Name: Liangfan Pang CWID:10453333

1.

1） Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data. This tradeoff in complexity is a the bias-variance trade-off.

Practical techniques to reduce bias:

Increase hypothesis space (i.e., model complexity)

Practical techniques to reduce variance:

resampling (e.g., bagging models like random forest)

bias of each tree is the same as full model but higher variance

averaging many trees decreases variance without increasing bias

theoretically the more trees, the less variance (if no computation limit)

2） K-Fold cross-validation is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. Let's take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

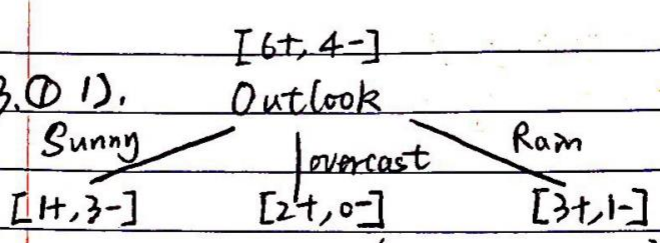
As there is never enough data to train your model, removing a part of it for validation poses a problem of underfitting. By reducing the training data, we risk losing important patterns/ trends in data set, which in turn increases error induced by bias. So, what we require is a method that provides ample data for training the model and also leaves ample data for validation. K Fold cross validation does exactly that.

2. ① $p = \frac{TP}{TP+FP} = \frac{50}{50+40} = \frac{5}{9}$ #

② $r = \frac{TP}{TP+FN} = \frac{50}{50+30} = \frac{5}{8}$

③ $F_1\text{-score} = \frac{2pr}{p+r} = \frac{2}{\frac{1}{p}+\frac{1}{r}} = \frac{2}{\frac{9}{5}+\frac{8}{5}} = \frac{34}{5} = 6.8$

3. ① 1).



$G(S,O) = E(S) - \frac{4}{10} \times E(\text{Sunny}) - \frac{2}{10} \times E(\text{overcast}) - \frac{4}{10} \times E(\text{Rain})$

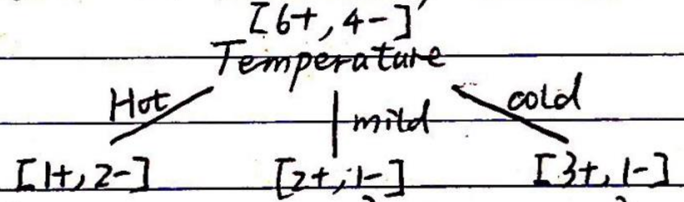
$E(S) = \frac{-6}{10} \times \log_2 \frac{6}{10} - \frac{4}{10} \times \log_2 \frac{4}{10}$ ~~= 0.971~~

~~E(high)~~ $E(\text{Sunny}) = \frac{-1}{4} \times \log_2 \frac{1}{4} - \frac{3}{4} \times \log_2 \frac{3}{4}$

$E(\text{Overcast}) = -1 \times \log_2 \frac{2}{2}$

$E(\text{Rain}) = \frac{-3}{4} \times \log_2 \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4}$

So $G(S,O) = 0.3219$



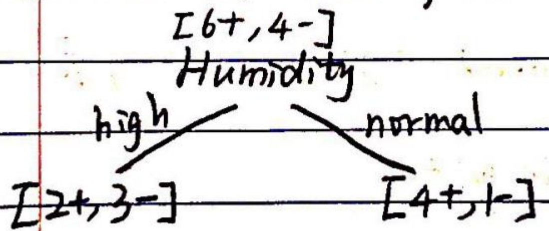
$G(S,T) = E(S) - \frac{3}{10} \times E(\text{Hot}) - \frac{3}{10} \times E(\text{mild}) - \frac{4}{10} \times E(\text{cold})$

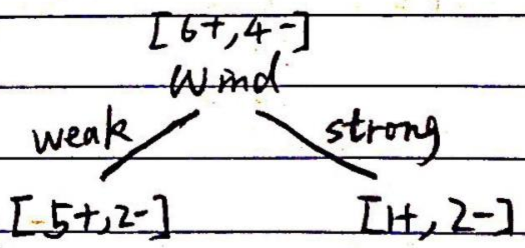
$E(\text{hot}) = \frac{-1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}$

$E(\text{mild}) = \frac{-2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3}$

$E(\text{cold}) = \frac{-3}{4} \times \log \frac{3}{4} - \frac{1}{4} \times \log_2 \frac{1}{4}$

So $G(S,T) = 0.0955$





$G(S,H) = E(S) - \frac{1}{2} \times E(high) - \frac{1}{2} \times E(normal)$

$E(high) = \frac{-2}{5} \times \log_2 \frac{2}{5} - \frac{3}{5} \times \log_2 \frac{3}{5}$

$E(normal) = \frac{-4}{5} \times \log_2 \frac{4}{5} - \frac{1}{5} \times \log_2 \frac{1}{5}$

So $G(S,H) = 0.1245$

$G(S,W) = E(S) - \frac{7}{10} \times E(weak) - \frac{3}{10} \times E(strong)$

$E(weak) = \frac{-5}{7} \times \log_2 \frac{5}{7} - \frac{2}{7} \times \log_2 \frac{2}{7}$

$E(strong) = \frac{-1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}$

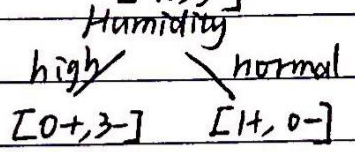
So $G(S,W) = 0.0913$

Since $G(S,O) > G(S,H) > G(S,T) > G(S,W)$

So the root node is Outlook.

2). $S_{Sunny} = \{D_1, D_2, D_8, D_9\}$ $S_{Rain} = \{D_4, D_5, D_6, D_{10}\}$ $S_{overcast} = \{D_3, D_7\}$

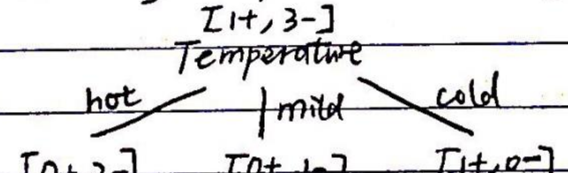
$G(S_{Sunny}, Humidity) = E(Sunny) - \frac{3}{4} \times E(high) - \frac{1}{4} \times E(normal) = E(Sunny) = 0.8113$



$E(high) = -0 \times \log_2 0 - 1 \times \log_2 \frac{3}{3} = 0$

$E(normal) = -1 \times \log_2 \frac{1}{1} - 0 \times \log_2 0 = 0$

$G(S_{sunny}, Temperature) = E(Sunny) - \frac{2}{4} \times E(hot) - \frac{1}{4} \times E(mild) - \frac{1}{4} \times E(cold) = 0.8113$

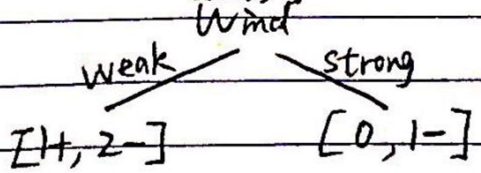


$E(hot) = 0$

$E(mild) = 0$

$E(cold) = 0$

$G(S_{sunny}, Wind) = E(Sunny) - \frac{3}{4} \times E(weak) - \frac{1}{4} \times E(strong) = 0.1226$

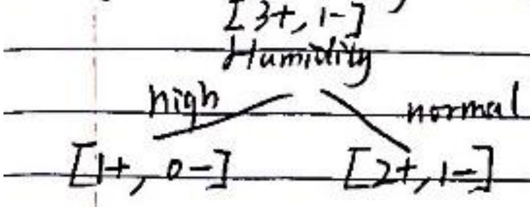


$E(weak) = -\frac{1}{3} \times \log_2 \frac{1}{3} - \frac{2}{3} \times \log_2 \frac{2}{3}$

$E(strong) = 0$

$\therefore G(S_{sunny}, H) = G(S_{sunny}, T) > G(S_{sunny}, W)$

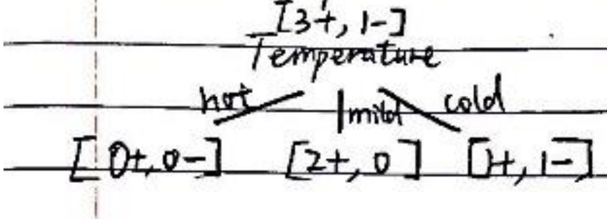
$G(S_{Rain}, Humidity) = E(Rain) - \frac{1}{4} \times E(high) - \frac{3}{4} \times E(normal) = 0.1226$



$E(high) = 0$

$E(normal) = -\frac{2}{3} \times \log_2 \frac{2}{3} - \frac{1}{3} \times \log_2 \frac{1}{3}$

$G(S_{Rain}, Temperature) = E(Rain) - 0 \times E(hot) - \frac{2}{4} \times E(mild) - \frac{2}{4} \times E(cold) = 0.3113$

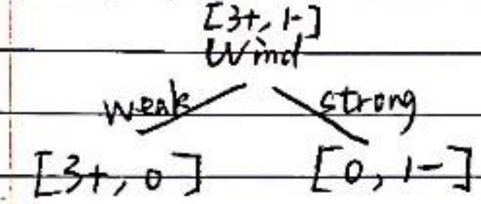


$E(mild) = 0$

$E(hot) = 0$

$E(cold) = -\frac{1}{2} \times \log_2 \frac{1}{2} - \frac{1}{2} \times \log_2 \frac{1}{2}$ #

$G(S_{Rain}, Wind) = E(Rain) - \frac{3}{4} \times E(weak) - \frac{1}{4} \times E(strong) = 0.8113$

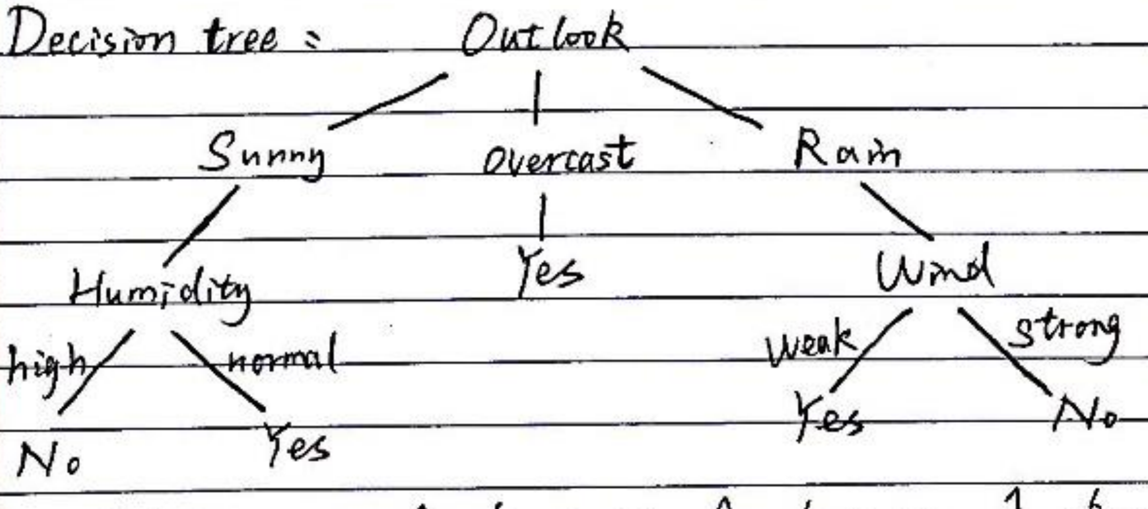


$E(weak) = 0$

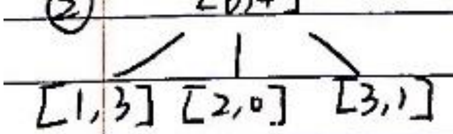
$E(strong) = 0$

So, $G(S_{Rain}, Wind) > G(S_{Rain}, Temperature) > G(S_{Rain}, Humidity)$.

Decision tree :



②



$\hat{p}_1 = \frac{6}{10} \times 4 = 2.4$ $\hat{p}_2 = \frac{6}{10} \times 2 = 1.2$ $\hat{p}_3 = \frac{6}{10} \times 4 = 2.4$

$\hat{n}_1 = \frac{4}{10} \times 4 = 1.6$ $\hat{n}_2 = \frac{4}{10} \times 2 = 0.8$ $\hat{n}_3 = \frac{4}{10} \times 4 = 1.6$

$$Q = \sum_{i} \frac{(p_i - \hat{p}_i)^2}{\hat{p}_i} + \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = 3.75$$

$df = 2$ Thus, $p(X^2 > CV) = 0.15$

4. The outputs of individual classifiers are 1 1 2

$p(w_1|d_{1,1}(x)=1) = \frac{40}{70}$ $\qquad$ $\hat{p}(w_2|d_{1,1}(x)=1) = \frac{30}{70}$

$\hat{p}(w_1|d_{2,1}(x)=1) = \frac{20}{40}$ $\qquad$ $\hat{p}(w_2|d_{2,1}(x)=1) = \frac{20}{40}$

$\hat{p}(w_1|d_{3,2}(x)=1) = \frac{0}{10}$ $\qquad$ $\hat{p}(w_2|d_{3,2}(x)=1) = \frac{10}{10}$

Thus, class 1 $= \frac{40}{70} \times \frac{20}{40} \times \frac{0}{10} = 0$

class 2 $= \frac{30}{70} \times \frac{20}{40} \times \frac{10}{10} = 0.21$

So the final decision is class 2.