ML CT1 Solution

1 Question 1

Consider a 3-class classification problem where your instances have two input features. You use linear classification with softmax. Suppose you get the following parameter values for the three classes after optimizing cross-entropy loss:

$$\Theta^1 = (1, 0, 4)$$

$$\Theta^2 = (0, 2, 3)$$

$$\Theta^3 = (1, 2, 0)$$

1.a

Give a rule to decide the probability distribution of belonging to the the classes for point $x \in \mathbf{R}^2$.

For a point $x = (x_1, x_2) \in \mathbf{R}^2$, the probability of belonging to a class c is given by:

$$p(y = c|x) = \frac{e^{\Theta_c^T x}}{e^{\Theta_1^T x} + e^{\Theta_2^T x} + e^{\Theta_3^T x}}$$

For a point $x = (x_1, x_2) \in \mathcal{R}^2$, the probabilities are:

$$p(y = 1|x) = \frac{e^{1+4x_2}}{e^{1+4x_2} + e^{2x_1+3x_2} + e^{1+2x_1}}$$

$$p(y = 2|x) = \frac{e^{2x_1+3x_2}}{e^{1+4x_2} + e^{2x_1+3x_2} + e^{1+2x_1}}$$

$$p(y = 3|x) = \frac{e^{1+2x_1}}{e^{1+4x_2} + e^{2x_1+3x_2} + e^{1+2x_1}}$$

1.b

Classify the following points: $(-1, 0), (1, 5), (5, 0)$.

For the point $(-1, 0)$, $\Theta_1^T x = 1 + 4x_2 = 1$, $\Theta_2^T x = 2x_1 + 3x_2 = -2$, $\Theta_3^T x = 1 + 2x_1 = -1$. This point belongs to Class **1**.

For the point $(1, 5)$, $\Theta_1^T x = 1 + 4x_2 = 21$, $\Theta_2^T x = 2x_1 + 3x_2 = 17$, $\Theta_3^T x = 1 + 2x_1 = 3$. This point belongs to Class **1**.

For the point $(5, 0)$, $\Theta_1^T x = 1 + 4x_2 = 1$, $\Theta_2^T x = 2x_1 + 3x_2 = 10$, $\Theta_3^T x = 1 + 2x_1 = 11$. This point belongs to Class **3**.

1.c Instances with the same probability for Class 1 and 2

Consider a point $x = (x_1, x_2) \in \mathbf{R}^2$

$$\Theta_1^T x = \Theta_2^T x$$

$$1 + 4x_2 = 2x_1 + 3x_2$$

$$2x_1 - x_2 = 1$$

All the points (x_1, x_2) on the straight line $2x_1 - x_2 = 1$ have the same probability for class 1 and 2.

1.d Instance(s) having same probability for all classes

Consider a point $x = (x_1, x_2) \in \mathbf{R}^2$, $\Theta_1^T x = \Theta_2^T x = \Theta_3^T x$

$$\Theta_1^T x = \Theta_2^T x$$

$$1 + 4x_2 = 2x_1 + 3x_2$$

$$2x_1 - x_2 = 1$$

$$\Theta_1^T x = \Theta_3^T x$$

$$1 + 4x_2 = 1 + 2x_1$$

$$x_1 = 2x_2$$

Solving the above two equations for x_1 and x_2, we get: $x_1 = 2/3, x_2 = 1/3$.
The point is $x = (2/3, 1/3)$

2 Question 2

The regularized loss function of regression in d variables is given by:

$$f(\theta) = \sum_{i=1}^{m} (y_i - \theta^T x_i)^2 + \lambda\Omega$$

2.a

Ω for Ridge (L2) regularization:

$$\Omega = \sum_{j=0}^{d} \theta_j^2$$

2.b Update Rule

$$f(\theta) = \sum_{i=1}^{m} (y_i - \theta^T x_i)^2 + \lambda \sum_{j=0}^{d} \theta_j^2$$

$$\frac{\partial f(\theta)}{\partial \theta_j} = -2 \sum_{i=1}^{m} (y_i - \theta^T x_i) x_{ij} + 2\lambda\theta_j$$

Update rule:

$$\theta_j \leftarrow \theta_j - \alpha(-2 \sum_{i=1}^{m} (y_i - \theta^T x_i) x_{ij} + 2\lambda\theta_j)$$

α is the learning rate.

2.c

As $\lambda \to \infty$, $\theta \to 0$

3 Question 3

Training Set

CGPA	Lab	Studied	Ace
L	F	F	F
L	F	T	T
M	T	F	F
M	F	T	T
L	T	F	T
H	T	T	T

3.a : Entropy at the Root Node

Given training set:

$$H(root) = -\left[\frac{4}{6} \times \log_2(\frac{4}{6}) + \frac{2}{6} \times \log_2(\frac{2}{6})\right]$$

Given: $\log_2(3) \approx 1.6$ and $\log_2(5) \approx 2.32$,

$$H(root) \approx 0.93$$

Entropy at the root node $= 0.93$

3.b Information Gain Calculation

1. Attribute: CGPA

Corresponding count for the attribute:

CGPA	T	F
L	2	1
M	1	1
H	0	1

$$H(root|CGPA) = -\frac{3}{6}\times\left[\frac{2}{3} \times \log_2(\frac{2}{3}) + \frac{1}{3} \times \log_2(\frac{1}{3})\right]-\frac{2}{6}\times\left[\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2})\right]-\frac{1}{6}\times 0 \approx 0.795$$

$$IG(root, CGPA) = 0.93 - 0.795 = 0.135$$

2. Attribute: Lab

Corresponding count for the attribute:

Lab	T	F
T	2	1
F	2	1

$$H(root|Lab) = -\frac{3}{6} \times \left[\frac{2}{3} \times \log_2(\frac{2}{3}) + \frac{1}{3} \times \log_2(\frac{1}{3})\right] - \frac{3}{6} \times \left[\frac{2}{3} \times \log_2(\frac{2}{3}) + \frac{1}{3} \times \log_2(\frac{1}{3})\right] \approx 0.93$$

$$IG(root, Lab) = 0.93 - 0.93 = 0$$

3. Attribute: Studied

Corresponding count for the attribute:

Studied	**T**	**F**
T	3	0
F	1	2

$$H(root|Studied) = -\frac{3}{6} \times \left[\frac{2}{3} \times \log_2(\frac{2}{3}) + \frac{1}{3} \times \log_2(\frac{1}{3})\right] + \frac{3}{6} \times 0 \approx 0.465$$

$$IG(root, Studied) = 0.93 - 0.465 = 0.465$$

Summary of Information Gain:
$IG(S, CGPA) = 0.135$
$IG(S, Lab) = 0$
$IG(S, Studied) = 0.465$

Root Attribute: Studied (Highest Information Gain of 0.465)

3.c Decision Tree Construction

After the first split at root node based on Studied attribute, Child nodes:

$Studied_T$	T	F
	3	0

$Studied_F$	T	F
	1	2

Now $Studied_T$ child node does not need any more split.
For $Studied_F$,

$$H(Studied_F) = -\left[\frac{1}{3} \times \log_2(\frac{1}{3}) + \frac{2}{3} \times \log_2(\frac{2}{3})\right] \approx 0.93$$

1. Attribute: CGPA

Corresponding count for the attribute:

CGPA	**T**	**F**
L	1	1
M	0	1
H	0	0

$$H(Studied_F|CGPA) = -\frac{2}{3} \times \left[\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2})\right] - \frac{1}{3} \times \left[0 + \frac{1}{1} \times \log_2(\frac{1}{1})\right] - \frac{1}{6} \times 0 \approx 0.67$$

$$IG(Studied_F, CGPA) = 0.93 - 0.67 = 0.26$$

2. Attribute: Lab

Corresponding count for the attribute:

Lab	**T**	**F**
T	1	1
F	0	1

$$H(Studied_F|Lab) = -\frac{2}{3} \times \left[\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2})\right] - \frac{1}{3} \times \left[0 + \frac{1}{1} \times \log_2(\frac{1}{1})\right] \approx 0.67$$

$$IG(root, Lab) = 0.93 - 0.67 = 0.26$$

As both of the attributes are giving same Information Gain, we choose randomly one of the attributes.

After the second split at $Studied_F$ node based on Lab attribute, Child nodes:

Lab_T	T	F
	1	1

Lab_F	T	F
	0	1

Next split based on CGPA attribute:

1. Attribute: CGPA

Corresponding count for the attribute:

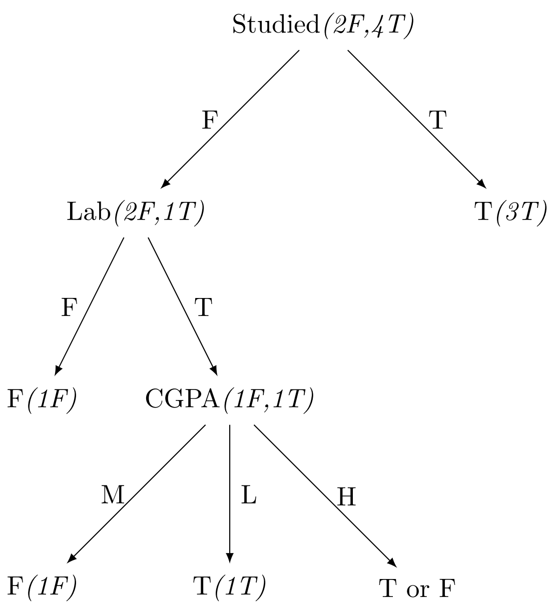
CGPA	**T**	**F**
L	1	0
M	0	1
H	0	0

After the third split at Lab_F node based on CGPA attribute, Child nodes:

$CGPA_L$	T	F
	1	0

$CGPA_M$	T	F
	0	1

Final Decision Tree:



Test Set

CGPA	Lab	Studied	Ace
L	T	T	T
H	T	F	F
M	F	T	T
H	F	F	F

3.d Test

Predictions:

$Datapoint_1$: T (Correctly classified)
$Datapoint_2$: Missing; T (Wrongly classified) or F (Correctly classified)
$Datapoint_3$: T (Correctly classified)
$Datapoint_4$: F (Correctly classified)

Accuracy: 75% or 100%