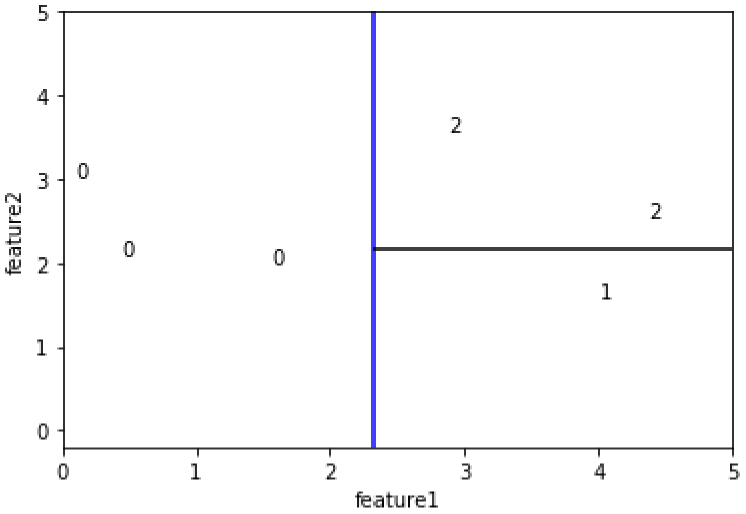
Machine Learning and Algorithms
(Session 6)

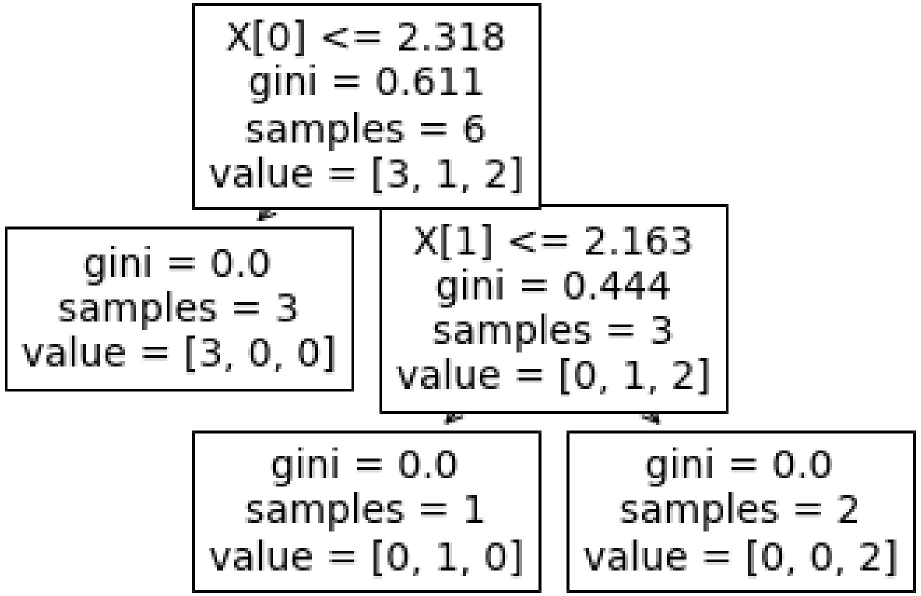
Yi Zhang

March 10, 2022

Review of decision tree classifier

- Try to split the domain into segments
  - Loop through each feature
    - For each feature, loop through each mid-point between values to get the split performance
  - Choose feature that gives the best performance





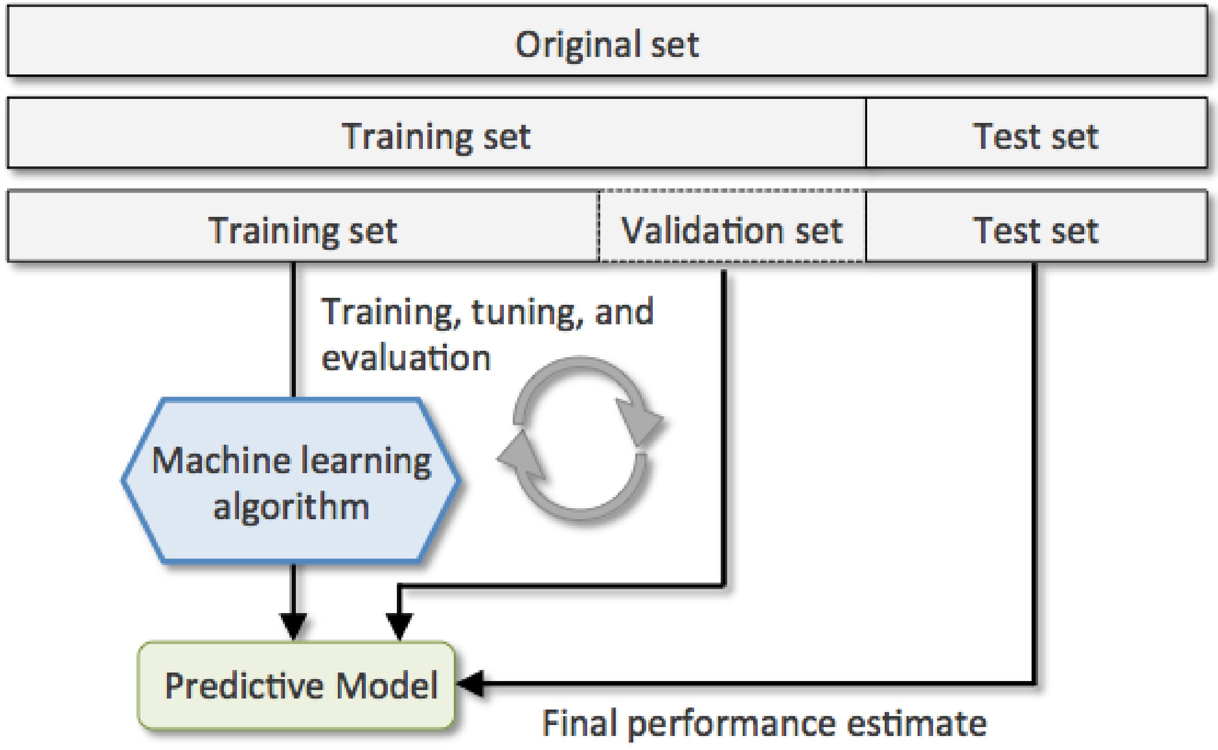
Review of Cross-Validation

- For Decision Tree classification, if we keep on splitting
  - We can get very high accuracy on a dataset if we keep splitting.
  - The model might perform poorly on new data.

- Solution: Split the data into two sets
  - Training (For train the model)
  - Test (For the model performance evaluation)

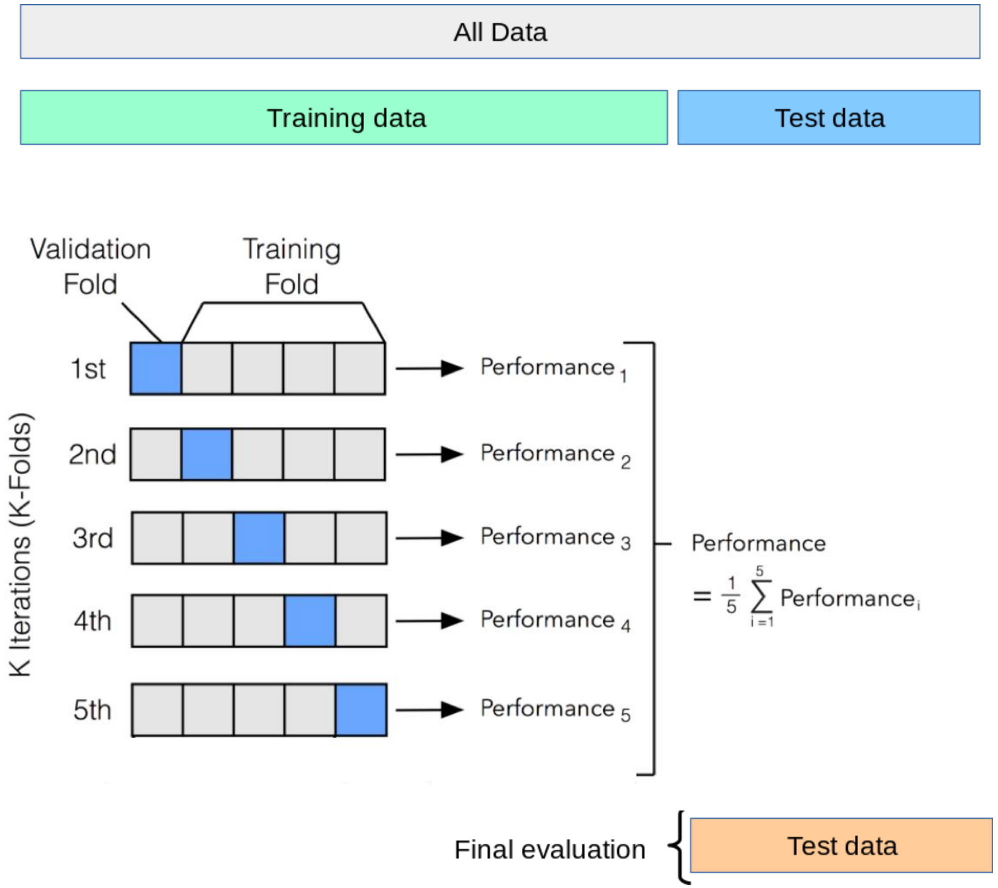
Training-validation-testing

- Training/validation/testing is a technique when we want to perform both model selection and report the accuracy for the best model



1. Split the data into training, validation, and testing
2. Use training to train each model
3. Use validation to validate each model
4. Choose final model that perform best on the validation model
5. Retrain the best model using training set and validation set combined
6. Evaluate accuracy on the test set

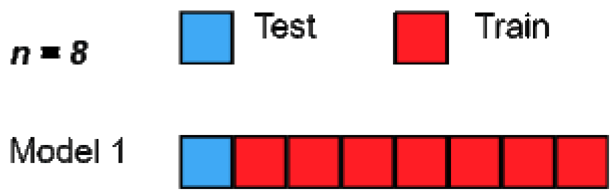
K-fold cross validation



- Split the data into training/testing
- Further split training data into k-folds
  - Each time, use a fold for measure model accuracy
  - Use the rest of the folds to train the model
- Choose the best model based on the average performance
- Re-train the best model on all training data
- Measure accuracy of the best model on the test data

Leave-one-out cross validation

- When we divide training set with n datapoints to n folds, it is called leave-one-out cross validation. Each time, we are only using one sample to evaluate the model performance.
- This method gives very good performance since each time, we use n-1 data points to train the model. However, it is very computational expensive



Decision tree regressor

- Decision tree classifier is used for categorial outcomes
- For continuous outcome, we use decision tree regressor (also called regression tree)
- The algorithm is as follows:
  - For each node
    - Loop through each feature
      - For each feature, loop through each possible split
        - The prediction for each segment is the average outcome
        - Evaluate the quality of the split based on Mean Square Error
    - Choose the best split based on the split quality

Mean square error

- Mean squared error (MSE) can be used to measure the prediction quality for continuous outcomes:

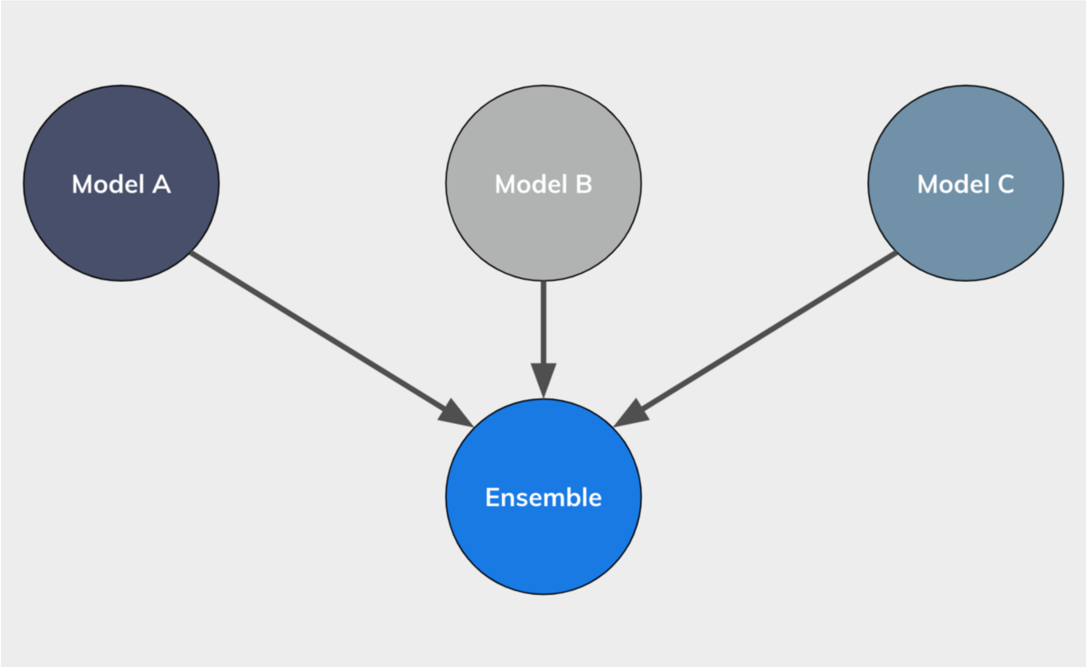
$$MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$

  - y_i: actual outcome for point i
  - $\hat{y}_i$: predicted outcome for point i
  - n: total number of points
  - MSE reaches 0 when $y_i = \hat{y}_i$ for every i (perfect prediction)
- For decision tree regression, prediction is equal to the average outcomes values of all the points belong to this segment

Datapoints	Prediction	MSE
$1,3,0,2,4$	$\frac{1}{5}(1+3+0+2+4) = 2$	$\frac{1}{5}\left((1-2)^2 + (3-2)^2 + (0-2)^2 + (2-2)^2 + (4-2)^2\right) = 2$
$2,6,0,4,-2$	$\frac{1}{5}(2+6+0+4-2) = 2$	$\frac{1}{5}\left((2-2)^2 + (6-2)^2 + (0-2)^2 + (4-2)^2 + (-2-2)^2\right) = 8$

Ensemble method

- Models might tend to become very complicated.
  - Very accuracy on the training set
  - Very inaccurate on the new data
- Use several models/datasets to increase robustness
  - Model averaging
  - Bagging
  - Random Forest
  - Boosting (Later lecture)



Model average

- Run several candidate models
- When predicting the outcome, combine the results of all
  - Regression: Take average
  - Classification: Majority vote

	Model 1 Error	Model 2 Error	Model Averaging
Data 1	-5	5	0
Data 2	-5	5	0
Data 3	0	5	2.5
Data 4	0	0	0
Data 5	5	-5	0
Data 6	0	-5	-2.5
MSE	12.5	20.83	2.08

Model average

- Especially good on models with similar accuracy but negatively correlated predictions

	Model 1 Error	Model 2 Error	Model Averaging
Data 1	-5	5	0
Data 2	-5	5	0
Data 3	0	5	2.5
Data 4	0	0	0
Data 5	5	-5	0
Data 6	0	-5	-2.5
MSE	12.5	20.83	2.08

Model average

- Especially good on models with similar accuracy but negatively correlated predictions

	Model 1	Model 2	Model Averaging
Data 1	-5	-1	-3
Data 2	-5	-1	-3
Data 3	0	0	0
Data 4	0	0	0
Data 5	5	0	2.5
Data 6	0	0	0
MSE	12.5	0.3333	4.04

Bagging (Bootstrap aggregating)

- Use sample with replacement (Bootstrap) to sample N sets of data
- Ran the model on each dataset
- Ensemble the model

X1	X2	y
0.98	0.97	3.53
0.44	0.33	1.3
0.01	0.12	0.49
0.82	0.54	2.8
0.77	0.95	3.24
0.86	0.52	1.97
0.17	0.86	2.75

X1	X2	y
0.86	0.52	1.97
0.86	0.52	1.97
0.82	0.54	2.8
0.01	0.12	0.49
0.77	0.95	3.24
0.77	0.95	3.24
0.98	0.97	3.53

X1	X2	y
0.44	0.33	1.3
0.86	0.52	1.97
0.77	0.95	3.24
0.17	0.86	2.75
0.77	0.95	3.24
0.86	0.52	1.97
0.77	0.95	3.24

Random Forest

- Random Forest has two randomness
- For each model
  - Bootstrap Data
  - Randomly select a subset feature for split
- Ensemble the result