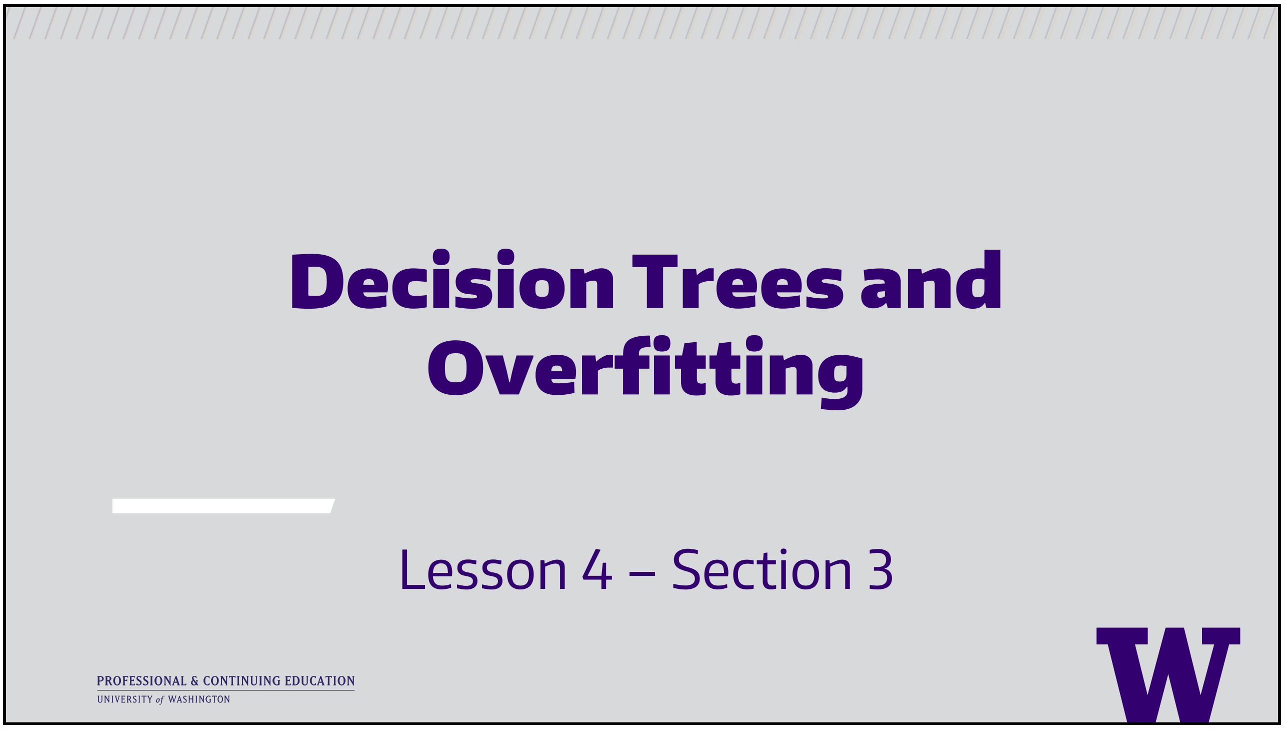
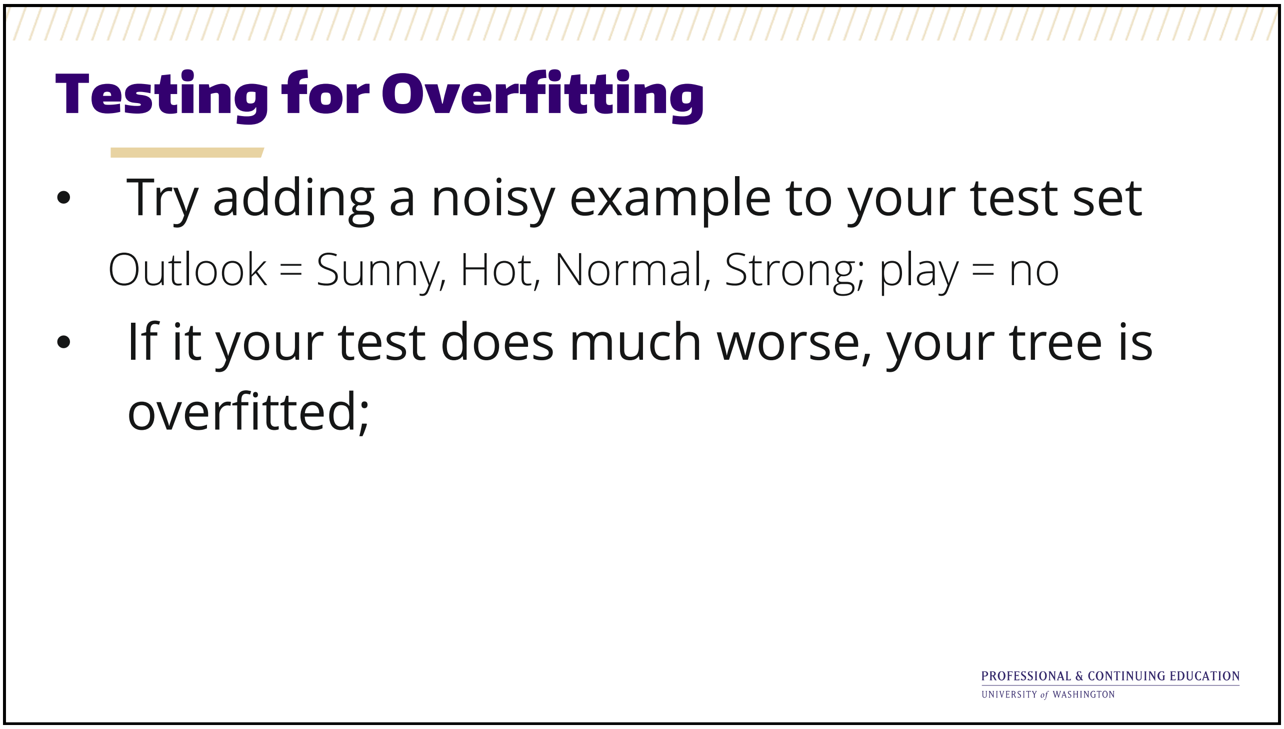


Decision Trees and Overfitting

Lesson 4 – Section 3

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Testing for Overfitting

- Try adding a noisy example to your test set

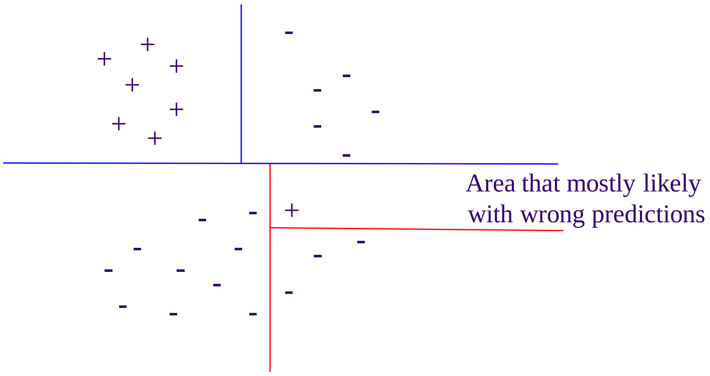
  Outlook = Sunny, Hot, Normal, Strong; play = no

- If it your test does much worse, your tree is overfitted;

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Reasons for Overfitting:

A small number of instances are associated with leaf nodes: In this case it is possible that for coincidental regularities to occur that are unrelated to the actual borders



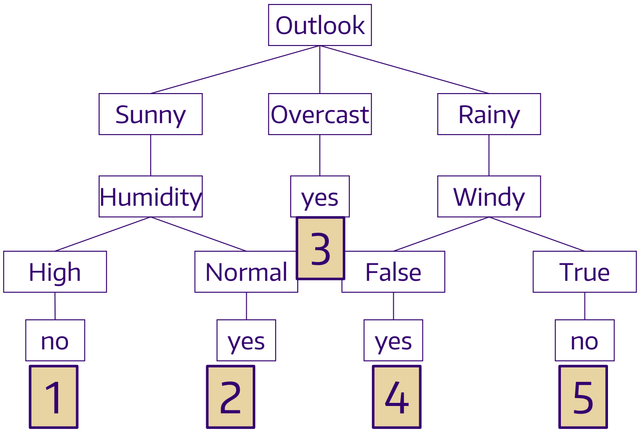
Area that mostly likely with wrong predictions

Approaches to Avoid Overfitting:

- **Pre-pruning:** stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data
- **Post-pruning:** Allow the tree to overfit the data, and then post-prune the tree.

Pre-Pruning:

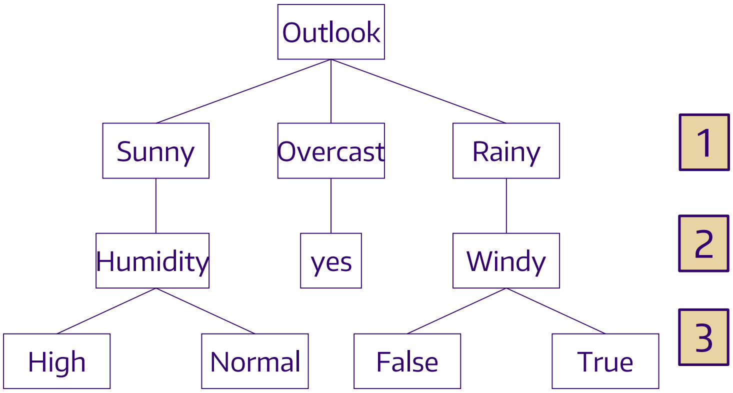
- It is challenging to determine when to stop growing the tree
- One thing to try is limited the number of leaf nodes get less than m training instances.
- max_leaf_nodes= 5



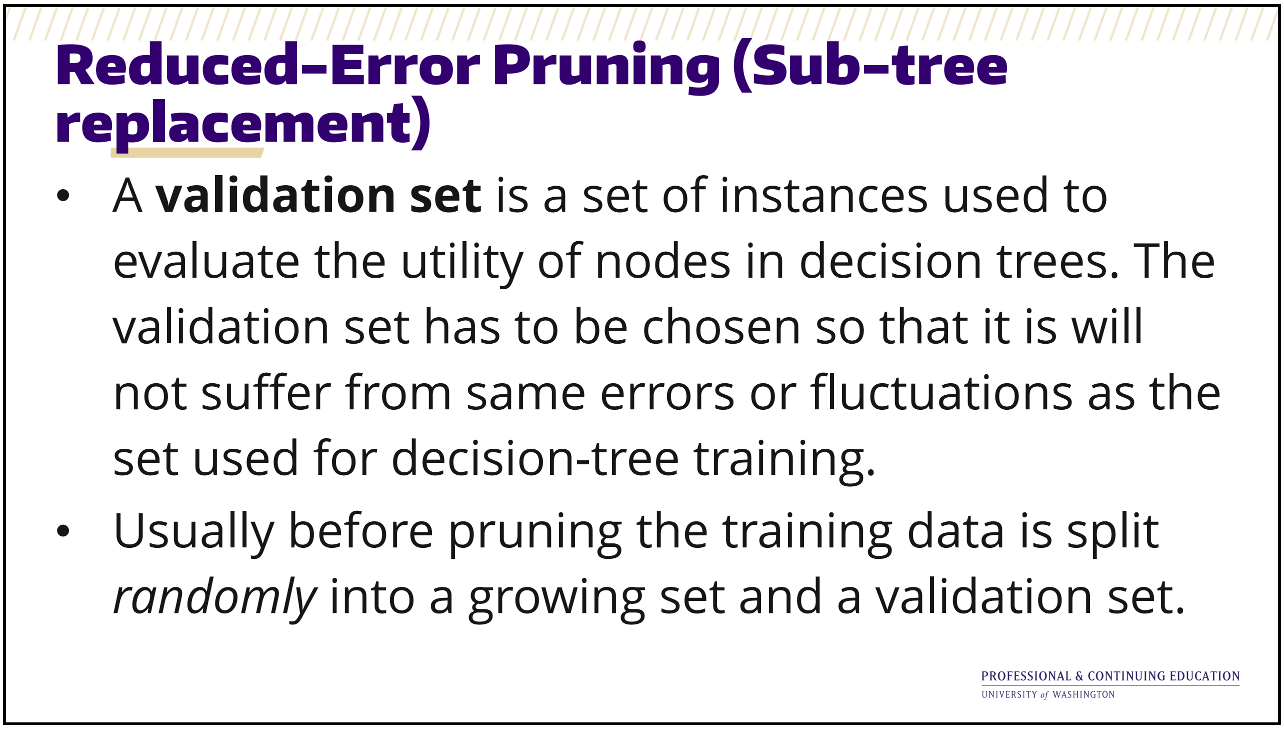
max_leaf_nodes : int or None, optional (default=None)

Pre-Pruning:

- You can also limit the depth (number of decision levels) of the tree
- E. g., max_depth = *3*



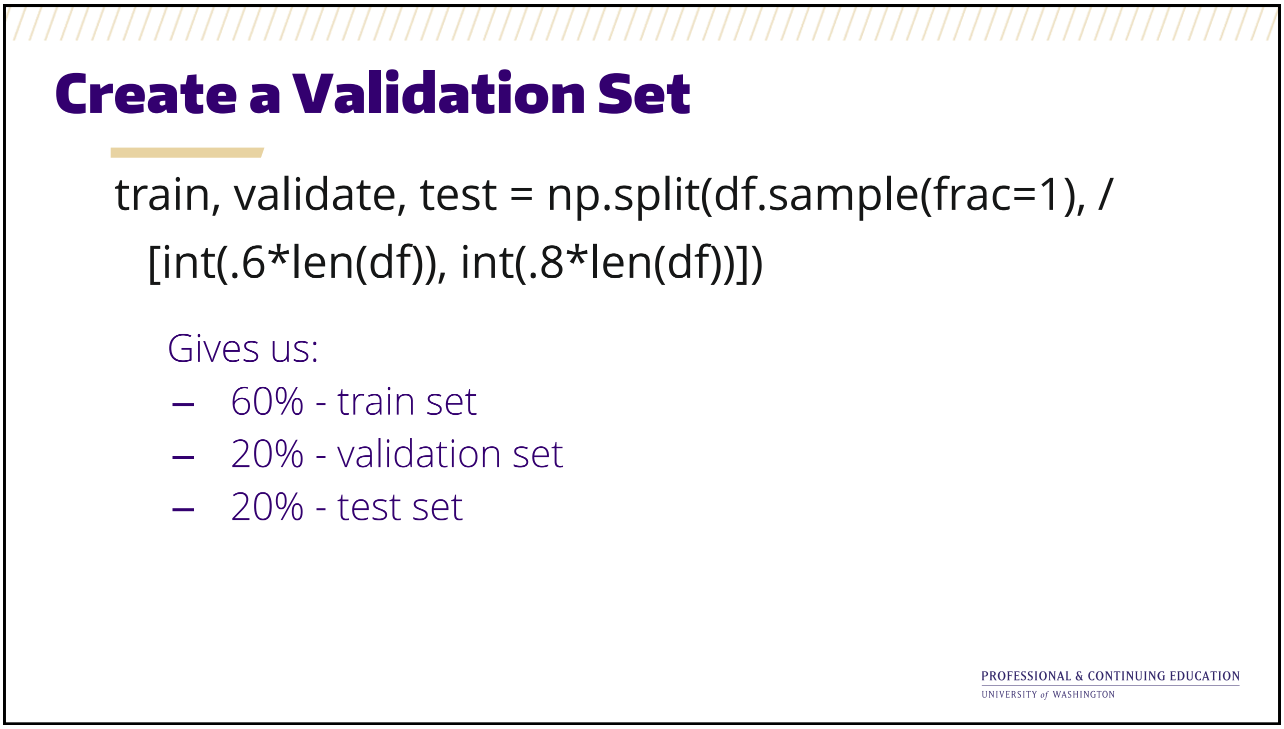
max_depth : int or None, optional (default=None)



Reduced-Error Pruning (Sub-tree replacement)

- A **validation set** is a set of instances used to evaluate the utility of nodes in decision trees. The validation set has to be chosen so that it is will not suffer from same errors or fluctuations as the set used for decision-tree training.
- Usually before pruning the training data is split *randomly* into a growing set and a validation set.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Create a Validation Set

```
train, validate, test = np.split(df.sample(frac=1), /
  [int(.6*len(df)), int(.8*len(df))])
```

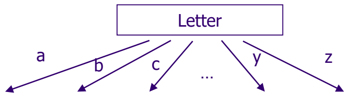
Gives us:

- 60% - train set
- 20% - validation set
- 20% - test set

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

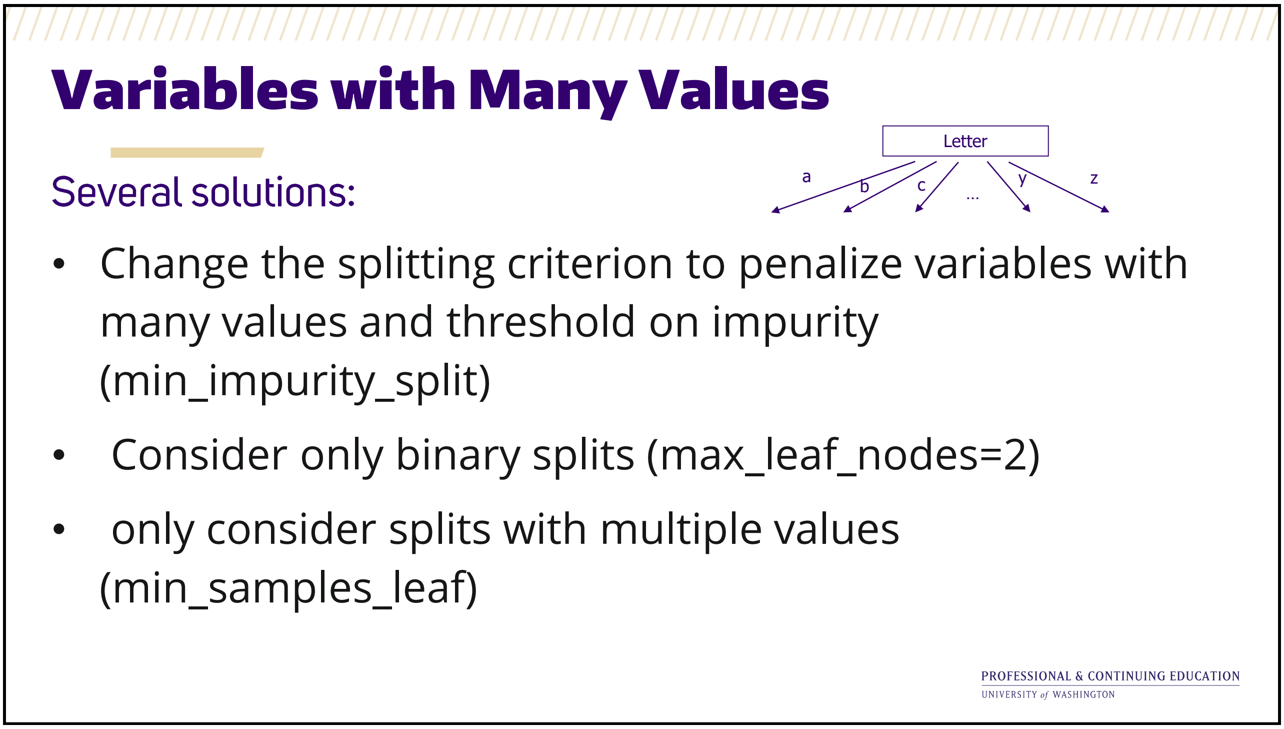
Data Problems for Decision Trees

Variables with Many Values



Problem

- Not good splits: they fragment the data too quickly, leaving insufficient data at the next level
- The reduction of impurity of such test is often high (example: split on the object id)



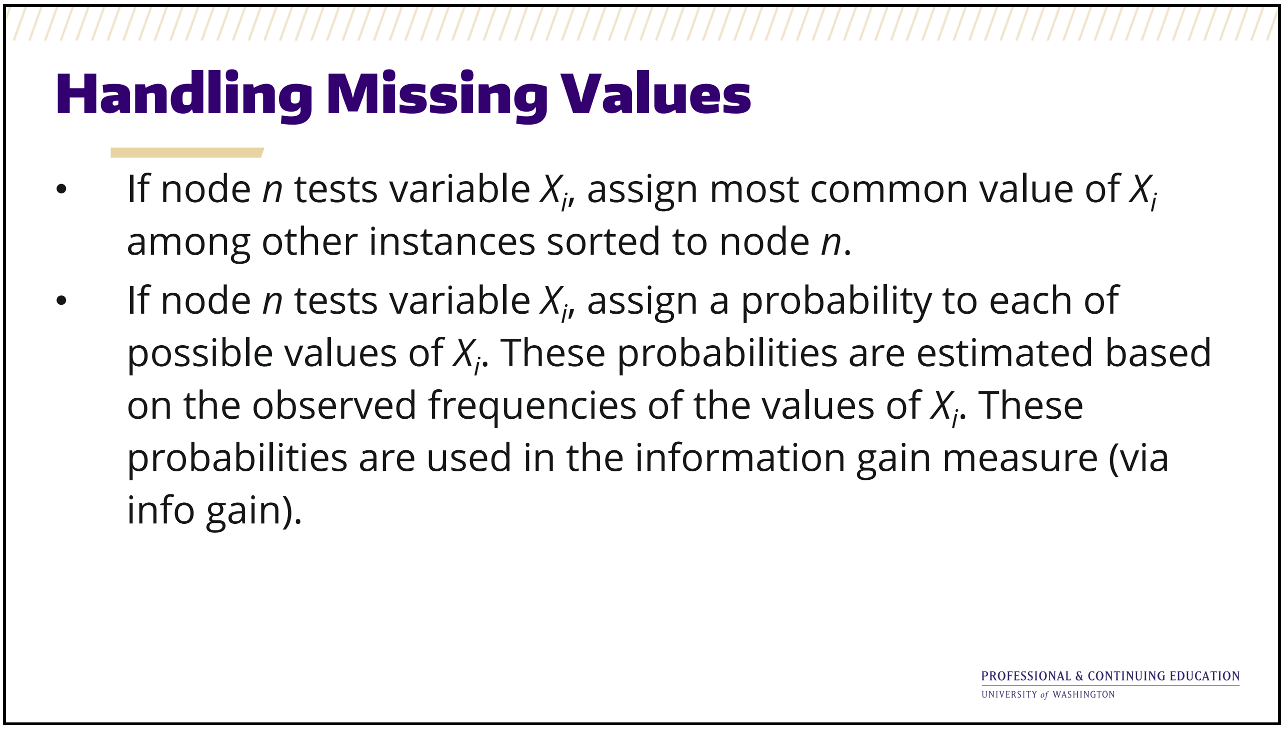
Letter
a
b
c
...
y
z

Variables with Many Values

Several solutions:

- Change the splitting criterion to penalize variables with many values and threshold on impurity (min_impurity_split)
- Consider only binary splits (max_leaf_nodes=2)
- only consider splits with multiple values (min_samples_leaf)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Handling Missing Values

- If node *n* tests variable X_i, assign most common value of X_i among other instances sorted to node *n*.
- If node *n* tests variable X_i, assign a probability to each of possible values of X_i. These probabilities are estimated based on the observed frequencies of the values of X_i. These probabilities are used in the information gain measure (via info gain).

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

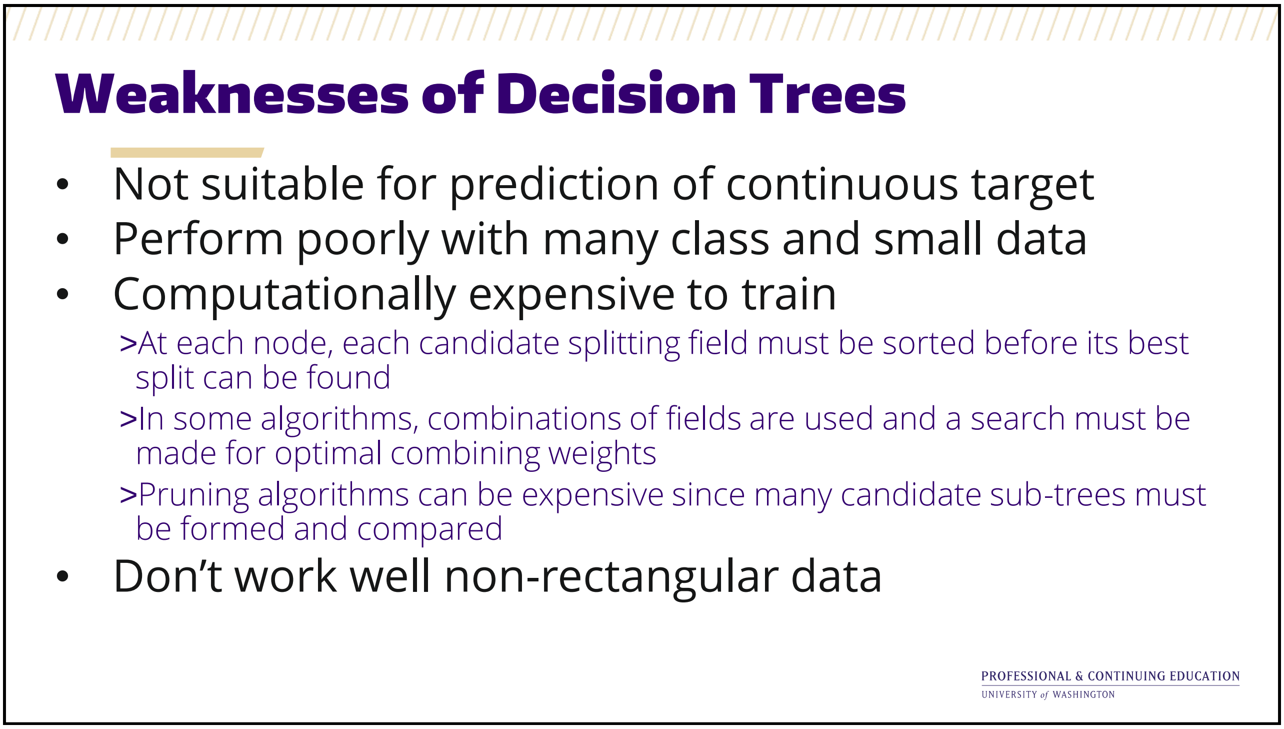
Summary

Decision Trees

Strengths of Decision Trees

- Generates understandable rules
- Performs classification without much computation
- Handles continuous and categorical variables
- Provides a clear indication of which fields are most important for prediction or classification

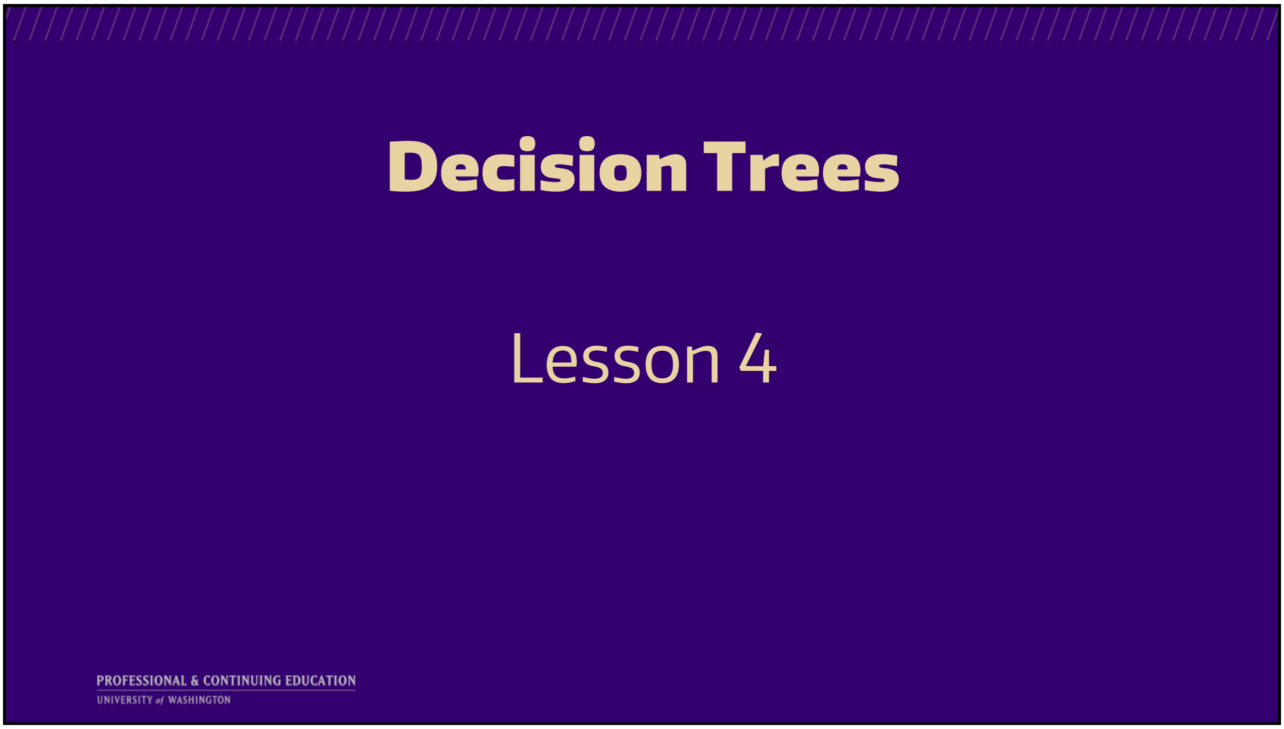
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON



Weaknesses of Decision Trees

- Not suitable for prediction of continuous target
- Perform poorly with many class and small data
- Computationally expensive to train
  - >At each node, each candidate splitting field must be sorted before its best split can be found
  - >In some algorithms, combinations of fields are used and a search must be made for optimal combining weights
  - >Pruning algorithms can be expensive since many candidate sub-trees must be formed and compared
- Don’t work well non-rectangular data

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON



Decision Trees

Lesson 4

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON