


# Q1-1: Which of the following statements is TRUE?

1. For generalization, we should grow the tree completely to get good training accuracy.
2. To get good training accuracy, we can grow the tree completely and then prune back.
3. We measure the performance of a tree by calculating the fraction of training instances that are correctly classified.
4. Test set should never be used in training the tree.

# Q1-1: Which of the following statements is TRUE?

1. For generalization, we should grow the tree completely to get good training accuracy.
2. To get good training accuracy, we can grow the tree completely and then prune back.
3. We measure the performance of a tree by calculating the fraction of training instances that are correctly classified.
4. Test set should never be used in training the tree. 

1. No. Generalization is not training accuracy
2. That is for generalization
3. Should be on future unseen data like test set
4. Right, so that we can get unbiased estimation of the accuracy

## Q2-1: Which of the following statements is TRUE?

1. If there is no noise, then there is no overfitting.
2. Overfitting may improve the generalization ability of a model.
3. Generalization error is monotone with respect to the capacity/complexity of a model.
4. More training data may help preventing overfitting.

# Q2-1: Which of the following statements is TRUE?

1. If there is no noise, then there is no overfitting.
2. Overfitting may improve the generalization ability of a model.
3. Generalization error is monotone with respect to the capacity/complexity of a model.
4. More training data may help preventing overfitting.



1. We can still have false correlation that leads to overfitting.
2. Overfitting would undermine the generalization ability.
3. Generalization error would first decrease and then increase as the model capacity increases.
4. Increasing training data size would help better approximate the true distribution.

## Q2-2: Which of the following may not lead to overfitting?

1. Limited training data.
2. Noise in the training data.
3. Too many hypotheses made for the model.
4. Increase the training data size if the model capacity is increased

## Q2-2: Which of the following may not lead to overfitting?

1. Limited training data.
2. Noise in the training data.
3. Too many hypotheses made for the model.
4. Increase the training data size if the model capacity is increased



(1), (2) and (3) are all from the lecture.

(4): Although the model become more complicated, which may lead to overfitting, increasing that training data size would help prevent overfitting as long as we keep certain increasing rate.

Q2-3: Are these statements true or false?

(A) Early stopping is a practical technique to reduce the training error.

(B) Pruning reduces the complexity of a decision tree but will improve its generalization ability.

1. True, True
2. True, False
3. False, True
4. False, False

Q2-3: Are these statements true or false?

(A) Early stopping is a practical technique to reduce the training error.

(B) Pruning reduces the complexity of a decision tree but will improve its generalization ability.

1. True, True
2. True, False
3. False, True
4. False, False



1. Early stopping is designed for improving generalization ability (reducing generalization error). Somehow it may cause the training error not so low.
2. Pruning reduces certain sub-trees thus reducing the complexity of a decision tree. By properly doing so, we can improve the generalization ability by preventing overfitting.



Q2-4: Are these statements true or false?

(A) A validation set is used for model selection.

(B) Post-pruning is usually based on the generalization error on the test set .

1. True, True
2. True, False
3. False, True
4. False, False

Q2-4: Are these statements true or false?

(A) A validation set is used for model selection.

(B) Post-pruning is usually based on the generalization error on the test set .

1. True, True
2. True, False
3. False, True
4. False, False



1. A validation set is usually used for evaluating the generalization error while training, so we can choose the best model of best generalization ability among many candidate models. On the contrary, the test set is only used to estimate the generalization performance of the chosen best model at last if there are validation sets.
2. Post-pruning is usually based on the generalization error on the validation set.

Q3-1: Are these statements true or false?

(A) Least square can be used in regression trees.

(B) CART can be used to construct regression trees.

1. True, True
2. True, False
3. False, True
4. False, False

Q3-1: Are these statements true or false?

(A) Least square can be used in regression trees.

(B) CART can be used to construct regression trees.

1. True, True



2. True, False

3. False, True

4. False, False

(1) is based on the lecture.

(2): Yes, it can be for classification and regression.

# Q3-2: Calculate information gain with lookahead. The value of choosing Humidity as our split is

Suppose we already know that

$$H_D(Y) = 0.999$$

$$H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind} = \text{strong}) = 0.971,$$

$$H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind} = \text{weak}) = 0.991,$$

$$H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature} = \text{high}) = 1,$$

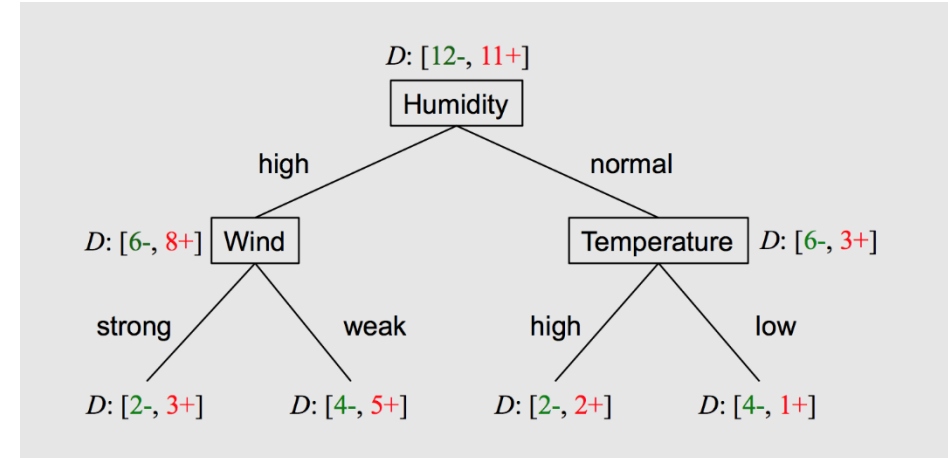
$$H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature} = \text{low}) = 0.722.$$

1.  $\sim 0.069$

2.  $\sim 0.151$

3.  $\sim 0.023$

4.  $\sim 0.255$



# Q3-2: Calculate information gain with lookahead.

## The value of choosing Humidity as our split is

Suppose we already know that

$$H_D(Y) = 0.999$$

$$H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind} = \text{strong}) = 0.971,$$

$$H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind} = \text{weak}) = 0.991,$$

$$H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature} = \text{high}) = 1,$$

$$H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature} = \text{low}) = 0.722.$$

1.  $\sim 0.069$



2.  $\sim 0.151$

3.  $\sim 0.023$

4.  $\sim 0.255$

The value is

$$H_D(Y) - \left[ \frac{14}{23} H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind}) + \frac{9}{23} H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature}) \right]$$

$$= 0.999$$

$$- \left[ \frac{5}{23} H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind} = \text{strong}) + \frac{9}{23} H_D(Y \mid \text{Humidity} = \text{high}, \text{Wind} = \text{weak}) + \right.$$

$$\left. \frac{4}{23} H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature} = \text{high}) + \frac{5}{23} H_D(Y \mid \text{Humidity} = \text{normal}, \text{Temperature} = \text{low}) \right]$$

$$\approx 0.999 - 0.930 \approx 0.069$$

