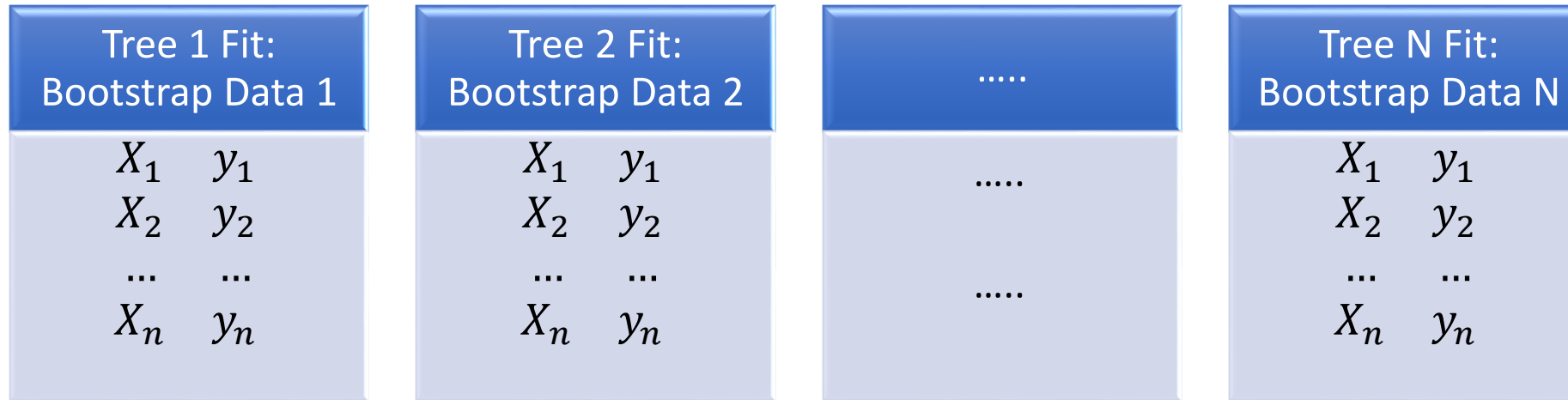


CS109A: HW8 Q6

Topic: Bagging

Note: All the members of this HW group give their consent for the usage of these slides for future classes by the CS109 faculty members, with attribution.

Bagging: Model Training Process



- Randomly generate N bootstrapped datasets, with either some or all observations from the training set
- Fit a simple decision tree (choose appropriate depth) on each of these datasets, so we get N trees

Training Predictions

Tree 1 Predictions: Bootstrap Data 1	Tree 2 Predictions: Bootstrap Data 2	Tree N Predictions: Bootstrap Data N	Train Predictions: Bagging N Trees
X_1 y_{pred1} X_2 y_{pred2} X_n y_{predn}	X_1 y_{pred1} X_2 y_{pred2} X_n y_{predn}	X_1 y_{pred1} X_2 y_{pred2} X_n y_{predn}	<ul style="list-style-type: none">• Regression: $y_{pred} = \frac{\sum_1^N y_{pred N}}{N}$• Classification: $y_{pred} = \max_1^N \{y_{pred N}\}$

- To evaluate the models accuracy on the training set, we use each of the models {1 to N} and predict the response vector y_{pred} for each of the bootstrapped datasets.
- Then we aggregate the results from all N trees by:
 - For regression, calculate the average for that observation for all N trees
 - For Classification, use that class that is predicted maximum number of times among the N trees

Test Predictions

Tree 1 Predictions: Test Data	Tree 2 Predictions: Test Data	Tree N Predictions: Test Data	Test Predictions: Bagging N Trees
X_1 y_{pred1} X_2 y_{pred2} X_n y_{predn}	X_1 y_{pred1} X_2 y_{pred2} X_n y_{predn}	X_1 y_{pred1} X_2 y_{pred2} X_n y_{predn}	<ul style="list-style-type: none">• Regression: $y_{pred} = \frac{\sum_1^N y_{pred N}}{N}$• Classification: $y_{pred} = \max_1^N \{y_{pred N}\}$

- To evaluate the models accuracy on the test set, we use each of the models {1 to N} on the same test data and predict the response vector y_{pred} for the test set.
- Then we aggregate the results from all N trees by:
 - For regression, calculate the average for that observation for all N trees
 - For Classification, use that class that is predicted maximum number of times among the N trees
- **Note:** For test predictions, we use the same test data each time.