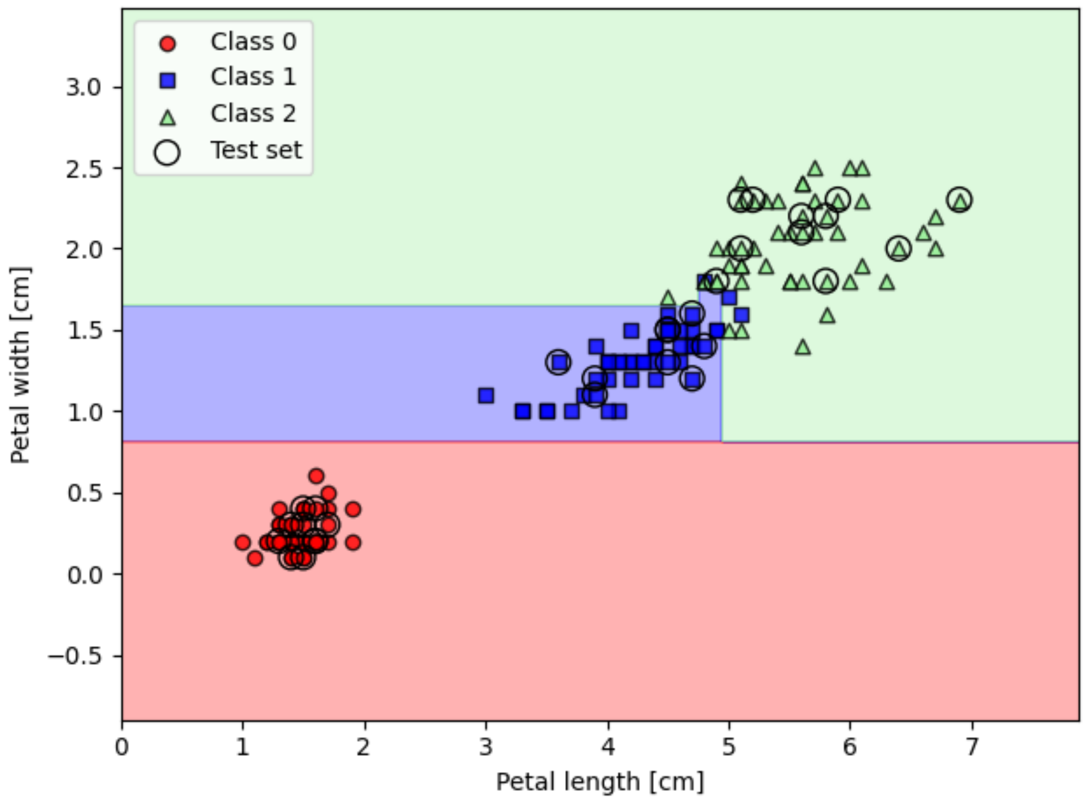


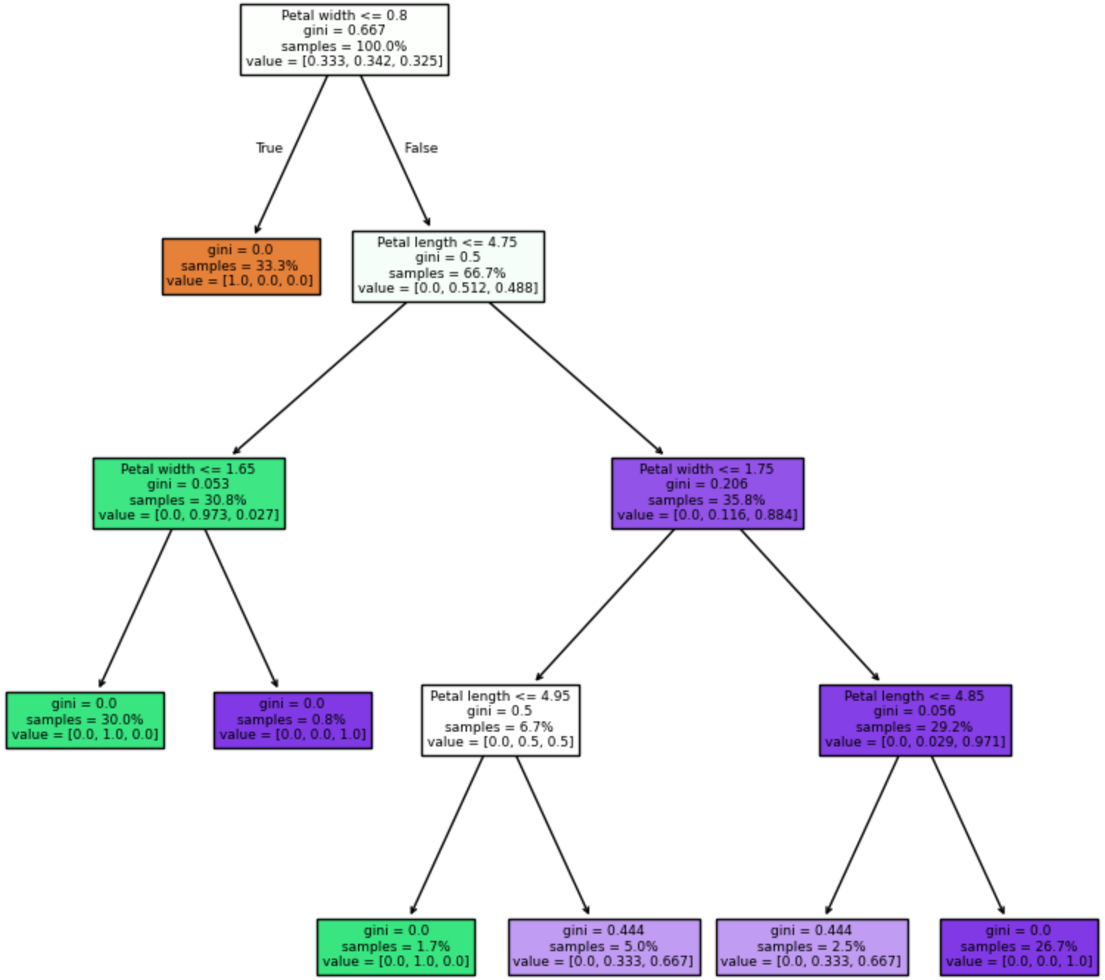
Demo with Iris dataset: Sepal Width and Length



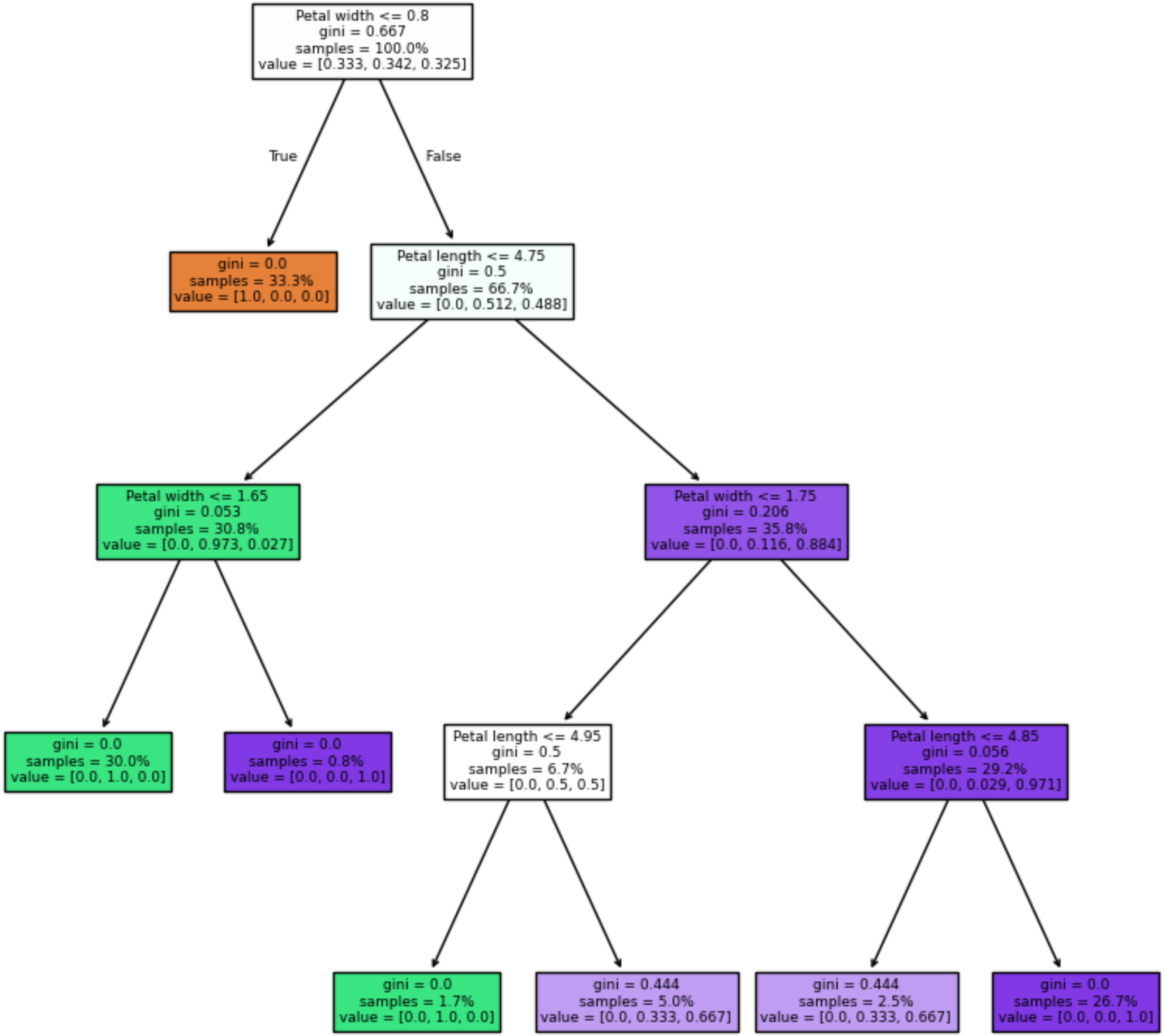


Demo with Iris dataset: Petal Width and Length





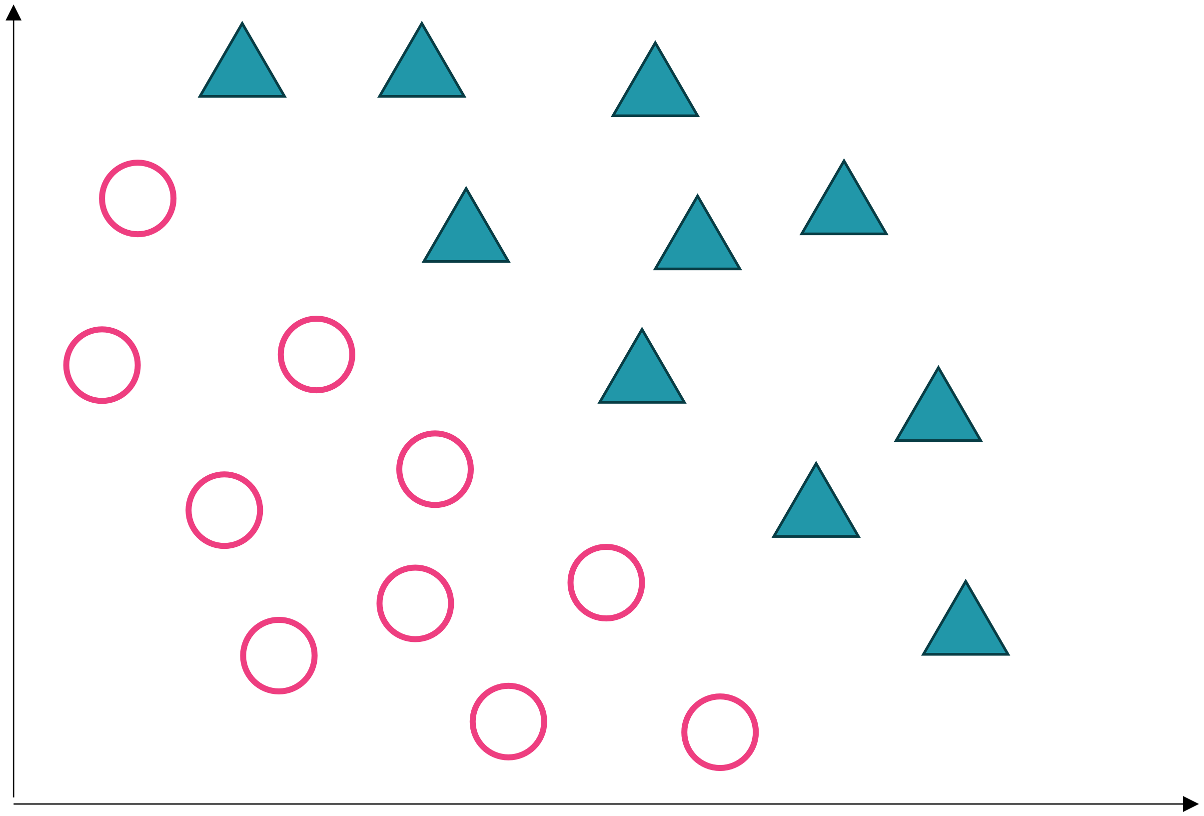
Demo with Iris dataset: All Features



Decision Trees Shortcomings

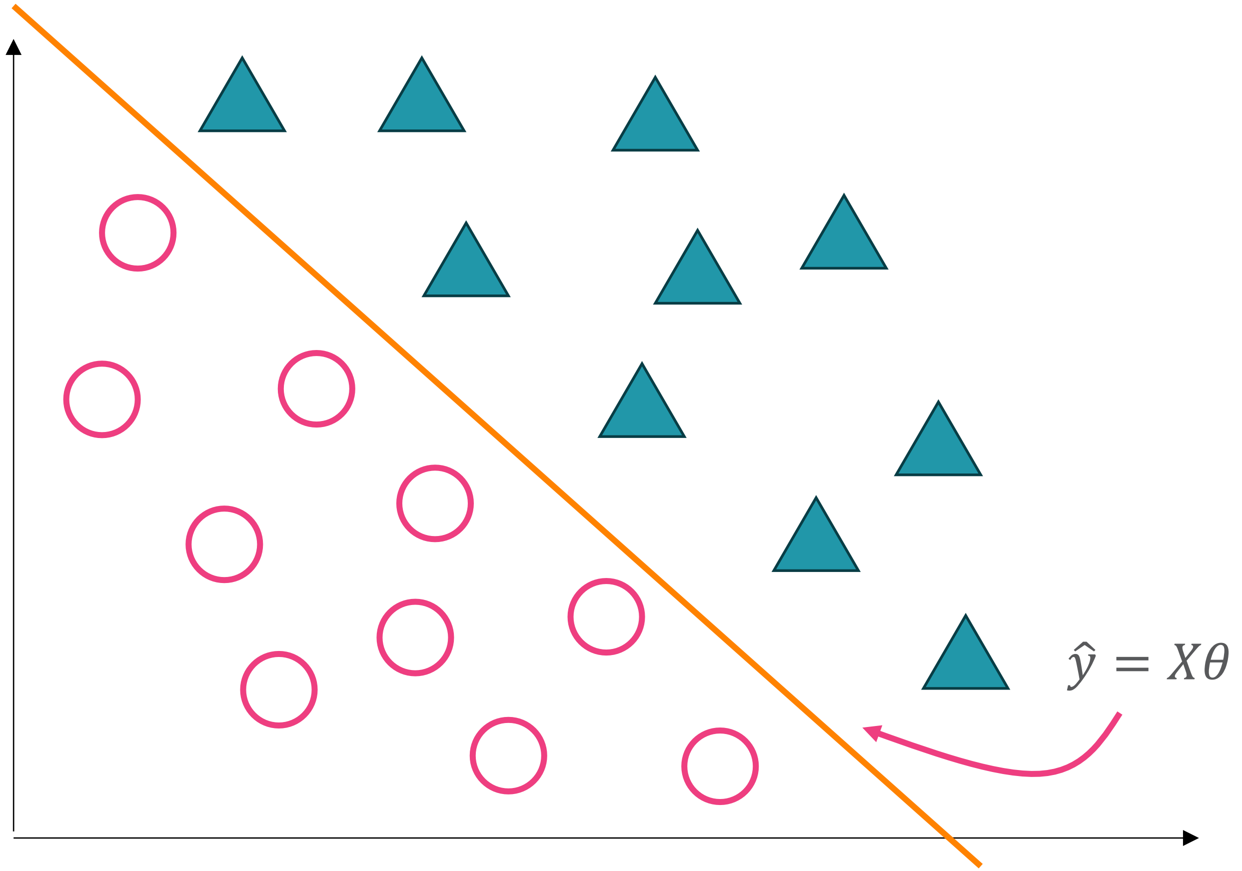
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Diagonal Boundaries



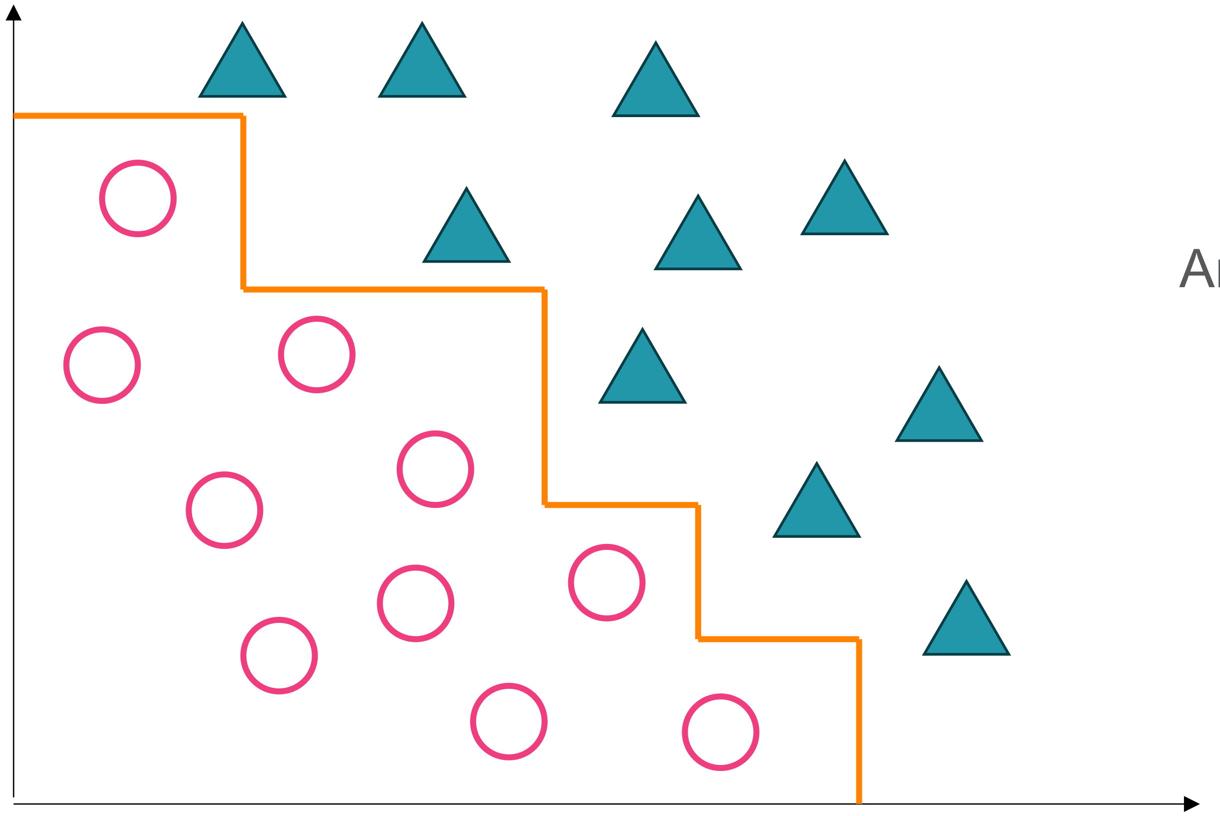
Slide credit: Dr. Raschka

Diagonal Boundaries



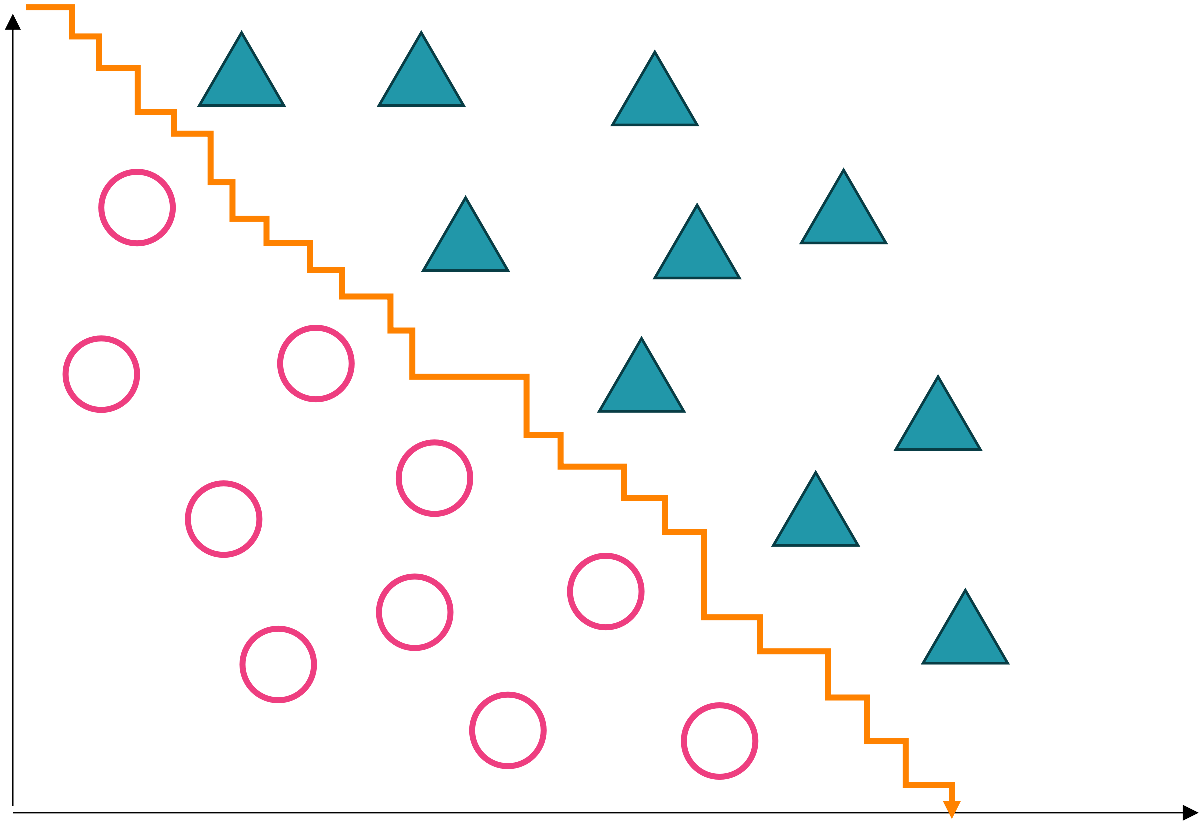
Slide credit: Dr. Raschka

Diagonal Boundaries



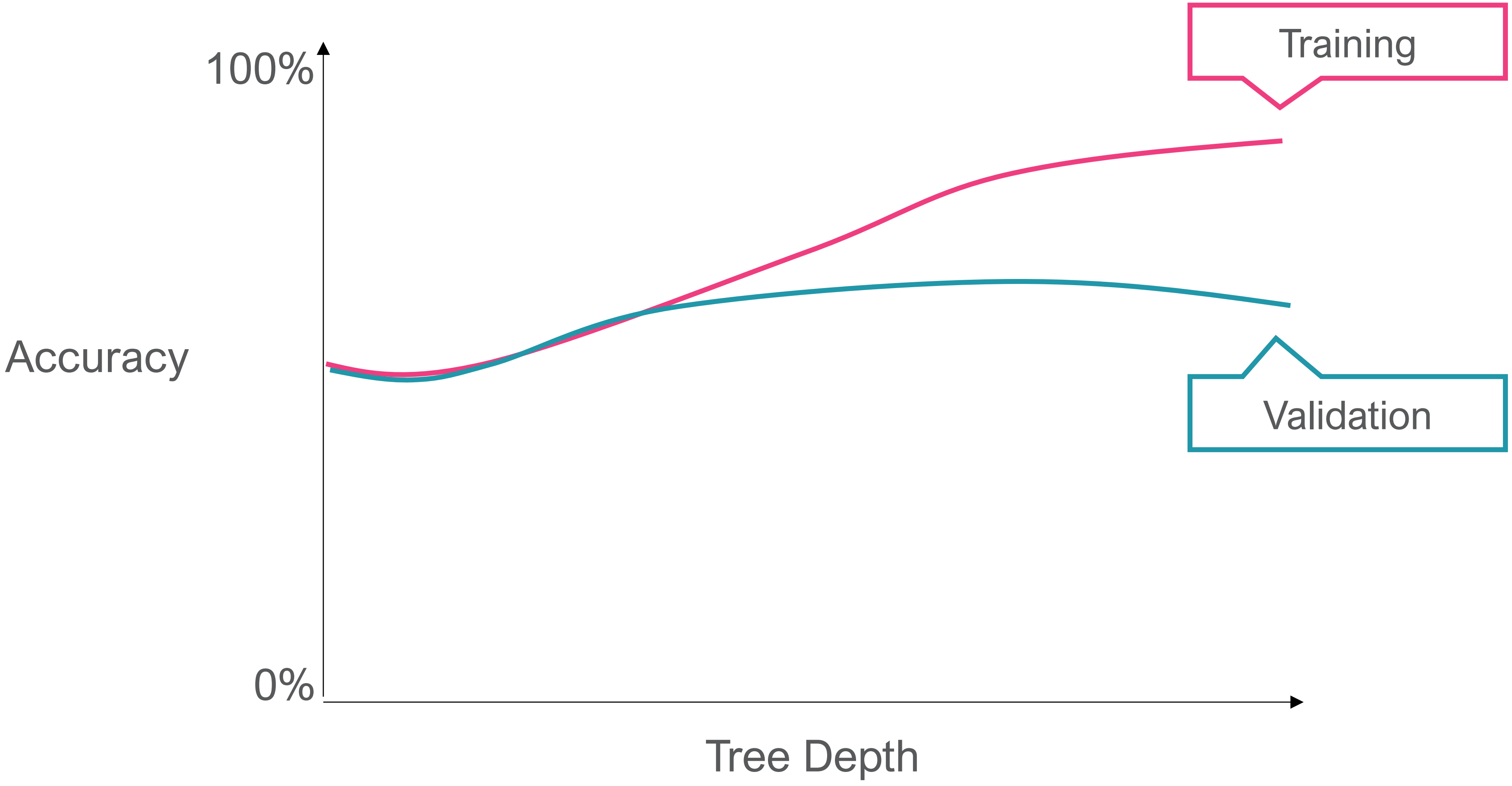
An internal node for each segment.

Diagonal Boundaries



Tree will become too large.

Overfitting



Slide credit: Dr. Raschka

Pop Quiz

Why (when) does the accuracy start at ~50% for a binary decision tree?

A. 100% of samples are from the positive class

B. 100% of samples are from the negative class

C. 50% of samples are from the positive class.

D. 50% of the samples are easy to classify.

Pop Quiz

Why (when) does the accuracy start at ~50% for a binary decision tree?

A. 100% of samples are from the positive class

B. 100% of samples are from the negative class

C. 50% of samples are from the positive class.

D. 50% of the samples are easy to classify.

Gain Ratio

- Addresses wide trees and helps with overfitting
- Penalizes node splits for features with several categories
  - E.g., Date column
- When the number of child nodes is 10x, SplitInfo is 2x

$$GainRatio(\mathcal{D}, V) = \frac{Gain(\mathcal{D}, V)}{SplitInfo(\mathcal{D}, V)}$$

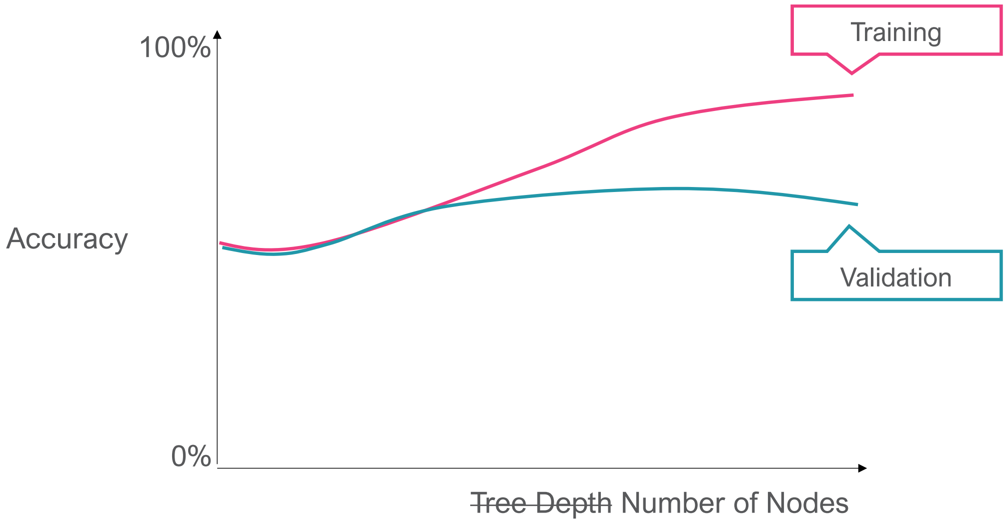
$$SplitInfo(\mathcal{D}, V) = -\sum_{v \in V} \frac{|\mathcal{D}_v|}{|\mathcal{D}|} \log_2 \left(\frac{|\mathcal{D}_v|}{|\mathcal{D}|} \right)$$

Pre-Pruning (Before we grow tree)

- Set a depth cut-off (maximum tree depth)
- Cost-complexity pruning, where we set a total number of nodes.
- Stop growing if split is not statistically significant
  - (e.g., χ^2 test)
- Set a minimum number of data points for each node
  - Addresses labeling errors
- Remove irrelevant attributes

Post-Pruning (After Training)

- Acquire more training data
- Grow full tree first, then remove nodes
- ***Reduced-error pruning:*** remove nodes via validation set evaluation
  - Requires a test set
  - Greedly remove node that most improves validation set accuracy



Regression Trees

- Variance reduction (CART algorithm)

- Given a node t

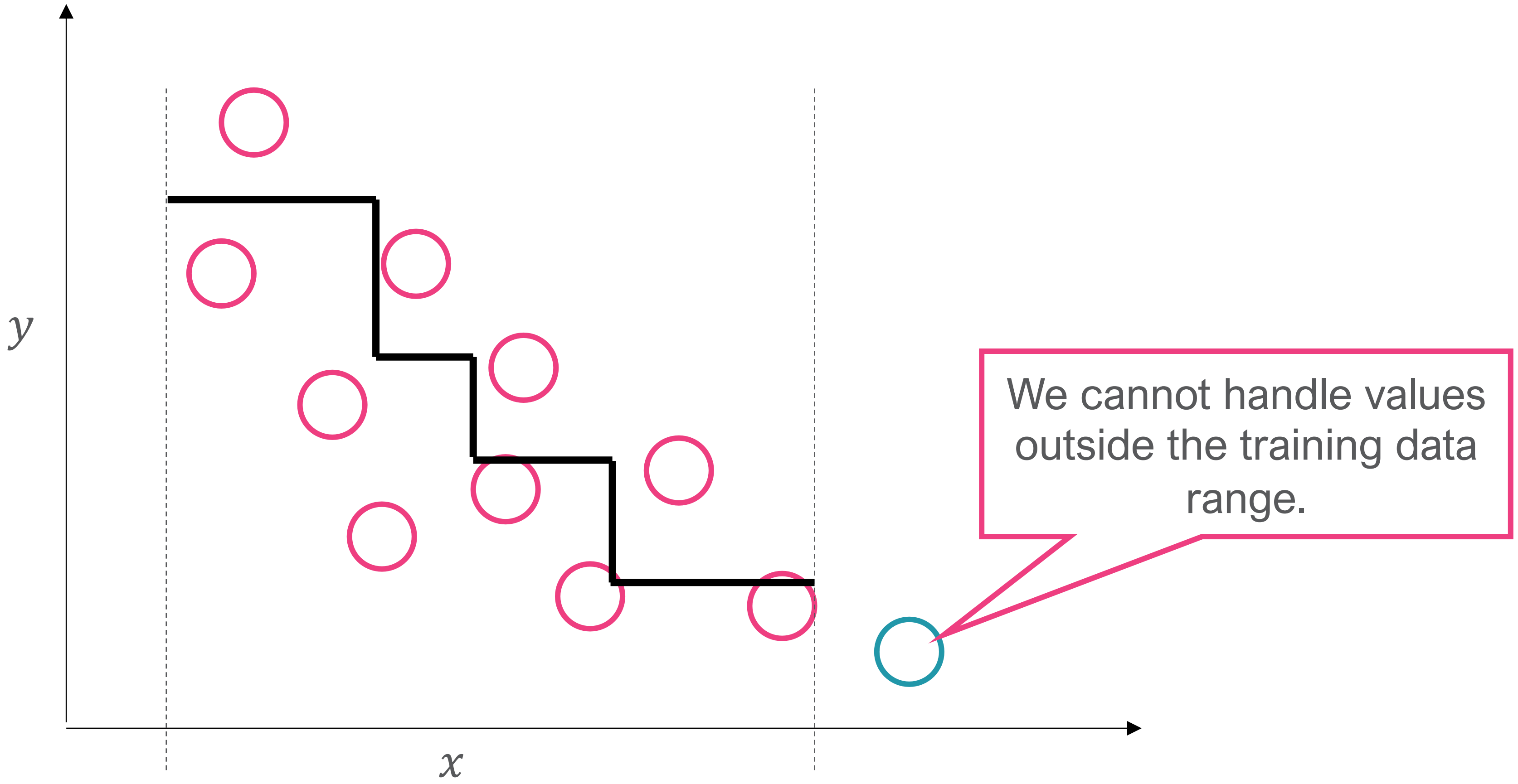
$$MSE(D_t) = \frac{1}{n_t} \sum_{i \in D_t}^{n_t} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

Minimize

$$\hat{y} = \frac{1}{n_t} \sum_{i \in D_t} y^{(i)}$$

Majority Voting

Shortcomings in Tree Regression



ID3 – Iterative Dichotomizer

- Early algorithm proposed by Quinlan, 1986.
- Cannot handle numeric values
- Prone to overfitting (no pruning)
- Produce short and wide trees
- Maximize information gain by minimizing entropy
- Support discrete features, binary and multi-category features

C4.5

- Continuous and discrete features, Quinlan 1993.
- Continuous is very expensive, because must consider all possible ranges
- Handles missing attributes (ignores them in gain compute)
- Post-pruning (bottom-up pruning)
- Gain Ratio stop criteria

CART

- **C**lassification **A**nd **R**egression **T**rees proposed by Breiman 1984.
- Handles continuous and discrete features
- Strictly uses binary splits (taller trees than ID3, C4.5)
- Trees produce better results that ID3 and C4.5 but are harder to interpret
- Tree growth
  - Variance reduction in regression trees
  - Gini impurity, also known as twoing.
- Cost complexity pruning

Review

- (+) Easy to interpret and communicate
- (+) Can represent "complete" hypothesis space
- (-) Easy to overfit
- (-) Elaborate pruning required
- (-) Expensive to just fit a "diagonal line"
- (-) Output range is bounded in regression trees by input range.

Next Lecture

- Ensemble techniques