COMS 4701 - Homework 4 - Written

Sankalp, Apharande spa2138

November 22, 2021

Question 1

Answer:

1. $GINI = 1 - p_+^2 - p_-^2$

   $GAIN(S, A) = Gini(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Gini(S_v)$

   1. $P(TRUE) = \frac{3}{4}$

      $P(FALSE) = \frac{1}{4}$

      GINI index (Root) $= 1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = \frac{6}{16} = 0.375$
   2. Gain(root, word1) $= Gini(root) - \frac{2}{4} \cdot 0 - \frac{2}{4} \cdot 0.5$
      Gain(root, word1) $= 0.375 - \frac{2}{4} \cdot 0 - \frac{2}{4} \cdot 0.5$
      Gain(root, word1) $= 0.125$

      Gain(root, word2) $= Gini(root) - \frac{1}{4} \cdot 0 - \frac{3}{4} \cdot \frac{4}{9}$

      Gain(root, word2) $= 0.375 - \frac{1}{4} \cdot 0 - \frac{3}{4} \cdot \frac{4}{9}$

      Gain(root, word2) $= 0.375 - \frac{1}{3} = 0.0416$
      $Gain(root, word1) > Gain(root, word2)$

      Hence word1 will be picked up as a feature at the root. Because Gain is maximum gain while splitting.

Question 2

1. **Answer:**
   $f(x)$ is predicted final score if midterm score is x.

   $f(x) = \beta_0 + \beta_1 \cdot x$
   $\beta_0 = -8$ and $\beta_1 = 1.2$
   $f(x) = -8 + 1.2 \cdot x$
   $f(80) = 1.2 \cdot 80 - 8 = 88$
   The predicted final score is 88.
2. $(x_1, y_1) = (55, 67)$
   $(x_1, y_1) = (60, 63)$
   $(x_1, y_1) = (66, 72)$
   $(x_1, y_1) = (72, 90)$
   $(x_1, y_1) = (85, 93)$
   $(x_1, y_1) = (90, 92)$
   $\beta_0 = -8$ and $\beta_1 = 1.2$

   $f(x) = 1.2 \cdot x - 8$

   $f(x_1) = 1.2 \cdot 55 - 8 = 58$
   $y_{pred} = 58$ & $y_{actual} = 67$

   $f(x_1) = 1.2 \cdot 60 - 8 = 64$
   $y_{pred} = 64$ & $y_{actual} = 63$

   $f(x_1) = 1.2 \cdot 66 - 8 = 71.2$
   $y_{pred} = 71.2$ & $y_{actual} = 72$

   $f(x_1) = 1.2 \cdot 72 - 8 = 78.4$
   $y_{pred} = 78.4$ & $y_{actual} = 90$

   $f(x_1) = 1.2 \cdot 85 - 8 = 94$
   $y_{pred} = 94$ & $y_{actual} = 93$

   $f(x_1) = 1.2 \cdot 90 - 8 = 100$
   $y_{pred} = 100$ & $y_{actual} = 92$

   $n = 6$ $Cost = \frac{1}{2n} \cdot \sum_{i=1}^{n} (y_i + 8 - 1.2 \cdot x_i)^2$
   $Cost = \frac{1}{12} \cdot \sum_{i=1}^{n} (y_{actual} - y_{predicted})^2$
   $Cost = \frac{1}{12} \cdot [(58 - 67)^2 + (64 - 63)^2 + (71.2 - 72)^2 + (78.4 - 90)^2 + (94 - 93)^2 + (100 - 92)^2]$
   $Cost = \frac{1}{12} \cdot [282.2] = 23.51666667$
   $Cost = 23.51666667$
3. a. Incorrect.
      Because if we have β_0 and β_1 zero, then equation becomes zero. i.e. $f(x) = \beta_0 + \beta_1 \cdot x = 0$.
      i.e $y_{pred} = 0$
      But $Cost = \frac{1}{2n} \cdot \sum_{i=1}^{n} (y_{actual} - y_{predicted})^2 = \frac{1}{2n} \cdot \sum_{i=1}^{n} (y_{actual})^2$ may not be zero.
   b. Correct.
      The linear regressor perfectly fit the data. Because $R = Cost = \frac{1}{12} \cdot \sum_{i=1}^{n} (y_{actual} - y_{predicted})^2$.
      It will be zero if and only if each term is zero. Hence our reggressor perfectly fit the data.
   c. Incorrect.
      We can't do a perfect prediction test set because its highly possible that our model has over-fitted the training set. And it is highly possible that it will give completely wrong prediction on new data point.

Question 3

Answer

$P(TRUE) = \frac{6}{10} = 0.6$

$P(FALSE) = \frac{4}{10} = 0.4$

$P(A|TRUE) = \frac{2}{6} = \frac{1}{3} = 0.3333$

$P(A|FALSE) = \frac{3}{4} = 0.75$

$P(\neg B|TRUE) = \frac{1}{6} = 0.166$

$P(\neg B|FALSE) = \frac{2}{4} = 0.5$

$P(C|TRUE) = \frac{3}{6} = \frac{1}{2} = 0.5$

$P(C|FALSE) = \frac{4}{4} = 1$

$P(TRUE) \cdot P(A|TRUE) \cdot P(\neg B|TRUE) \cdot P(C|TRUE) = 0.6 \cdot 0.333 \cdot 0.166 \cdot 0.5 = 0.016667$

$P(FALSE) \cdot P(A|FALSE) \cdot P(\neg B|FALSE) \cdot P(C|FALSE) = 0.4 \cdot 0.75 \cdot 0.5 \cdot 1 = 0.15$

We assign the label which maximises the $p(y) \prod_j p(a_j|y)$
Here P(new$_label = True) < P(new_label = False)$
Hence the label for $(A = 1, B = 0, C = 1)$ **is** $FALSE$