Machine Learning Report

Homework I - Correlation and Decision Trees

ist1114964 - Axel Carapinha
ist1106565 - Martim Gordino

September 26, 2024

Contents

1 Correlation

1.1 a)

① a) $x_1 = (-4, -2, 0, 2, 4)$ $\quad \overline{x_1} = 0$

$x_2 = (-1, -0.5, 0, 0.5, 1)$ $\quad \overline{x_2} = 0$

$f(-4) = 0.25 \times (-4) = -1$

$f(-2) = 0.25 \times (-2) = -0.5$

$f(0) = 0$

$f(2) = 0.25 \times 2 = 0.5$

$f(4) = 0.25 \times 4 = 1$

Correlation (Pearson's)

$$PCC(x_1, x_2) = \frac{\sum_{i=1}^{m} (a_{i1} - \overline{x_1})(a_{i2} - \overline{x_2})}{\sqrt{\sum_{i=1}^{m} (a_{i1} - \overline{x_1})^2} \sqrt{\sum_{i=1}^{m} (a_{i2} - \overline{x_2})^2}} =$$

$$= \frac{(-4)(-1) + (-2)(-0.5) + (0)(0) + (2)(0.5) + (4)(1)}{\sqrt{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2} \sqrt{(-1)^2 + (-0.5)^2 + 0^2 + (0.5)^2 + 1^2}} = 1$$

Spearman's Rank

x_1	rank	x_2	rank
-4	1	-1	1
-2	2	-0.5	2
0	3	0	3
2	4	0.5	4
4	5	1	5

→ we calculate spearman's rank directly by the formula because there were no ties in the ranks

$$\text{Spearman}(x_1, x_2) = PCC(R(x_1), R(x_2)) \Leftrightarrow$$

$$\Leftrightarrow \text{Spearman}(x_1, x_2) = \frac{\sum_{i=1}^{m} (a_{i1} - \overline{R(x_1)})(a_{i2} - \overline{R(x_2)})}{\sqrt{\sum_{i=1}^{m} (a_{i1} - \overline{R(x_1)})^2} \sqrt{\sum_{i=1}^{m} (a_{i2} - \overline{R(x_2)})^2}}$$

$$\overline{R(x_1)} = \frac{5+4+3+2+1}{5} = 3 = \overline{R(x_2)}$$

$$R(x_2) = R(x_1) = [1,2,3,4,5]$$

$$\text{Spearman}\,(x_1, x_2) = \frac{\cancel{\sum_{i=1}^{n} (a_{i1} - \overline{R(x_1)})^2}}{\cancel{\left[\sqrt{\sum_{i=1}^{n} (a_{i1} - \overline{R(x_1)})^2}\right]^2}} = 1$$

Both correlations are equal because a linear function is also a monotonic function, so both correlations described the linearity of the data, by approximating (perfectly) with a linear function (Pearson), and with a monotonic function (Spearman's).

1.2 b)

1 b) $x_1 = (-4, -2, 0, 2, 4)$ $\quad x_2 = (0, 0, 1, 1, 1)$

$\overline{x_1} = 0$ $\quad \overline{x_2} = 0.6$

$-4 < 0 \Rightarrow f(-4) = 0$

$-2 < 0 \rightarrow f(-2) = 0$

$0 = 0 \Rightarrow f(0) = 1$

$2 > 0 \Rightarrow f(2) = 1$

$4 > 0 \Rightarrow f(4) = 1$

Pearson's Correlation

$$PCC(x_1, x_2) = \frac{\sum_{i=1}^{n} (a_{i1} - \overline{x_1})(a_{i2} - \overline{x_2})}{\sqrt{\sum_{i=1}^{n} (a_{i1} - \overline{x_1})^2}\sqrt{\sum_{i=1}^{n} (a_{i2} - \overline{x_2})^2}} =$$

$$= \frac{(-4)(-0.6) + (-2)(-0.6) + (0)(0.4) + (2)(0.4) + (4)(0.4)}{\sqrt{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2} \cdot \sqrt{(-0.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (0.4)^2}} =$$

$= 0.866$

Spearman's Correlation

x_1	rank	x_2	rank
-4	1	0	1.5
-2	2	0	1.5
0	3	1	4
2	4	1	4
4	5	1	4

$$\overline{R(x_1)} = \frac{1+2+3+4+5}{5} = 3 \qquad \overline{R(x_2)} = \frac{1.5+1.5+4+4+4}{5} = 3$$

$$\text{Spearman}(x_1, x_2) = \frac{\sum_{i=1}^{m}(a_{i1}-3)(a_{i2}-3)}{\sqrt{\sum_{i=1}^{m}(a_{i1}-3)^2}\sqrt{\sum_{i=1}^{m}(a_{i2}-3)^2}} =$$

$$= \frac{(1-3)(1.5-3)+(2-3)(1.5-3)+(3-3)(4-3)+(4-3)(4-3)+(5-3)(4-3)}{\sqrt{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}\cdot\sqrt{(1.5-3)^2+(1.5-3)^2+3\cdot(4-3)^2}} =$$

$$= 0.866$$

→ Both ~~cross~~ correlation values are similar because the monotonicity of the sign() function is consistent with the linear correlation between data, creating a synchrony between these two behaviours. However, the linearity and the monotonicity are not perfect, leading to a worse value than in a).

1.3 c)

$x_1 = (-4, -2, 0, 2, 4)$

$\bar{x}_1 = \frac{-4+(-2)+0+2+4}{5} = 0$

$x_2 = \left(\frac{1}{1+e^4}, \frac{1}{1+e^2}, \frac{1}{2}, \frac{1}{1+e^{-2}}, \frac{1}{1+e^{-4}}\right)$

$\bar{x}_2 = \frac{\frac{1}{1+e^4} + \frac{1}{1+e^2} + \frac{1}{2} + \frac{1}{1+e^{-2}} + \frac{1}{1+e^{-4}}}{5} = \frac{1}{2}$

$f(-4) = \frac{1}{1+e^4}$

$f(-2) = \frac{1}{1+e^2}$

$f(0) = \frac{1}{1+e^0} = \frac{1}{2}$

$f(2) = \frac{1}{1+e^{-2}}$

$f(4) = \frac{1}{1+e^{-4}}$

Pearson's correlation:

$$PCC(x_1, x_2) = \frac{\sum_{i=1}^{m}(a_{i1}-0)(a_{i2}-0.5)}{\sqrt{\sum_{i=1}^{m}(a_{i1})^2}\sqrt{\sum_{i=1}^{m}(a_{i2}-0.5)^2}} =$$

$$= \frac{(-4)\left(\frac{1}{1+e^4}-0.5\right)+(-2)\left(\frac{1}{1+e^2}-0.5\right)+(0)\left(\frac{1}{2}-0.5\right)+(2)\left(\frac{1}{1+e^{-2}}-0.5\right)+(4)\left(\frac{1}{1+e^{-4}}-0.5\right)}{\sqrt{(-4)^2+(-2)^2+0+2^2+4^2} \times \sqrt{\left(\frac{1}{1+e^4}-0.5\right)^2+\left(\frac{1}{1+e^2}-0.5\right)^2+(0)+\left(\frac{1}{1+e^{-2}}-0.5\right)^2+\left(\frac{1}{1+e^{-4}}-\frac{1}{2}\right)^2}}$$

$$= \frac{5.3773}{\sqrt{40} \times 0.8687} \approx 0.9791$$

Spearman's Correlation

x_1	rank	x_2	rank
-4	1	$\frac{1}{1+e^{4}}$	1
-2	2	$\frac{1}{1+e^{2}}$	2
0	3	$1/2$	3
2	4	$\frac{1}{1+e^{-2}}$	4
4	5	$\frac{1}{1+e^{-4}}$	5

As in 1a), Spearman $(x_1, x_2) = 1$

→ Here the correations differ, due to the nature of the sigmoid curve, that is not linear, so a correlation between data (Pearson's) is underepresented, contrary to a correlation between ranks (Spearman's). In addition, Spearman's correlation is higher because the distribution of data can be better aproximated by a non-linear monotonic function rather than by a linear function.

2 Decision Trees

2.1 a)

② a) $E_{start}\left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\right) = -\sum_{i=1}^{m} P_i \log_2(P_i) =$

$$= -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 1.5219 \text{ bit}$$

—//—

$\boxed{\text{Weekend}}$

Weekend = Yes	Weekend = No
#{What to do? = Go for a walk} = 1	#{What to do? = TV} = 1
#{What to do? = Reading} = 1	#{What to do? = Reading} = 1
	#{What to do? = Go for a walk} = 1

$$E_{weekend} = \frac{2}{5} E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{5} E\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) =$$

$$= \frac{2}{5} + 0.951 = 1.351 \text{ bit}$$

$IG(\text{Weekend}) = E_{start} - E_{weekend} = 0.17095$ bit

—//—

$\boxed{\text{Weather}}$

Weather = Sunny	Weather = Rain	Weather = Cloudy
#{Go for a walk} = 1	#{TV} = 1	#{Go for a walk} = 1
	#{Reading} = 1	#{Reading} = 1

$$E_{weather} = \frac{1}{5} E(1) + \frac{2}{5} E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{2}{5} E\left(\frac{1}{2}, \frac{1}{2}\right) = 0.8 \text{ bit}$$

$IG(\text{Weather}) = 1.5219 - 0.8 = 0.7219$ bit

$\boxed{\text{Tired}}$

Tired = Yes	Tired = No
#{TV} = 1	#{Go for a walk} = 2
#{Reading} = 1	#{Reading} = 1

$$E_{tired} = \frac{2}{5} \times E\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{5} \times E\left(\frac{2}{3}, \frac{1}{3}\right) =$$

$$= \frac{2}{5} \times 1 + \frac{3}{5} \times 0.918 = 0.9508 \text{ bit}$$

$$IG(tired) = 1.5219 - 0.9508 = 0.5711 \text{ bit}$$

As the information gain is greater ~~[illegible]~~ with the attribute "weather", the root of the tree must be "weather".

2.2 b)

Weather = Sunny

Weekend	Tired	What to do?
yes	no	Go for a walk

done! (there is no more uncertainty)

Weather = Rain

Weekend	Tired	What to do?
No	Yes	TV
No	No	Reading

The partition "weather = rain" still has uncertainty.

Weather = Cloudy

Weekend	Tired	What to do?
No	No	Go for a walk
Yes	Yes	Reading

The partition "weather = cloudy" still has uncertainty.

With "Weather = Rain", the attribute tired is the only one that leads to a split of the dataset, so we will choose it:

Weather = rain	
Tired = Yes	Tired = No
What to do? TV	What to do? Reading

Done! (there's no more uncertainty)

Both attributes "Tired" and "Weekend" lead to uncertainty with "weather = cloudy". So, according to the problem statement, the attribute "weekend" will be chosen to split the dataset.

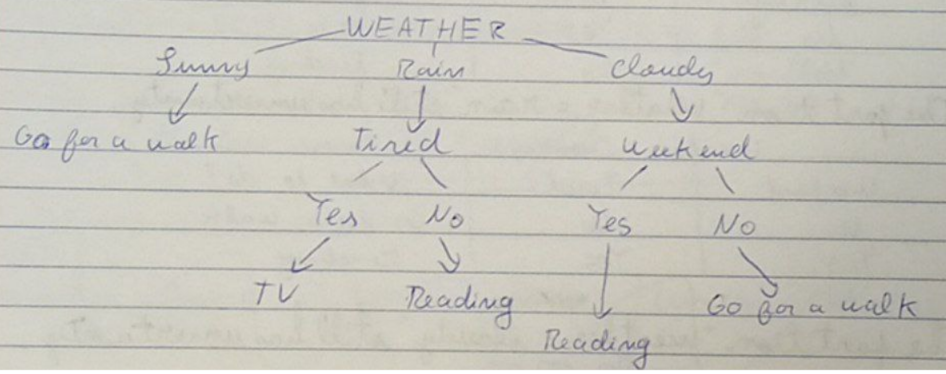
Weather = Cloudy

Weekend = Yes		Weekend = No	
Tired	What to do?	Tired	What to do?
Yes	Reading	No	Go for a walk

Done! Done!

Done!

So, the decision tree will be (as there's no more uncertainty):



2.3 c)

c)

predicted

true	Go for a walk	TV	Reading
Go for a walk	1	0	0
TV	0	1	0
Reading	2	1	0

3 Software Experiments

The group number is 9 (gn=9).

3.1 a)

Results for `value = 0.1`
train size: 17
test size: 161
accuracy on testing set: 0.68
depth: 2
number of leaves: 3

Results for `value = 0.9`
train size: 160
test size: 18
accuracy on testing set: 0.83
depth: 4
number of leaves: 8

This happens because the training was done considering a greater percentage of data, so the specificity of the tree augmented, and so did, consequently, its depth and its accuracy, thus avoiding underfitting, and the bias-variance trade-off (strong assumptions with weak basis). In other words, the tree could capture more patterns during the training with a *train_size* of 0.9, generalizing better and responding more correctly to unseen data, but also not overfiting the training data.

3.2 b)

The tree is less accurate due to overfitting, which results in worse generalization. This issue arises from not stratifying the data (i.e., cross-validation). As a consequence, some classes are underrepresented, meaning the proportion of classes in the training set does not align with that in the testing set.