Session 2
AI and Machine Learning
Hult International Business School

Michael de la Maza
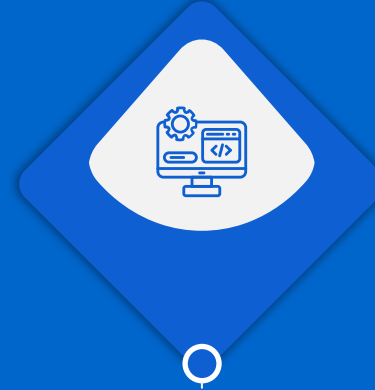
Version 1.1

Advanced Decision Trees: Agenda

Regularization

Boosting

Bagging

Application using XGBoost

What is a decision tree?

- A tree-like structure in which:

  - Internal nodes represent tests on an attribute

  - Each branch represents the outcome of the test

  - Each leaf node represents a class label
    (in the case of a classifier) or a number
    (in the case of a regressor)

Regularization

Regularization intuition: Reduce the risk of overfitting by preferring 'small' models

- Two general approaches:

  - Penalize 'large' models

    Example: In linear regression, prefer models with fewer input variables.

  - Use hyperparameters to 'force' models to be small

    Example: Limit the depth of decision trees

Regularization hyperparameters in decision trees

max_depth:

- If set, then the maximum depth of the tree cannot exceed this parameter.

- If not set, then tree continues to split until nodes are all of the same class (or until the number of samples is less than min_samples_split)

min_samples_split:

- The minimum number of samples needed for a node to be split. If a node has fewer than this number of examples, it will not be split.

- The default value is 2. So, by default, it will continue to split until there is only one sample.
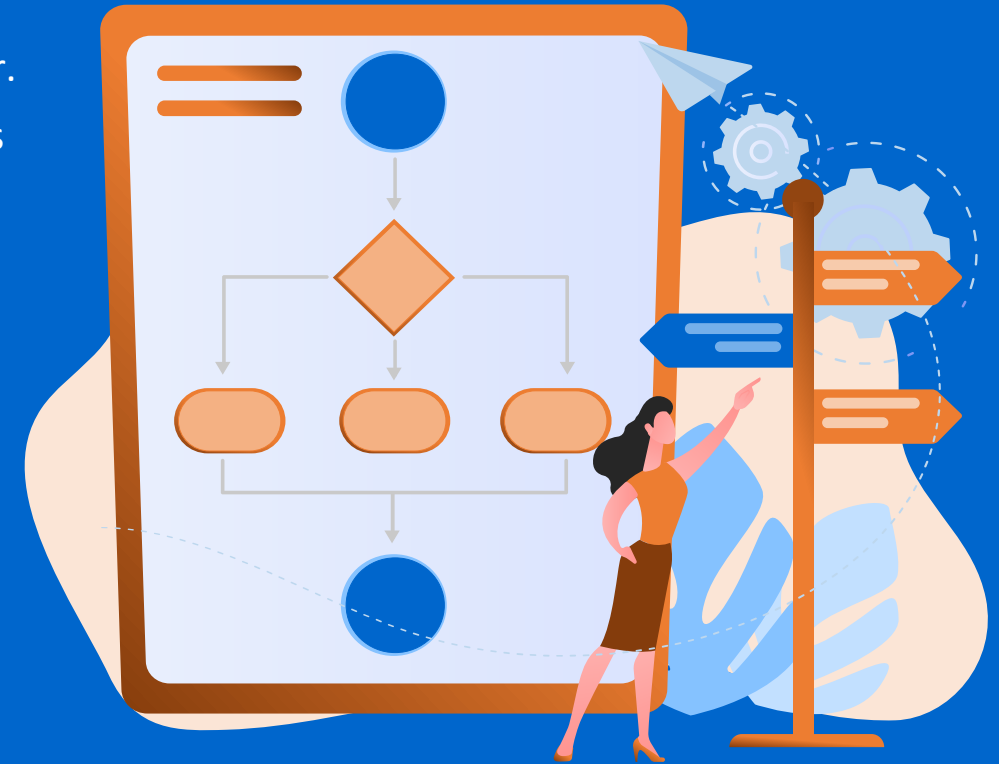
max_leaf_nodes: Sets the maximum number of leaf nodes.

min_impurity_decrease: A node will not be split if the purity decrease is less than this value.

There are several other regularization hyperparameters.

Note: Make sure to add hyperparameter to your glossary

Let's look at code

- Optimize max_depth (just like last session)

- Optimize min_weight_fraction_leaf (require a minimum number of samples at each leaf)

- Optimize min_impurity_decrease (require impurity to decline by a minimum amount)

End result is a smaller tree with same performance as original tree

5 minute exercise

- Select another regularization hyperparameter and train a decision tree with a specific value for that hyperparameter

- Examples:
  - min_samples_split = 20
  - min_samples_leaf = 10
  - max_leaf_nodes = 6

5 minute exercise

Optimize the hyperparameter
(e.g., check values between 1 and 10)

Boosting

Boosting Intuition

Train predictors (classifiers/regressors) in sequence with each subsequent predictor attempting to correct errors made by the previous predictors.

Predictors are combined by weighing predictions according to accuracy: the higher the accuracy, the higher the weight the predictor receives.

Example:

- Train decision tree A on data.
- Train decision tree B on data and increase the weight of instances that A gets wrong.

AdaBoost Classifier

- AdaBoostClassifier

  DecisionTreeClassifier(max_depth=1), n_estimators=5,
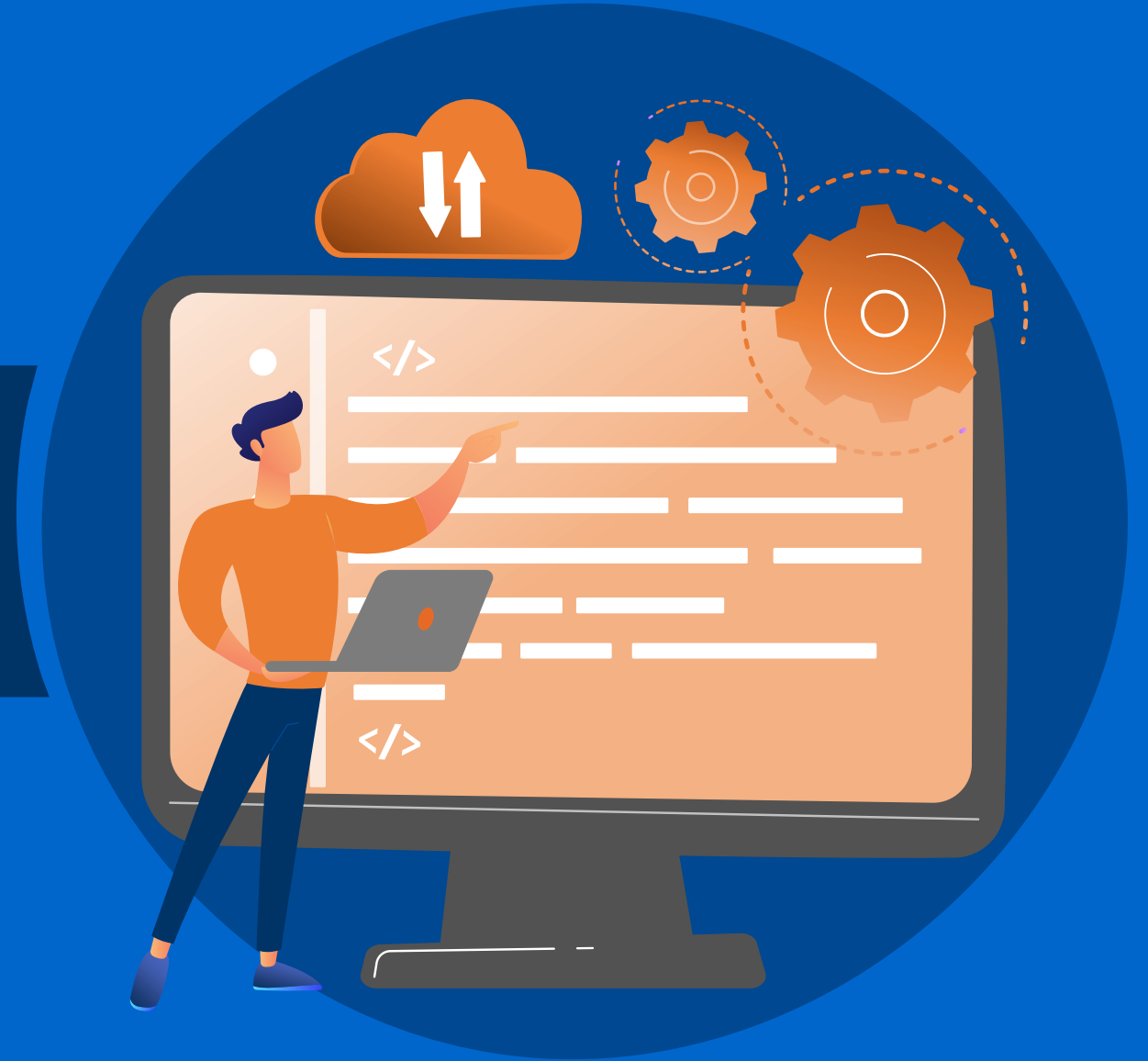
  learning_rate=0.5, random_state=42)

Note: It is normal to have 50+ estimators. We use 5 so we can visualize.

5 minute exercise: Boosting

Try AdaBoost with different n_estimators = 2, 4, 6, 8

- How does accuracy change?
- Which one would you choose?

Advanced: Write a for loop over n_estimators from 1 to 10

- How do you plot the decision trees?
- What is the best value of n_estimator?
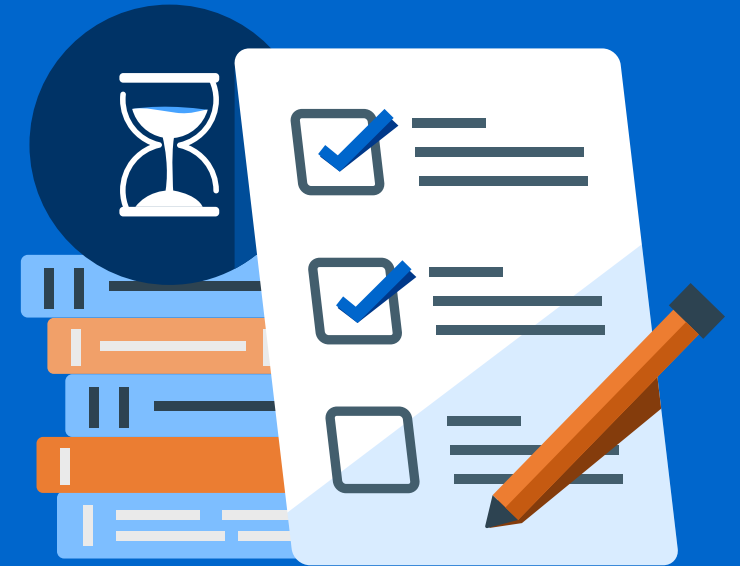
Bagging

Intuition

Train same classifier on different subsets of the training set.

- Sample instances
- Sample features

Produce estimate by aggregating predictions:

- For classifiers, choose the most popular class (or aggregate probabilities).
- For regressors, average the predictions.

5 minute exercise

What do you expect the result to be when:

- n_estimators = 1
- max_samples = 142
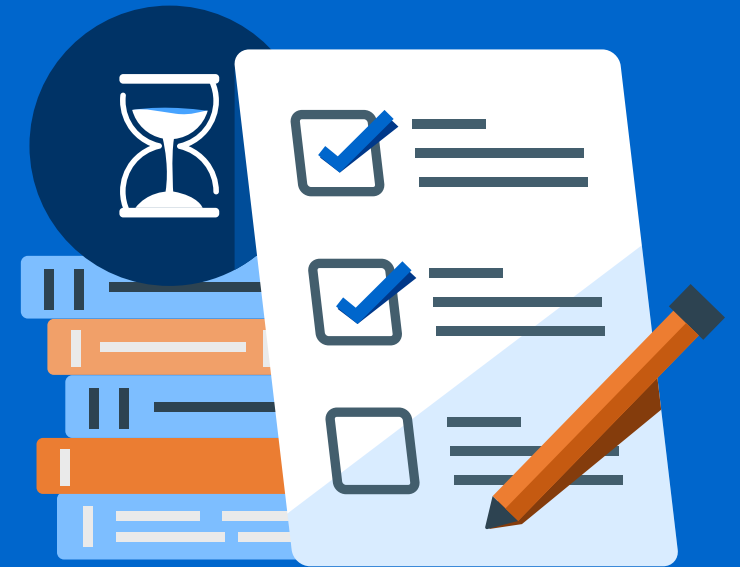
XGBoost

XGBoost: 'extreme' decision trees

- Combines many good ideas for creating decision trees. Focuses on boosting, as the name implies.

- Was the core algorithm in many winning machine learning contests.

- XGBoost is the default model for structured classification problems.
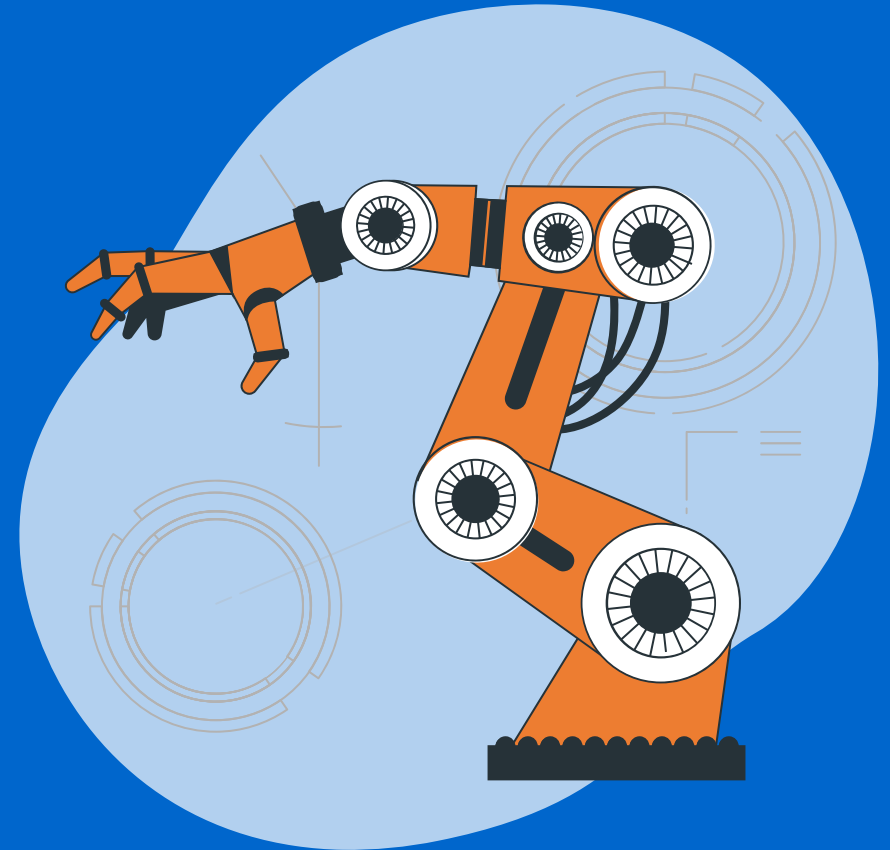
Decision Tree Review

What is a decision tree?

A. A method for optimizing model parameters

B. A tree-like structure used to make predictions based on input features

C. A technique for finding the optimal solution in a search space
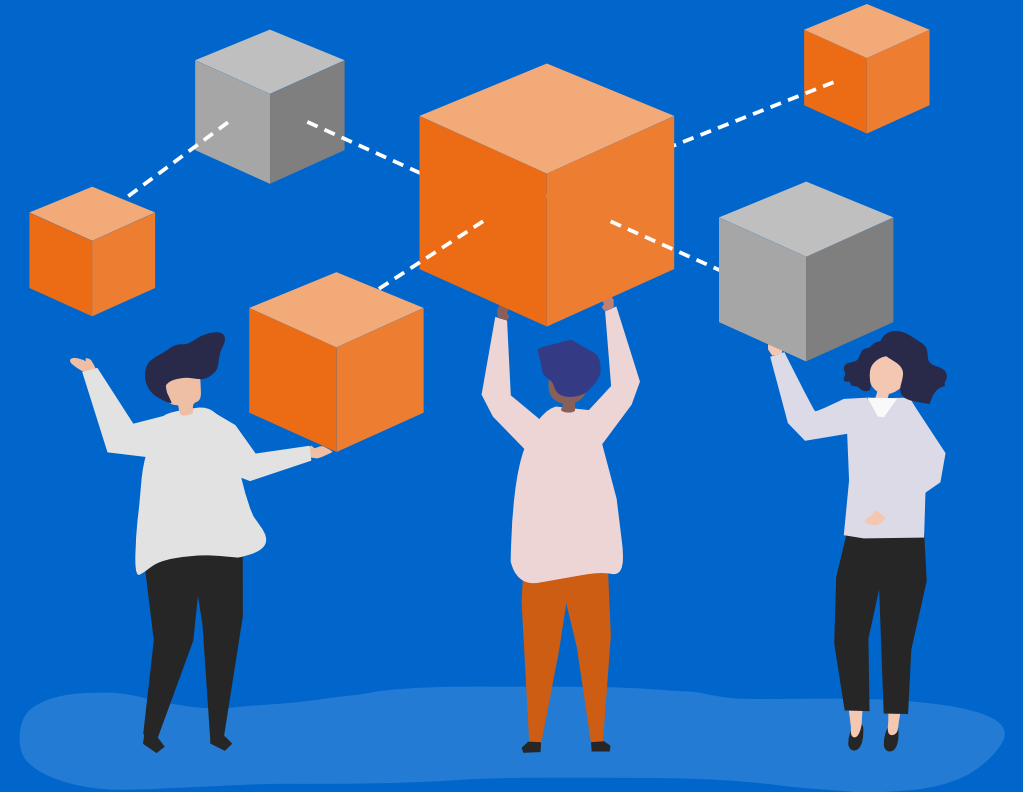
D. A tool for data visualization

What is the primary advantage of using decision trees in machine learning?

A. They are easy to interpret and visualize

B. They can handle missing data

C. They have high predictive accuracy

What are the two types of nodes in decision trees?

A. Red nodes and blue nodes

B. Round nodes and square nodes

C. Decision nodes and leaf nodes

What does this graph indicate?

A. Overfitting

B. Underfitting

C. The training error decreases as the validation error increases

D. The training error increases as the validation error decreases

Source: https://developers.google.com/machine-learning/decision-forests/overfitting-gbdt