AAI 695 - HW

Q.1 Bias-variance tradeoff

Bias is the inability of a machine learning model to capture the true relationship between data points due to its iherent assumptions.

eg: A linear regression model assumes a linear relationship between variables

Variance is the difference in how a model fits over new test data & training data, an indicator of overfitting.

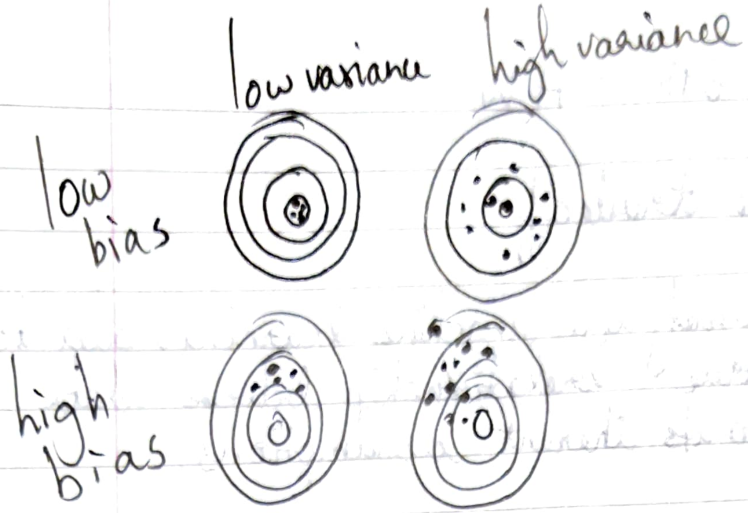
High bias: assuming more about the rel^n b/w variables

low bias: fewer assumptions by the learning algo

High bias + low variance: underfitting
(can't capture the rel^n b/w variables)

low bias + high variance: overfitting
(can't generalize over new data)

We need to find a good balance between the bias and variance of the model, without overfitting or underfitting, this is called the bias-variance tradeoff.



- to fix bias, adding more i/p s will help the model fit better.
- adding more polynomicals can make it more complexity to the model
- to fix variance, reducing i/p features or using features with more importance to reduce overfitting
- more training data will also help.

Q.2

actual

50	30
40	60

predicted

$$\text{precision} = \frac{TP}{TP+FP} = \frac{50}{40+50} = 0.555$$

$$\text{recall} = \frac{TP}{TP+FN} = \frac{50}{50+30} = 0.625$$

$$f_1\text{-score} = 2 \times (0.555 \times 0.625) / (0.625 + 0.555)$$
$$= 0.587$$

Ans. 3 target → play (6 P, 4 N)

$$\text{entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$\text{info gain} = \text{entropy}(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|}\,\text{entropy}(S_v)$$

$$\text{entropy of play} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$$
$$= 0.973$$

$$\text{entropy of sunny} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$
(1P, 3N)
$$= 0.817$$

$$\text{entropy of overcast} = -\frac{2}{2}\log 1 - 0 = 0$$
(2 P, 0N)

$$\text{entropy of rain} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$
$$= 0.817$$

$$\text{gain}(s, \text{outlook}) = E_S - \sum \frac{|S_v|}{|S|}\,\text{entropy}\ S_v$$
$$= 0.973 - \frac{4}{10} \times 0.817 - \frac{4}{10} \times 0.817$$
$$= \boxed{0.3194}$$

$$E_{Hot}(1P, 2N) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$
$$= 0.93$$

$$E_{mild}(2P, 1N) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$$
$$= 0.93$$

$$E_{cool}(2P, 1N) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.93$$

$$E_{High}(2P, 3N) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$
$$\cancel{= 0.159} \quad = 0.973$$

$$E_{Normal}(4P, 1N) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$$
$$= 0.72$$

$$gain(S, temp) = 0.973 - \frac{3}{10} \times 0.93 \times 3 = \boxed{0.136}$$

$$gain(S, humidity) = 0.973 - \frac{5}{10} \times 0.973 - \frac{5}{10} \times 0.72$$
$$= \boxed{0.126}$$

$$E_{weak}(5P, 2N) = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7}$$
$$= 0.87$$

$$E_{strong}(1P, 2N) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$
$$= 0.93$$

$$gain(S, wind) = 0.973 - \frac{7}{10} \times 0.876 - \frac{3}{10} \times 0.93$$
$$= \boxed{0.081}$$

gain (S, outlook) has highest gain

outlook

sunny outcast rain

yes

sunny, humidity:

$$E_{high}(0P, 3N) = 0, \quad E_{normal}(1P, 0N) = 0$$

$$gain(sun, humid) = 0.817$$

$$E_{Hot}(0P, 2N) = 0, \quad E_{mid}(0P, 1N) = 0, \quad E_{cool}(1P, 0N) = 0$$

$$gain(sun, temp) = 0.817$$

$$E_{weak}(1P, 2N) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$

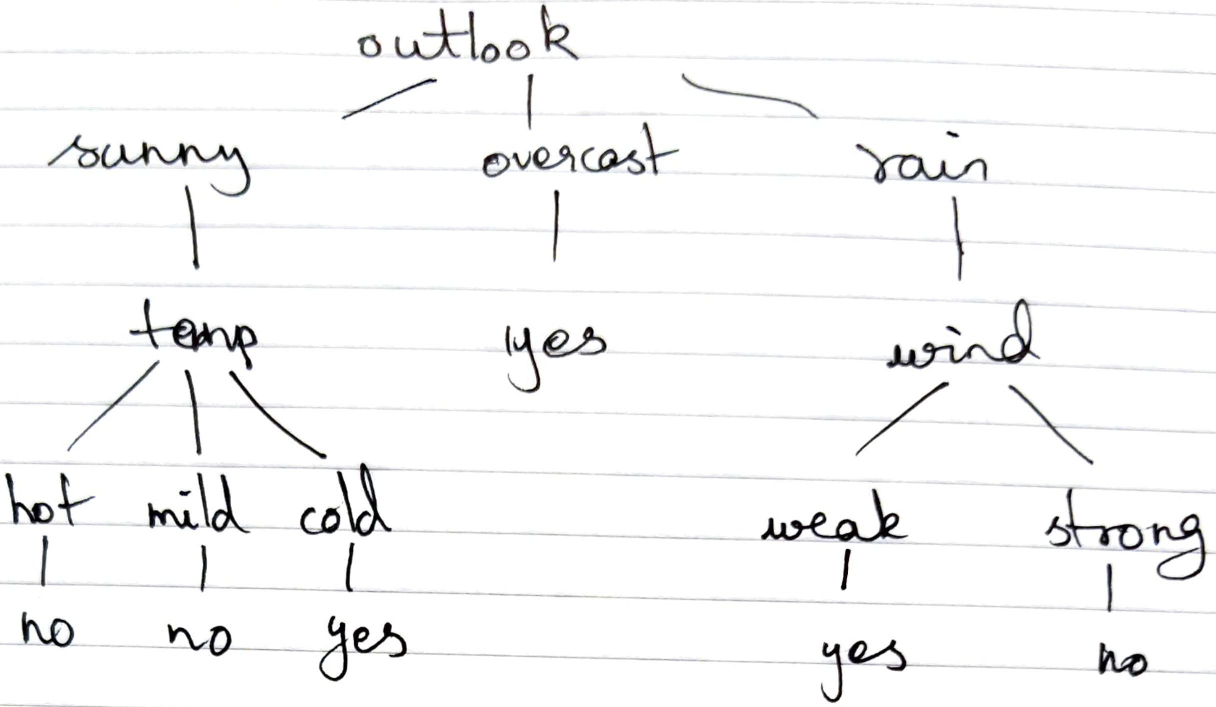
$$= 0.93$$

$$E_{strong}(3P, 1N) = 0$$

$$gain(sunny, windy) = 0.817 - \frac{3}{4} \times 0.93$$

$$= 0.1195$$

gain (sunny, humid) + gain (sunny, temp) will have info gain



Ans 4

$$P(\omega_1 | d_{11}(x)) = \frac{40}{40+30} = 0.57$$

$$P(\omega_1 | d_{21}(x)) = \frac{20}{40} = 0.5$$

$$P(\omega_1 | d_{32}(x)) = 0/10 = 0$$

$$\omega(\text{class } 1) = 0.57 \times 0.5 \times 0 = 0$$

$$P(\omega_2 | d_{11}(x)) = \frac{30}{30+40} = 0.428$$

$$P(\omega_2 | d_{21}(x)) = \frac{20}{40} = 0.5$$

$$P(\omega_2 | d_{32}(x)) = \frac{10}{10+0} = 1$$

$$\omega(\text{class } 2) = 0.48 \times 0.5 \times 1 = 0.214$$

$\therefore$ class 2 has higher preference