HW3

Thursday, February 24, 2022 12:18 PM

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

The bias-variance trade-off is a common problem in machine learning models. Firstly, bias is the difference between the average prediction of the model & the actual value. Models with high bias lead to underfitting in which the model is not complex enough & end up with high training & testing error. Variance is the variability or the spread of model predictions. Models with high variance lead to overfitting in which the model is too complex & performs very well on the training data but not on the testing data.

Bias & variance are inversely correlated, meaning a model with high variance will have low bias & a model with low variance will have high bias. Hence, there is a tradeoff & a good middle ground must be found. If a model is too simple, it will likely underfit the data resulting in high bias & low variance. However, if a model is too complex, it will likely overfit the data resulting in low bias & high variance.

If a model is too simple & has high bias, the bias can be reduced by adding more parameters, adding more epochs, & overall making the model more complex. On the otherhand, if a model is too complex & has high variance, the variance can be reduced by removing parameters/features, adding regularization, & overall making the model less complex.

Question 2: [6 points] Assume the following confusion matrix of a classifier. Please compute its
1) precision,
2) recall, and
3) F_1-score.

Class 1 = +
Class 2 = −

Predicted results

Actual values	Class 1	Class 2
Class 1	50 TP	30 FN
Class 2	40 FP	60 TN

1) Precision

$$p = \frac{TP}{TP+FP} \rightarrow p = \frac{50}{50+40} = \frac{50}{90} = \boxed{0.556}$$

2) Recall

$$r = \frac{TP}{TP+FN} \rightarrow r = \frac{50}{50+30} = \frac{50}{80} = \boxed{0.625}$$

3) F_1-score

$$F_1 = \frac{2pr}{p+r} \rightarrow F_1 = \frac{2(0.556)(0.625)}{0.556+0.625} = \frac{0.695}{1.181} = \boxed{0.588}$$

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

yes = p
no = n

Entropy $\rightarrow H = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$

Conditional Entropy $\rightarrow H(S|A) = \sum \frac{p_i+n_i}{p+n} \cdot \text{Entropy}(A)$

Information Gain $\rightarrow \text{Gain}(S,A) = H(S) - H(S|A)$

To decide which attribute to split on first we can find the information gain of each & then split on the attribute with the highest.

$$H = \frac{-6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right) = 0.971$$

$$H(S|Outlook) = \frac{4}{10}\left(\frac{-1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) + \frac{2}{10}\left(\frac{-2}{2}\log_2\left(\frac{2}{2}\right)\right) + \frac{4}{10}\left(\frac{-3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.650$$

$$H(S|Temp) = \frac{3}{10}\left(\frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$$

$+ \frac{}{10}\left(\frac{}{4}\log_2\left(\frac{}{4}\right) - \frac{}{4}\log_2\left(\frac{}{4}\right)\right) = 0.650$

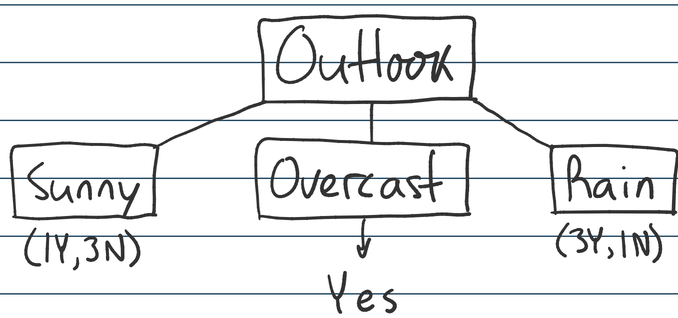
$H(S|Temp) = \frac{3}{10}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$
$+ \frac{3}{10}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)$
$+ \frac{4}{10}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.875$

$H(S|Humidity) = \frac{5}{10}\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$
$+ \frac{5}{10}\left(-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right) = 0.846$

$H(S|Wind) = \frac{7}{10}\left(-\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right)$
$+ \frac{3}{10}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.879$

$Gain(S, Outlook) = 0.971 - 0.650 = 0.321$
$Gain(S, Temp) = 0.971 - 0.875 = 0.096$
$Gain(S, Humidity) = 0.971 - 0.846 = 0.125$
$Gain(S, Wind) = 0.971 - 0.879 = 0.092$

Outlook has highest information gain so we will split on that first.



Need to find next best attribute to split on (highest information gain).

$H(Outlook=Sunny) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.811$

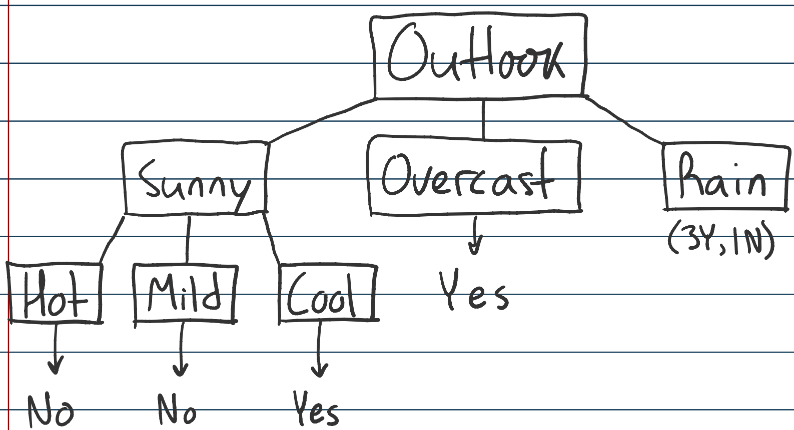
$H(S, Temp) = \frac{2}{4}\left(-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)$
$+ \frac{1}{4}\left(-\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right)$
$+ \frac{1}{4}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$

$H(S, Humidity) = \frac{3}{4}\left(-\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right)$
$+ \frac{1}{4}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$

$H(S, Wind) = \frac{3}{4}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$
$+ \frac{1}{4}\left(-\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0.689$

$Gain(S, Temp) = 0.811 - 0 = 0.811$
$Gain(S, Humidity) = 0.811 - 0 = 0.811$
$Gain(S, Wind) = 0.811 - 0.689 = 0.122$

Temperature & Humidity have highest information gain → choose either.



Only node left to split on is from Outlook = Rain.

$H(Outlook = Rain) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$

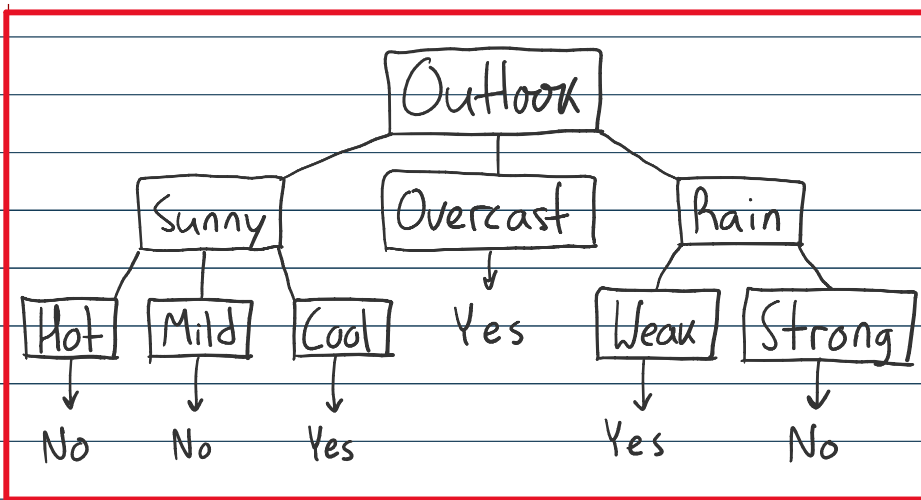
$$H(S, Temp) = \frac{2}{4}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{2}{4}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

$$H(S, Humidity) = \frac{1}{4}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) + \frac{3}{4}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) = 0.689$$

$$H(S, Wind) = \frac{3}{4}\left(-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right) + \frac{1}{4}\left(-\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0$$

$Gain(S, Temp) = 0.811 - 1 = -0.189$
$Gain(S, Humidity) = 0.811 - 0.684 = 0.122$
$Gain(S, Wind) = 0.811 - 0 = 0.811$

Wind has highest information gain so we will split on that.



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

Bayes $\rightarrow$ $P(\omega|x) = \frac{p(x|\omega)\,P(\omega)}{P(x)}$

$\mu_j(x) \propto \prod_{i=1}^{L} \hat{P}(\omega_j | d_{i,j}(x)=1)$

Classifier 1 : $\hat{P}(\omega_1 | d_{1,1}(x)=1) = \frac{40}{70}$, $\hat{P}(\omega_2 | d_{1,1}(x)=1) = \frac{30}{70}$

Classifier 2 : $\hat{P}(\omega_1 | d_{2,2}(x)=1) = \frac{20}{40}$, $\hat{P}(\omega_2 | d_{2,2}(x)=1) = \frac{20}{40}$

Classifier 3 : $\hat{P}(\omega_1 | d_{3,1}(x)=1) = \frac{0}{10}$, $\hat{P}(\omega_2 | d_{3,1}(x)=1) = \frac{10}{10}$

$\Rightarrow$ Class 1 $= \frac{40}{70} \cdot \frac{20}{40} \cdot \frac{0}{90} = 0$ $\Rightarrow$ $\boxed{\text{Class2}}$

Class 2 $= \frac{30}{70} \cdot \frac{20}{40} \cdot \frac{10}{10} = \underline{0.214}$