Open Questions:

1. Decision Tree (15 points)

You will be using a decision tree to classify whether an advertisement was clicked based on its size, position, and whether it played a sound.

1. Assume that Position is chosen for the root of the decision tree. What is the information gain associated with this attribute? (8 points)
2. Draw the full decision tree learned from this data (without any pruning). (7 points)

* $H(C) = -\frac{5}{10}\log\frac{5}{10} - \frac{5}{10}\log\frac{5}{10} = 1$

* $H(C \mid P = \text{Top}) = -\frac{0}{3}\log(\frac{0}{3}) - \frac{3}{3}\log(\frac{3}{3}) = 0$ [illegible] : $\frac{3}{10}$

* $H(C \mid P = \text{Middle}) = -\frac{2}{4}\log(\frac{2}{4}) - \frac{2}{4}\log(\frac{2}{4}) = 1$ [illegible] : $\frac{4}{10}$

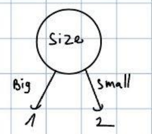
* $H(C \mid P = \text{bottom}) = -\frac{0}{3}\log\frac{0}{3} - \frac{3}{3}\log\frac{3}{3} = 0$ [illegible] : $\frac{3}{10}$

$IG(C \mid P) = 1 - \frac{3}{10}\cdot 0 - \frac{4}{10}\cdot 1 - \frac{3}{10}\cdot 0 = 0.6$

נתחיל מלבצע חישובים לכולם:

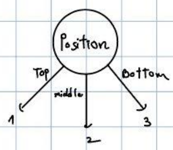
2.

	Leaf	N	p	H(p)
Big	1	5	$\frac{2}{5}, \frac{3}{5}$	$-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}$
Small	2	5	$\frac{3}{5}, \frac{2}{5}$	$-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$



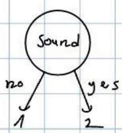
Total: $\frac{5}{10} \cdot 0.9709 + \frac{5}{10} \cdot 0.9709 = 0.9709$

	Leaf	N	p	H(p)
Top	1	3	$\frac{0}{3}, \frac{3}{3}$	0
middle	2	4	$\frac{2}{4}, \frac{2}{4}$	1
Bottom	3	3	$\frac{0}{3}, \frac{3}{3}$	0



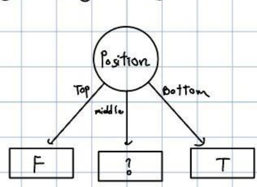
Total = $\frac{4}{10} \cdot 1 + 0 \cdot \frac{3}{10} + 0 \cdot \frac{3}{10} = \frac{4}{10}$

	Leaf	N	p	H(p)
no	1	6	$\frac{4}{6}, \frac{2}{6}$	$-\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0.9182$
yes	2	4	$\frac{1}{4}, \frac{3}{4}$	$-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.8112$



Total = $\frac{6}{10} \cdot 0.9182 + \frac{4}{10} \cdot 0.8112 = 0.8754$

נבחר את ה-node עם התוצאה הנמוכה ביותר

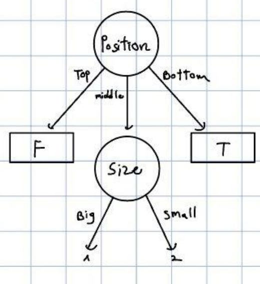


F	Small	Middle	Yes
F	Small	Middle	Yes
T	Small	Middle	No
T	Small	Middle	No

עכשיו נבחר את ה-node הבא:

נבצע את החישובים מחדש:

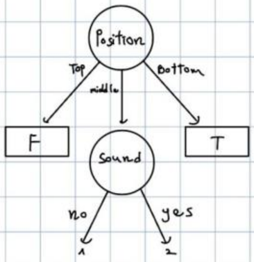
	Leaf	N	p	H(p)
Big	1	0	0	0
Small	2	4	$\frac{2}{4}, \frac{2}{4}$	$-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$

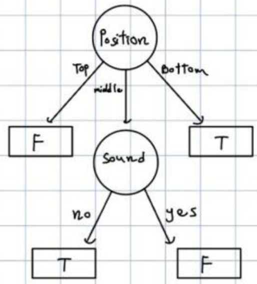


Total = $\frac{4}{4} \cdot 1 = 1$

	Leaf	N	P	H(P)
No	1	2	$\frac{2}{2}, \frac{0}{2}$	0
yes	2	2	$\frac{2}{2}, \frac{0}{2}$	0

Total = 0





לכן העץ של Sound:

2. Naive Base (10 points)

For the same data Using Naïve Base what is the prediction of the new Sample (big,Middle,No).

size	Big	small
F	3/5	2/5
T	2/5	3/5

position	Top	Middle	Bottom
F	3/5	2/5	0
T	0	2/5	3/5

sound	Yes	NO
F	3/5	2/5
T	1/5	4/5

New = (big, Middle, No)

$p(F) = 0.5 \qquad p(T) = 0.5$

$$p(F \mid New) = \frac{p(F) \cdot p(size = big \mid F) \cdot p(position = Middle \mid F) \cdot p(sound = No \mid F)}{p(New)} =$$

$$= 0.5 \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{2}{5} = 0.048$$

$$p(T \mid New) = \frac{p(T) \cdot p(size = big \mid T) \cdot p(position = Middle \mid T) \cdot p(sound = No \mid T)}{p(New)} =$$

$$= 0.5 \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} = 0.064$$

$p(T \mid New) > p(F \mid New)$

class (New) = T לכן

3. Understanding (16 points).

1. Describe the analytical solution for linear regression with MSE as a distance function. (4 points)
2. What is the problem with information gain? Describe any solution for it. (4 points)
3. Why do we use Gradient Descent or Neotun Roffson for Linear Regression? (4 points)
4. Explain how a Decision tree is used for regression problems. (4 points)

Answers:

1. The analytical solution for linear regression with Mean Squared Error (MSE) as the distance function involves finding the parameters (coefficients) that minimize the MSE between the predicted values and the actual values in the dataset. To minimize the MSE - MSE=1/N*∑ (Yi−Yi^)^2, we take the derivative of the loss function with respect to the coefficients and set it to zero. Solving these equations will give the analytical solution for the coefficients.
2. Problem: Information Gain in decision trees tends to favor attributes with more unique values as they can potentially create more splits, leading to overfitting.

   Solution: One solution is to use the Gain Ratio instead of Information Gain. Gain Ratio normalizes the information gain by considering the intrinsic information of a split, penalizing attributes with many values. It is calculated as the ratio of Information Gain to Split Information.
3. Gradient Descent: Gradient Descent is an iterative optimization algorithm used for finding the minimum of a function. In linear regression, it updates the parameters θ in the opposite direction of the gradient of the cost function with respect to θ. It is effective for large datasets and is suitable when the analytical solution is computationally expensive.

   Newton-Raphson: Newton-Raphson is an optimization algorithm that uses the second derivative (Hessian matrix) of the cost function. It converges faster than gradient descent but may be computationally expensive for large datasets due to the calculation of the Hessian matrix and its inverse.
4. A decision tree for regression predicts continuous values instead of discrete classes. The tree is built by recursively splitting the dataset based on features to minimize the variance of the target variable within each leaf node. The prediction for a new data point is the average of the target values in the leaf node it belongs to. The splitting criterion is often based on reducing the variance of the target variable. The process continues until a stopping criterion is met, such as a maximum depth or a minimum number of samples per leaf.