Group 37 – Enzo Nunes 106336, Tomás Barros 106588

Machine Learning - Homework1

Ex1 – Correlation

Pearson Correlation:

$$r = \frac{\sum(x_1 - \overline{x}_1)(x_2 - \overline{x}_2)}{\sqrt{\sum(x_1 - \overline{x}_1)^2 \sum(x_2 - \overline{x}_2)^2}}$$

Spearman's rank Correlation (general case):

$$\rho = \frac{\sum(R[x_1] - \overline{R[x_1]})(R[x_2] - \overline{R[x_2]})}{\sqrt{\sum(R[x_1] - \overline{R[x_1]})^2 \sum(R[x_2] - \overline{R[x_2]})^2}}$$

a)

Variable sets:

$$x_1 = (-4, -2, 0, 2, 4), \qquad R[x_1] = (1, 2, 3, 4, 5)$$
$$x_2 = f(x_1) = 0.25x_1 = (-1, -0.5, 0, 0.5, 1), \qquad R[x_2] = (1, 2, 3, 4, 5)$$

Calculate mean values:

$$\overline{x}_1 = 0, \qquad \overline{R[x_1]} = 3$$
$$\overline{x}_2 = 0, \qquad \overline{R[x_2]} = 3$$

Calculate deviations:

$$x_1 - \overline{x}_1 = (-4, -2, 0, 2, 4), \qquad R[x_1] - \overline{R[x_1]} = (-2, -1, 0, 1, 2)$$
$$x_2 - \overline{x}_2 = (-1, -0.5, 0, 0.5, 1), \qquad R[x_2] - \overline{R[x_2]} = (-2, -1, 0, 1, 2)$$

Calculate products:

$$(x_1 - \overline{x}_1)(x_2 - \overline{x}_2) = (4, 1, 0, 1, 4), \qquad \left(R[x_1] - \overline{R[x_1]}\right)\left(R[x_2] - \overline{R[x_2]}\right) = (4, 1, 0, 1, 4)$$
$$(x_1 - \overline{x}_1)^2 = (16, 4, 0, 4, 16), \qquad \left(R[x_1] - \overline{R[x_1]}\right)^2 = (4, 1, 0, 1, 4)$$
$$(x_2 - \overline{x}_2)^2 = (1, 0.25, 0, 0.25, 1), \qquad \left(R[x_2] - \overline{R[x_2]}\right)^2 = (4, 1, 0, 1, 4)$$

Calculate sums:

$$\sum(x_1 - \overline{x}_1)(x_2 - \overline{x}_2) = 10, \qquad \sum\left(R[x_1] - \overline{R[x_1]}\right)\left(R[x_2] - \overline{R[x_2]}\right) = 10$$
$$\sum(x_1 - \overline{x}_1)^2 = 40, \qquad \sum\left(R[x_1] - \overline{R[x_1]}\right)^2 = 10$$
$$\sum(x_2 - \overline{x}_2)^2 = 2.5, \qquad \sum\left(R[x_2] - \overline{R[x_2]}\right)^2 = 10$$

Final values:

$$r = \frac{10}{\sqrt{40 \times 2.5}} = 1$$
$$\rho = \frac{10}{\sqrt{10 \times 10}} = 1$$

The Pearson and Spearman correlations are both 1 because x_2 is a linear transformation of x_1 through $f(x) = 0.25x$. This preserves both the perfect linear relationship (Pearson) and the rank order (Spearman), resulting in both correlations being 1.

b)

Variable sets:

$$x_1 = (-4, -2, 0, 2, 4), \qquad R[x_1] = (1, 2, 3, 4, 5)$$
$$x_2 = f(x_1) = \begin{cases} 1, & x_1 \geq 0 \\ 0, & x_1 < 0 \end{cases} = (0, 0, 1, 1, 1), \qquad R[x_2] = (1.5, 1.5, 3, 4, 5)$$

Calculate mean values:

$$\overline{x}_1 = 0, \qquad \overline{R[x_1]} = 3$$

$$\overline{x}_2 = 0.6, \qquad \overline{R[x_2]} = 3$$

Calculate deviations:

$$x_1 - \overline{x}_1 = (-4, -2, 0, 2, 4), \qquad R[x_1] - \overline{R[x_1]} = (-2, -1, 0, 1, 2)$$

$$x_2 - \overline{x}_2 = (-0.6, -0.6, 0.4, 0.4, 0.4), \qquad R[x_2] - \overline{R[x_2]} = (-1.5, -1.5, 0, 1, 2)$$

Calculate products:

$$(x_1 - \overline{x}_1)(x_2 - \overline{x}_2) = (2.4, 1.2, 0, 0.8, 1.6), \qquad \left(R[x_1] - \overline{R[x_1]}\right)\left(R[x_2] - \overline{R[x_2]}\right) = (3, 1.5, 0, 1, 4)$$

$$(x_1 - \overline{x}_1)^2 = (16, 4, 0, 4, 16), \qquad \left(R[x_1] - \overline{R[x_1]}\right)^2 = (4, 1, 0, 1, 4)$$

$$(x_2 - \overline{x}_2)^2 = (0.36, 0.36, 0.16, 0.16, 0.16), \qquad \left(R[x_2] - \overline{R[x_2]}\right)^2 = (2.25, 2.25, 0, 1, 4)$$

Calculate sums:

$$\sum (x_1 - \overline{x}_1)(x_2 - \overline{x}_2) = 6, \qquad \sum \left(R[x_1] - \overline{R[x_1]}\right)\left(R[x_2] - \overline{R[x_2]}\right) = 9.5$$

$$\sum (x_1 - \overline{x}_1)^2 = 40, \qquad \sum \left(R[x_1] - \overline{R[x_1]}\right)^2 = 10$$

$$\sum (x_2 - \overline{x}_2)^2 = 1.2, \qquad \sum \left(R[x_2] - \overline{R[x_2]}\right)^2 = 9.5$$

Final values:

$$r = \frac{6}{\sqrt{40 \times 1.2}} \simeq 0.866$$

$$\rho = \frac{9.5}{\sqrt{10 \times 9.5}} \simeq 0.866$$

In this case, both Pearson and Spearman correlations result in the same value because the unit step function preserves the overall structure of the relationship between x_1 and x_2. Pearson still detects a consistent relationship between negative and non-negative values of x_1 and x_2, while Spearman reflects the monotonic nature of the transformation, as the rank order of x_1 is preserved.

c)

Variable sets:

$$x_1 = (-4, -2, 0, 2, 4), \qquad R[x_1] = (1, 2, 3, 4, 5)$$
$$x_2 = f(x_1) = \frac{1}{1 + e^{-x_1}} \simeq (0.018, 0.119, 0.5, 0.881, 0.982), \qquad R[x_2] = (1, 2, 3, 4, 5)$$

Calculate mean values:

$$\overline{x}_1 = 0, \qquad \overline{R[x_1]} = 3$$
$$\overline{x}_2 = 0.5, \qquad \overline{R[x_2]} = 3$$

Calculate deviations:

$$x_1 - \overline{x}_1 = (-4, -2, 0, 2, 4), \qquad R[x_1] - \overline{R[x_1]} = (-2, -1, 0, 1, 2)$$
$$x_2 - \overline{x}_2 = (-0.482, -0.381, 0, 0.381, 0.482), \qquad R[x_2] - \overline{R[x_2]} = (-2, -1, 0, 1, 2)$$

Calculate products:

$$(x_1 - \overline{x}_1)(x_2 - \overline{x}_2) = (1.928, 0.762, 0, 0.762, 1.928), \qquad \left(R[x_1] - \overline{R[x_1]}\right)\left(R[x_2] - \overline{R[x_2]}\right) = (4, 1, 0, 1, 4)$$
$$(x_1 - \overline{x}_1)^2 = (16, 4, 0, 4, 16), \qquad \left(R[x_1] - \overline{R[x_1]}\right)^2 = (4, 1, 0, 1, 4)$$
$$(x_2 - \overline{x}_2)^2 \simeq (0.232, 0.145, 0, 0.145, 0.232), \qquad \left(R[x_2] - \overline{R[x_2]}\right)^2 = (4, 1, 0, 1, 4)$$

Calculate sums:

$$\sum (x_1 - \overline{x}_1)(x_2 - \overline{x}_2) = 5.38, \qquad \sum \left(R[x_1] - \overline{R[x_1]}\right)\left(R[x_2] - \overline{R[x_2]}\right) = 10$$
$$\sum (x_1 - \overline{x}_1)^2 = 40, \qquad \sum \left(R[x_1] - \overline{R[x_1]}\right)^2 = 10$$
$$\sum (x_2 - \overline{x}_2)^2 = 0.754, \qquad \sum \left(R[x_2] - \overline{R[x_2]}\right)^2 = 10$$

Final values:

$$r = \frac{5.38}{\sqrt{40 \times 0.754}} \simeq 0.98$$
$$\rho = \frac{10}{\sqrt{10 \times 10}} = 1$$

In this last case, the Pearson and Spearman correlation values are different, even though they are close to each other, because x_2 is not a linear transformation of x_1, as the function, $f(x)$, used in this case does not represent a linear transformation. Even though the function is monotonically increasing, which causes a positive Spearman correlation, the Pearson correlation does not capture the strength of this increasement accurately, which results in a disparity between both values.

Ex2 – Decision Trees

Entropy:

$$\mathrm{E}(p_1, p_2 \dots p_n) = \sum_{i=1}^{n} -p_i \log_2(p_i)$$

Information Gain:

$$IG(What\ to\ do \mid X) = E(What\ to\ do) - E(What\ to\ do \mid X)$$

a)

Entropy Calculations:

$$\begin{aligned}\mathrm{E}(What\ to\ do) &= -p(Go\ for\ a\ walk)\log_2\big(p(Go\ for\ a\ walk)\big) - p(TV)\log_2\big(p(TV)\big)\\ &\quad - p(Reading)\log_2\big(p(Reading)\big)\end{aligned}$$

$$E(What\ to\ do) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) \cong 1.5219\ bits$$

$$\begin{aligned}&E(What\ to\ do \mid Weekend)\\ &\quad = p(Yes) \times E(Go\ for\ a\ walk,\ Reading) + p(No) \times E(TV,\ Reading,\ Go\ for\ a\ walk)\\ &\quad = \frac{2}{5}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{3}{5}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)\\ &\quad = 1.3510\ bits\end{aligned}$$

$$\begin{aligned}&E(What\ to\ do \mid Weather)\\ &\quad = p(Sunny) \times E(Go\ for\ a\ walk) + p(Rain) \times E(TV, Reading)\\ &\quad + p(Cloudy) \times E(Go\ for\ a\ walk,\ Reading)\\ &\quad = \frac{1}{5}(-1\log 2\,(1)) + \frac{2}{5}\left(-\frac{1}{2}\log 2\left(\frac{1}{2}\right) - \frac{1}{2}\log 2\left(\frac{1}{2}\right)\right) + \frac{2}{5}\left(-\frac{1}{2}\log 2\left(\frac{1}{2}\right) - \frac{1}{2}\log 2\left(\frac{1}{2}\right)\right)\\ &\quad = 0.8\ bits\end{aligned}$$

$$\begin{aligned}&E(What\ to\ do \mid Tired)\\ &\quad = p(Yes) \times E(TV,\ Reading) + p(No) \times E(Go\ for\ a\ walk,\ Reading,\ Go\ for\ a\ walk)\\ &\quad = \frac{2}{5}\left(-\frac{1}{2}\log 2\left(\frac{1}{2}\right) - \frac{1}{2}\log 2\left(\frac{1}{2}\right)\right) + \frac{3}{5}\left(-\frac{2}{3}\log 2\left(\frac{2}{3}\right) - \frac{1}{3}\log 2\left(\frac{1}{3}\right)\right) = 0.9510\ bits\end{aligned}$$

Information Gain Calculations:

$$IG(What\ to\ do \mid Weekend) = E(What\ to\ do) - E(What\ to\ do \mid Weekend) = 1.5219 - 1.3510$$

$$IG(What\ to\ do \mid Weather) = E(What\ to\ do) - E(What\ to\ do \mid Weather) = 1.5219 - 0.8$$

$$IG(What\ to\ do \mid Tired) = E(What\ to\ do) - E(What\ to\ do \mid Tired) = 1.5219 - 0.9510$$

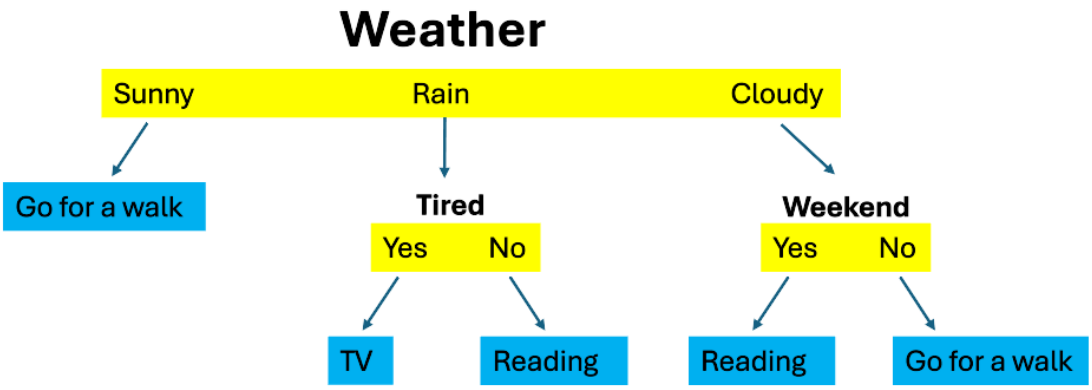
Final Conclusions:

$$E(What\ to\ do \mid Weekend) > E(What\ to\ do \mid Tired) > E(What\ to\ do \mid Weather)$$

$$IG(What\ to\ do \mid Weather) > IG(What\ to\ do \mid Tired) > IG(What\ to\ do \mid Weekend)$$

The root of the tree, according to the ID3 algorithm, should be Weather.

b)



c)

		TRUE		
		Reading	*TV*	*Go for a Walk*
Predicted	*Reading*	0	0	0
	TV	1	1	0
	Go for a walk	2	0	1

Ex3 – Software Experiments

Group Number (used for the RNG): 37

a)

- For the train_size value 0.1, we have a depth of 2 and an accuracy of 0.73.
- For the train_size value 0.9, we have a depth of 3 and an accuracy of 0.89.

We have a higher accuracy for the value 0.9 because this means we are using 90% of the data to train the model and only 10% for testing. This means the model has way more examples to learn from, which usually results in better performance. Also, as we have more training data, the decision tree needs more depth to capture more intricate patterns in the data.

b)

- For the train_size value 0.9, without the stratify option, we have a depth of 5 and an accuracy of 0.83.

When we remove the stratify option from the train_test_split function call, the accuracy of the model decreases. This is because the split between the data is random without stratification, which results in a risk that some classes may be underrepresented or overrepresented in the training or testing sets. Also, if the training set does not adequately represent the distribution of the classes, the model may not learn the patterns effectively, leading to poor generalization and lower accuracy on the test set.