Week 2 Assignments

Topic: Linear regression, Decision trees, overfitting

1. In regression the output is

A) Discrete.

B) Continuous and always lies in a finite range.

[✓ **C) Continuous.**]

D) May be discrete or continuous.

2. In linear regression the parameters are

A) strictly integers

B) always lies in the range [0,1]

[✓ **C) any value in the real space**]

D) any value in the complex space

3. Which of the following is true for a decision tree?

A) Decision tree is an example of linear classifier.

[✓ **B) The entropy of a node typically decreases as we go down a decision tree.**]

C) Entropy is a measure of purity.

D) An attribute with lower mutual information should be preferred to other attributes.

Decision tree is not a linear classifier.

Entropy is a measure of impurity. Entropy reaches maximum value when all classes in the table are equally probable. And Entropy of a pure table (only one class) is zero.

4. Given a list of 14 examples including 9 positive and 5 negative examples. The entropy of the dataset with respect to this classification is

A) 0.940

B) 0.06

C) 0.50

D) 0.22

$$\text{Entropy} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.940.$$

5.

Outlook	Temperature	Humidity	Wind	Play tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	strong	No

The decision on whether tennis can be played or not is based on the following features: Outlook ∈ {Sunny, Overcast, Rain}, Temperature ∈ {Hot, Mild, Cool}, Humidity ∈ {High, Normal} and Wind ∈ {Weak, Strong}. The training data is given above.

5.1) The entropy of the entire dataset is

A) 1

B) 0.94

C) 0

D) 0.72

Entropy of the entire dataset

$$= -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$

$$= 0.530 + 0.409$$

$$= 0.939 = 0.94 \text{ (approx.)}$$

5.2) Which attribute will be the root of the decision tree and how much is the information gain due to the attribute.

[✓ **A) Outlook, 0.246**]

B) Humidity, 0.5

C) Temperature, 0.306

D) Humidity, 0.48

$$\text{Outlook}: \quad IG = 0.94 - \Big[\frac{5}{14}\Big(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\Big) + \frac{4}{14}\Big(-\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4}\Big)\Big]$$

$$= 0.246.$$

Using the same procedure check the Information Gain of other attributes. For temperature IG = 0.

Humidity = 0.151.

Wind = 0.048.

6. ISRO wants to discriminate between Martians (M) and Humans (H) based on the following features: Green ∈ {N,Y}, Legs ∈ {2,3}, Height ∈ {S,T}, Smelly ∈ {N,Y}. The training data is as follows:

Species	Green	Legs	Height	Smelly
M	N	3	S	Y
M	Y	2	T	N
M	Y	3	T	N
M	N	2	S	Y
M	Y	3	T	N
H	N	2	T	Y
H	N	2	S	N
H	N	2	T	N
H	Y	2	S	N
H	N	2	T	Y

6.1) Which attribute will be the root of the decision tree:

A) Green
✓ ~~B)~~ **Legs**
C) Height
D) Smelly

As, 'Legs' has the hieghest information gain (IG).

6.2) how much is the information gain due to the attribute found in the previous question?

A) 0.45
✓ ~~B)~~ **0.40**
C) 0.80
D) 0.70

The entropy of the entire data set

$$= -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$$

IG for Green : $1-\left[4/10\left(-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}\right)+\frac{6}{10}\left(-\frac{2}{6}\log_2\frac{2}{6}-\frac{4}{6}\log_2\frac{4}{6}\right)\right]$

$= 1-[0.3245+0.551] = 0.1245$

Legs: $1-\left[\frac{3}{10}\left(-\frac{3}{3}\log_2\frac{3}{3}-\frac{0}{3}\log_2\frac{0}{3}\right)+\frac{7}{10}\left(-\frac{2}{7}\log_2\frac{2}{7}-\frac{5}{7}\log_2\frac{5}{7}\right)\right]$

$= 1-[0+0.6] = 0.40$. (Maximum)

Height : $1-\left[\frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}\right)+\frac{6}{10}\left(-\frac{3}{6}\log_2\frac{3}{6}-\frac{3}{6}\log_2\frac{3}{6}\right)\right]$

$= 1-1=0$. For, Smelly IG $=0$.

7. The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute back-ache. Find the equation of the best fit line for this data.

Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

A) y = 3.39x + 11.62
B) y = 4.69x + 12.58
[✓ **C) y = 4.59x + 12.58**]
D) y = 3.59x + 10.58

For each x calculate the value of y using the given equations. Then calculate <u>error</u> for each equation. Equation with lowest error is the desired answer.

hints.

x	$y = 4.59x + 12.58$	actual y	error

8. Decision trees can be used for the following type of datasets:

I. The attributes are categorical
II. The attributes are numeric valued and continuous
III. The attributes are discrete valued numbers

A) In case I only
B) In case II only
C) In cases II and III only
[✓ **D) In cases I, II and III**]