Alden Park Homework 3

Created	@July 6, 2023 2:39 PM
Class	Machine Learning
Type	
Materials	
Reviewed	☐

Question 1: [4 points] Explain what is the bias-variance trade-off? Describe few techniques to

reduce bias and variance respectively

- As you increase the variance in the training dataset, you will tend to have a overfitted training model. However, higher variance results in lower bias in a training dataset. A highly bias training model will result in underfitting the data. The goal is to find an optimal model that balances variance and bias. If one of these measures is too high, you will end up with a sub-optimal model that overfits (high variance) or underfits (high bias) the test data.
- To reduce bias, you can
  - Increase the complexity of the model by adding relevant features to increase the overall sensitivity of the model.
  - You may also reduce the strength of regularization in the model to capture more patterns in the dataset
- To reduce variance:
  - Use a larger dataset, which will give the model more examples to learn from and reduce the sensitivity to new data noise and fluctuations.
  - You may also use cross-validation to evaluate the performance of the model by using k-fold and make sure that the models perform consistently with new dataset

Question 2: [6 points] Assume the following confusion matrix of a classifier. Please compute its

1) precision,

2) recall, and

3) F1-score.

		Predicted results	
Actual values		Class 1	Class 2
	Class 1	50	30
	Class 2	40	60

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

1. Precision = 50/(50 + 30) = `0.62`
2. Recall = 50/(50+40) = `0.55`
3. F1-Score =2((0.62 * 0.55) / (0.62 + 0.55)) = 2 * (0.341/1.17) = `0.582`

Question 3: [10 points] Build a decision tree using the following training instances (using

information gain approach):

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

$$Entropy = -\sum_{i}^{n} log_2(P_i)$$

n= # of features

i = features

P = probability of i

Entropy of the root node (play tennis):

- Probability (yes) = 6/10 = 0.6
- Probability (no) = 4/10 = 0.4

$$E = -(P_y * log_2(P_y) + P_n * log_2(P_n))$$

$$E = -(0.6 * log_2(0.6) + 0.4 * log_2(0.4)) = 0.97$$

Information Gain:

$$IG = Entropy_p - weightedAvg * Entropy_c$$

Outlook:

Sunny: 4 total

- 1 yes
- 3 no

`(0.25 log2(0.25)+ 0.75 log2(0.75)) = Entropy`

Entropy = 0.811

Overcast: 2 total

- 2 yes
- 0 no

Entropy = 0

Rain: 4 total

- 3 yes
- 1 no

Entropy=0.811

$$ChildrenEntropy = 0.4(0.81) + 0 + 0.4(0.81) = 0.64$$

$$IG = 0.97 - 0.64 = 0.33$$

`Information Gain with outlook = 0.33`

Temperature:

$$ChildrenEntropy = 1$$

$$IG = 0.97 - 1 = -0.03$$

`Information Gain with Temperature = -0.03`

Humidity:

$$ChildrenEntropy = 0.84$$

$$IG = 0.97 - 0.84 = 0.13$$

`Information Gain with Humidity = 0.13`

Wind:

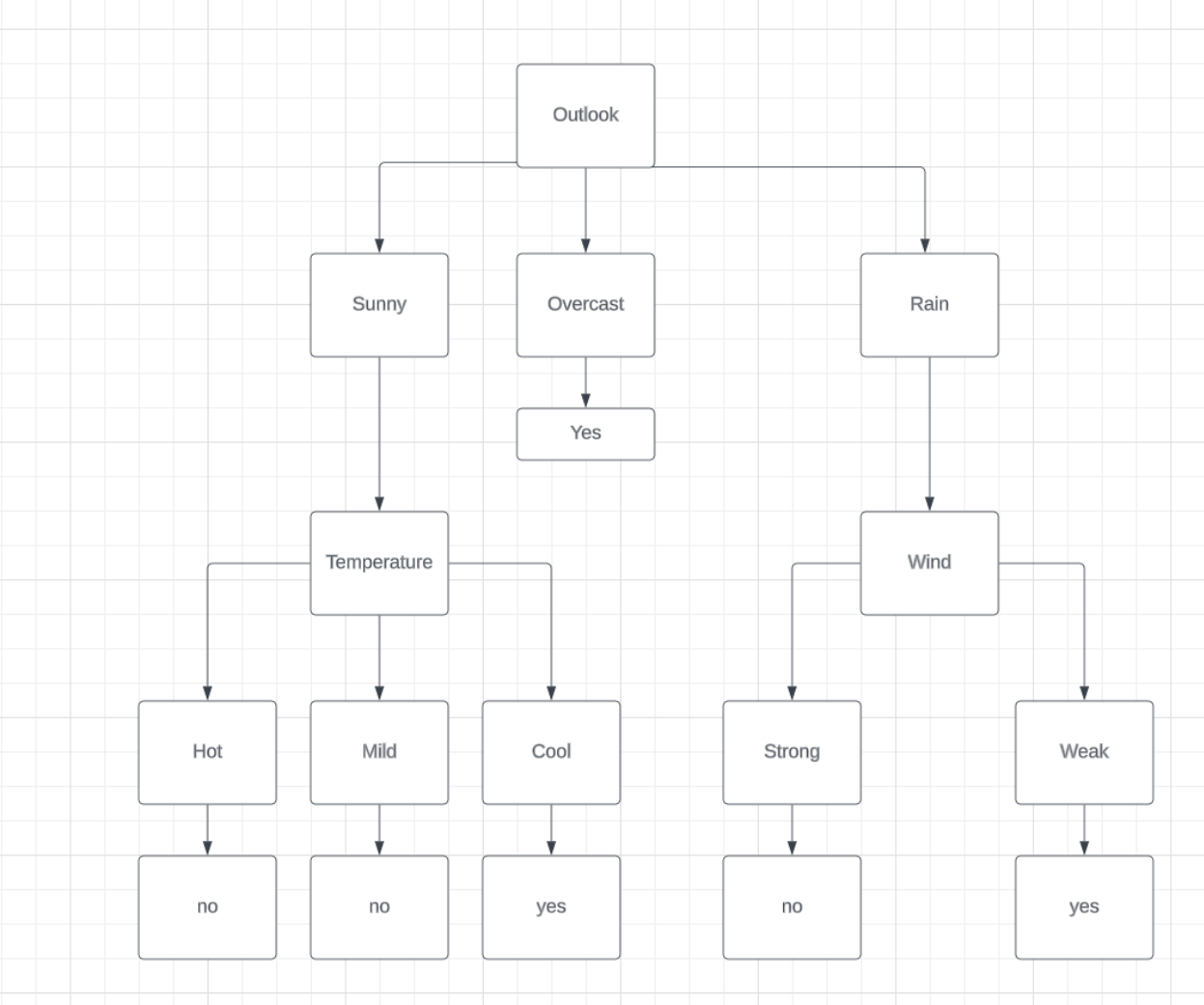
$$ChildernEntropy = 0.87$$

$$IG = 0.97 - 0.87 = 0.1$$

`Information Gain with Wind = 0.1`

We choose based on the highest information gain for each children node to make the following children branches on the decision tree.

- Outlook had the highest IG
- We then looked at each of the child nodes and perform another information gain
- For sunny, temperature had the highest IG and therefore led to the next branch
- Overcast only had one option, resulting in yes
- Rain had wind as the highest IG and therefore led to the next branch



Question 4. [10 points] The naïve Bayes method is an ensemble method as we learned in
Module 5. Assuming we have 3 classifiers, and their predicted results are given in the table 1.
The confusion matrix of each classifier is given in table 2. Please give the final decision using the
Naïve Bayes method:

Table 1 Predicted results of each classifier

Sample x	Result
Classifier 1	Class 1
Classifier 2	Class 1
Classifier 3	Class 2

Table 2 Confusion matrix of each classifier

i) Classifier 1

	Class1	Class2
Class1	40	10
Class2	30	20

ii) Classifier 2

	Class1	Class2
Class1	20	30
Class2	20	30

iii) Classifier 3

	Class1	Class2
Class1	50	0
Class2	40	10

Classifier 1:

$$TP = \frac{40}{60} = 0.66$$

$$FP = \frac{10}{60} = 0.16$$

$$Accuracy = 60/100 = 0.6$$

Classifier 2:

$$TP = 20/50 = 0.4$$

$$FP = 30/50 = 0.6$$

$$Accuracy = 50/100 = 0.5$$

Classifier 3:

$$TP = 0.83$$

$$FP = 0$$

`Therefore Classifier 1 is the best one`