Name: Janmejay Mohanty **Course:** Applied Machine Learning **Homework Assignment**: 3
CWID: 20009315 **Course Number:** AAI 695

Solutions

Ans1: The bias-variance trade-off is one of the fundamental concepts of machine learning which is related to the performance of a model.

When the error is caused by approximating a real-world problem using a simplified model it is known as Bias. When the model not able to get correct relationship between output variable and input features, then this type of error gets occurred. High bias can lead to underfitting, when the model is too simple and cannot capture the underlying patterns in the data.

When the error is caused by model sensitivity to variations in training data it is known as Variance. This is due to variability of predictions computed by model for different training data. High variance can lead to overfitting, when the model is too complex and captures noise in training data, which leads to bad performance on new data.

Therefore, the goal is to find a model that gives optimal balance between bias and variance, which results in good performance for both training and testing data.

To reduce bias, we can use these following techniques:

1. Increasing features
2. Increasing model complexity
3. Performing feature engineering
4. Reducing the alpha parameter of regularization

To reduce variance, we can use these following techniques:

1. Increasing training set
2. Reducing features
3. Increasing Lambda
4. Dropout out some of the neurons in a neural network during training

Ans2: For Class 1

Predicted results

Actual values

	Class 1	Class 2
Class 1	50 (TP)	30 (FN)
Class 2	40 (FP)	60 (TF)

1) Precision

$$precision = \frac{TP}{TP + FP}$$
$$= \frac{50}{50 + 40}$$
$$= \frac{50}{90}$$
$$= 0.5556$$

2) Recall

$$recall = \frac{TP}{TP + FN}$$
$$= \frac{50}{50 + 30}$$

$$= \frac{50}{80}$$
$$= 0.625$$

3) F1 score

$$f1\ score\ = \frac{2(precision * recall)}{precision + recall}$$
$$= \frac{2(0.5556 * 0.625)}{0.5556 + 0.625}$$
$$= 2 * \frac{0.3473}{1.1806}$$
$$= 2 * 0.2942$$
$$= 0.5884$$

For Class 2

Predicted results

Actual values

	Class 1	Class 2
Class 1	50 (TN)	30 (FP)
Class 2	40 (FN)	60 (TP)

4) Precision

$$precision = \frac{TP}{TP + FP}$$
$$= \frac{60}{60 + 30}$$
$$= \frac{60}{90}$$
$$= 0.6667$$

5) Recall

$$recall = \frac{TP}{TP + FN}$$
$$= \frac{60}{60 + 40}$$
$$= \frac{60}{100}$$
$$= 0.6$$

6) F1 score

$$f1\ score\ = \frac{2(precision * recall)}{precision + recall}$$
$$= \frac{2(0.6667 * 0.6)}{0.6667 + 0.6}$$
$$= 2 * \frac{0.4}{1.2667}$$
$$= 2 * 1.3596$$
$$= 2.7192$$

Ans3: Taking root node PlayTennis as PT, calculating the entropy for complete dataset and information gain for each attribute.

$$E(PT) = -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= -\frac{6}{10} * -0.737 - \frac{4}{10} * -1.3219$$

$$= 0.4422 + 0.5288$$

$$= 0.971$$

$$Gain(Outlook) = E(PT) - E(Outlook, PT)$$

$$= 0.971 - \left\{\frac{4}{10}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] + \frac{2}{10}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{4}{10}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]\right\}$$

$$= 0.971 - \left\{2 * \frac{4}{10}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] + \frac{2}{10}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right]\right\}$$

$$= 0.971 - \left\{2 * \frac{4}{10}[-(0.75)\,(-0.415) - (0.25)(-2)] + \frac{1}{5}[0]\right\}$$

$$= 0.971 - \left\{2 * \frac{4}{10}[0.3113 + 0.5]\right\}$$

$$= 0.971 - \frac{4}{5} * 0.8113$$

$$= 0.971 - 0.64904$$

$$= -0.322$$

$$Gain(Temperature) = E(PT) - E(Temperature, PT)$$

$$= 0.971 - \left\{\frac{3}{10}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] + \frac{3}{10}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] + \frac{4}{10}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]\right\}$$

$$= 0.971 - \{0.3[(-0.6667 * -0.5849) - (0.3333 * -1.5851)] + 0.3[(-0.6667 * -0.5849) - (0.3333 * -1.5851)] + 0.4[(-0.75 * -0.415) - (0.25 * -2)]\}$$

$$= 0.971 - \{0.3[0.39 + 0.5283] + 0.3[0.39 + 0.5283] + 0.4[0.3113 + 0.5]\}$$

$$= 0.971 - \{2 * 0.3[0.39 + 0.5283] + 0.4[0.3113 + 0.5]\}$$

$$= 0.971 - \{2 * 0.3 * 0.9183 + 0.4 * 0.8113\}$$

$$= 0.971 - \{0.551 + 0.3245\}$$

$$= 0.971 - 0.8755$$

$$= 0.0955$$

$$Gain(Humidity) = E(PT) - E(Humidity, PT)$$

$$= 0.971 - \left\{\frac{5}{10}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{5}{10}\left[-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right]\right\}$$

$$= 0.971 - \left\{\frac{5}{10}[(-0.4 * -1.3219) - (0.6 * -0.737)] + \frac{5}{10}[(-0.8 * -0.3219) - (0.2 * -2.3219)]\right\}$$

$$= 0.971 - \left\{\frac{5}{10}[0.5288 + 0.4422] + \frac{5}{10}[0.2575 + 0.4644]\right\}$$

$$= 0.971 - \left\{\frac{5}{10}(0.971) + \frac{5}{10}(0.7219)\right\}$$

$$= 0.971 - \{0.4855 + 0.361\}$$

$$= 0.971 - 0.8465$$

$$= 0.1245$$

$$Gain(Wind) = E(PT) - E(Wind, PT)$$

$$= 0.971 - \left\{\frac{7}{10}\left[-\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right] + \frac{3}{10}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right]\right\}$$

$$= 0.971 - \{0.7[(-0.7143 * -0.4854) -(0.2857 * -1.8074)] + 0.3[(-0.3333 * -1.5851) - (0.6667 * -0.5849)]\}$$

$$= 0.971 - \{0.7[(0.3467 + 0.5164)] + 0.3[(0.5283) - (0.39)]\}$$

$$= 0.971 - \{0.7 * 0.8631 + 0.3 * 0.2060\}$$

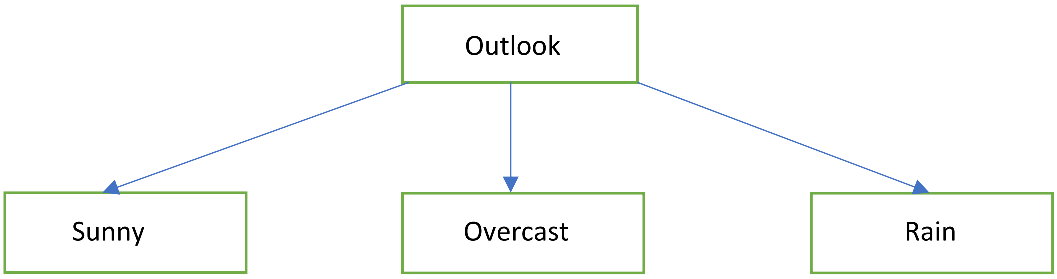
$$= 0.971 - \{0.6042 + 0.0618\}$$

$$= 0.971 - 0.666$$

$$= 0.305$$

From the above results, we can conclude that $Gain(Outlook)$ is the highest.

Hence Outlook is root node.



Now for second layer,

$$E(Sunny) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$= (-0.25 * -2) - (0.75 * -0.415)$$

$$= 0.5 + 0.3113$$

$$= 0.8113$$

$$E(Rain) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$$= (0.75 * -0.415) - (-0.25 * -2)$$

$$= 0.3113 + 0.5$$

$$= 0.8113$$

$$Gain(Sunny, Temperature) = 0.8113 - \left\{\frac{2}{4}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{1}{4}[-\log_2 1] + \frac{1}{4}[-\log_2 1]\right\}$$

$$= 0.8113 - \left\{\frac{2}{4}[-1(0)] + \frac{1}{4}(0) + \frac{1}{4}(0)\right\}$$

$$= 0.8113$$

$$Gain(Sunny, Humidity) = 0.8113 - \left\{\frac{3}{4}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] + \frac{1}{4}[-\log_2 1]\right\}$$

$$= 0.8113 - \left\{\frac{3}{4}[-1(0)] + \frac{1}{4}(0)\right\}$$

$$= 0.8113$$

$$Gain(Sunny, Wind) = 0.8113 - \left\{\frac{3}{4}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] + \frac{1}{4}[-\log_2 1]\right\}$$

$$= 0.8113 - \left\{\frac{3}{4}\left[(-0.3333 * -1.5851) - \frac{2}{3}[-0.5849]\right] + \frac{1}{4}(0)\right\}$$

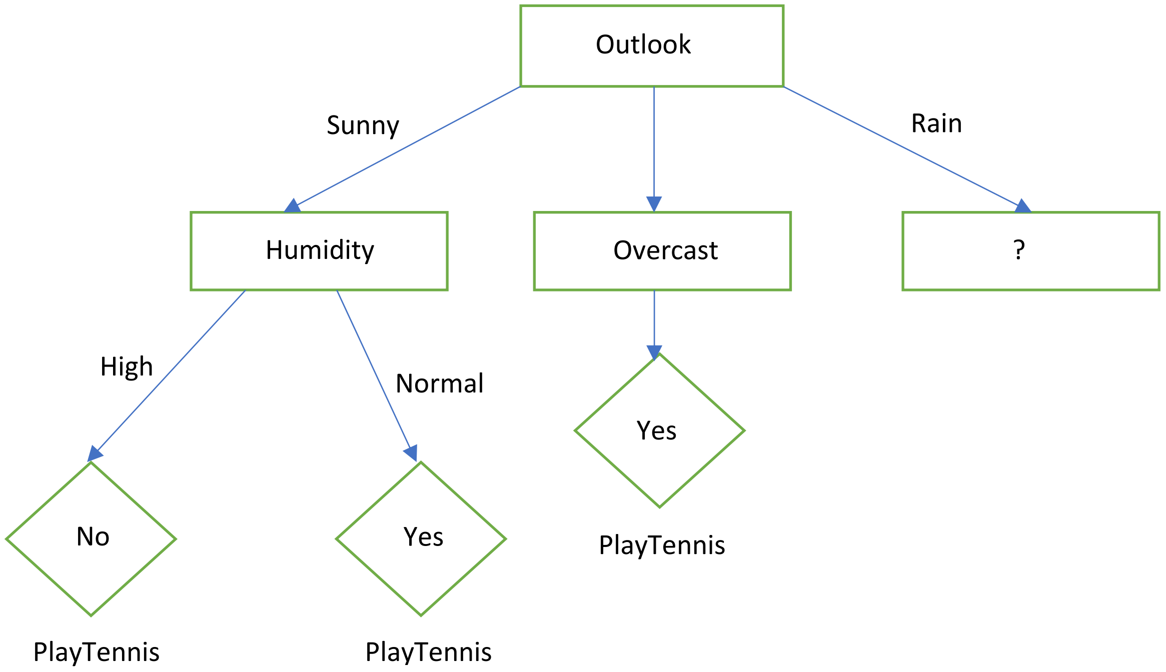
$$= 0.8113 - \left\{\frac{3}{4}[0.5283 + 0.39]\right\}$$

$$= 0.8113 - \frac{3}{4}(0.9183)$$

$$= 0.8113 - 0.6887$$

$$= 0.1226$$

As, both $Gain(Sunny, Temperature)$ and $Gain(Sunny, Humidity)$ are equal but Humidity has more no of same type of data points classifying as target class. Therefore, we are taking $(Sunny\ and\ Humidity)$



Now calculating for Rain,

$$Gain(Rain, Wind) = 0.8113 - \left\{\frac{3}{4}\left[\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] + \frac{1}{4}[-\log_2 1]\right\}$$

$$= 0.8113 - \left\{\frac{3}{4}[0] + \frac{1}{4}[0]\right\}$$

$$= 0.8113$$

$$Gain(Rain, Temperature) = 0.8113 - \left\{\frac{2}{4}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]\right\}$$

$$= 0.8113 - \left\{\frac{2}{4}[0] + \frac{2}{4}\left[-2 * \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]\right\}$$

$$= 0.8113 - \left\{\frac{1}{2}[-1 * -1]\right\}$$

$$= 0.8113 - 0.5$$

$$= 0.3113$$

$$Gain(Rain, Humidity) = 0.8113 - \left\{\frac{3}{4}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] + \frac{1}{4}[-\log_2 1]\right\}$$

$$= 0.8113 - \left\{\frac{3}{4}[(-0.6667 * -0.5849) -(0.3333 * -1.5851)] + \frac{1}{4}[0]\right\}$$

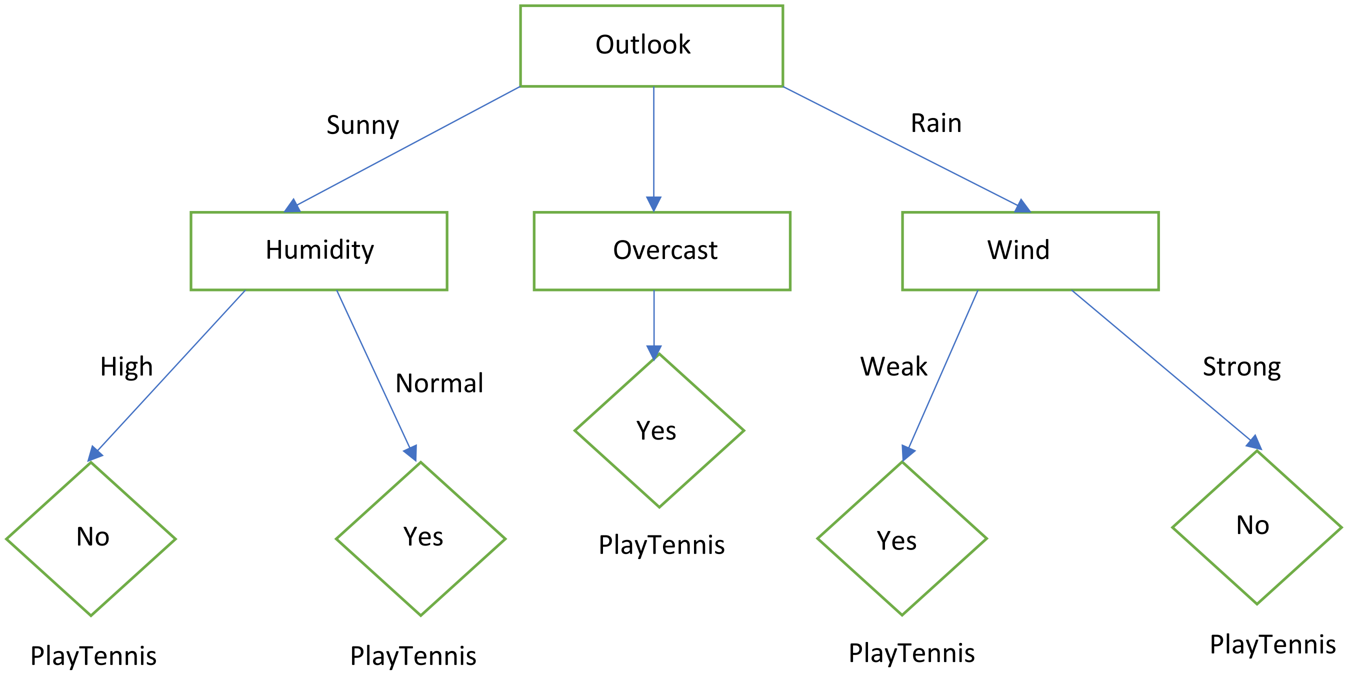
$$== 0.8113 - \left\{\frac{3}{4}[(0.39 +0.5283]\right\}$$

$$= 0.8113 - \frac{3}{4}(0.9183)$$

$$= 0.8113 - 0.6887$$

$$= 0.1226$$

Hence $Gain(Rain, Wind)$ is highest among others,



Hence, we got all classified leaf nodes as per target value.

Therefore, the above decision diagram is based upon our findings.

Ans4: The individual classifiers outputs are as follows:

Classifier 1 = Class 1

Classifier 2 = Class 1

Classifier 3 = Class 2

$$\hat{p}\left(w_1 \middle| d_{1,1}(x) = 1\right) = \frac{40}{70} = 0.5714$$

$$\hat{p}\left(w_1 \middle| d_{2,1}(x) = 1\right) = \frac{20}{40} = 0.5$$

$$\hat{p}\left(w_1 \middle| d_{3,2}(x) = 1\right) = \frac{0}{10} = 0$$

$$=> Class\ 1 = 0$$

$$\hat{p}\left(w_2 \middle| d_{1,1}(x) = 1\right) = \frac{30}{70} = 0.4286$$

$$\hat{p}\left(w_2 \middle| d_{2,1}(x) = 1\right) = \frac{20}{40} = 0.5$$

$$\hat{p}\left(w_2 \middle| d_{3,2}(x) = 1\right) = \frac{10}{10} = 1$$

$$=> Class\ 2 = 0.214$$

As, per comparison between the Class 1 and Class 2 probability. The Class 2 has higher probability than Class 1.

Therefore, we go with Class 2.