TÉCNICO LISBOA

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework I
Deadline 29/9/2024 20:00
*Submit on Fenix as pdf*

# I) Correlation (2 pts)

Compute the correlation (Pearson correlation) and **Spearman's rank** for two variables **x**1 and **x**2

(a) (1 pts)
**x**1 = (1, 3, 4, 6)
**x**2 = (-30, -10, 0, 20)
Solution, we assume it. Is a sample
(if you assume it is a population you will get the same results for correlation):

*Mean(x1)=7/2, Mean(x2)=-5*

Covariance for Sample *cov(x1,x2)=130/3*

Standard Deviation for sample: *s1=Sqrt(13/3), s2=10* Sqrt(13/3)*

*corr(x1,x2)=cov(x1,x2)/(s1*s2)=1*

*R1=(1,2,3,4)*
*R2=(1,2,3,4)*

*Mean(R1)=5/2, Mean(R2)=5/2*

Covariance for Sample *cov(R1,R2)=5/3*

Standard Deviation for sample: *s1=Sqrt(5/3), s2=Sqrt(5/3)*

*corr(R1,R2)=cov(R1,R2)/(s1*s2)=1*

Why are the same? *Points are on the line.*

(b) (1pts)
**x**1 = (1, 3, 4, 6)
**x**2 = (-3, -0.5, 29, 30)

*Mean(R1)=7/2, Mean(R2)=13.87*

Covariance for Sample *cov(R1,R2)=32.4167*

Standard Deviation for sample: *s1=Sqrt(13/3), s2=18.075*

*corr(x1,x2)=cov(x1,x2)/(s1*s2)=0.861516*

TÉCNICO LISBOA

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework I
Deadline 29/9/2024 20:00
*Submit on Fenix as pdf*

*R1=(1,2,3,4)*
*R2=(1,2,3,4)*

Rank is the same as before…

*corr(R1,R2)=cov(R1,R2)/(s1*s2)=1*

Why are they different? *Points are ordered but are not on the line.*

## II) Decision Trees (5 pts)

| *F1* | *F2* | *F3* | *F4* | *Output* |
|---|---|---|---|---|
| *c* | *a* | *b* | *x* | *n* |
| *a* | *a* | *c* | *a* | *t* |
| *a* | *b* | *b* | *a* | *t* |
| *c* | *b* | *c* | *x* | *m* |
| *a* | *b* | *b* | *c* | *f* |

( a ) (2 pts)
Determine the root of decision tree using the ID3 algorithm with the target "Output". Indicate the calculation.

Solution:

*p(n)=1/5 p(t)=2/5, p(m)=1/5, p(f)=1/5*

*Log2[x]= Log[x]/Log[2]*

*I(table)= -3*1/5*Log2[1/5] - 2/5*Log2[2/5]* = 1.92193 bits

$$E(P) = \sum_{i=1}^{n} \frac{|C_i|}{|C|} I(C_i) \qquad gain(P) = I(C) - E(P)$$

*F1 Ca=(t,t,f), Cc=Cx=(n,m),*
*I(Ca)= -2/3*Log2[2/3] – 1/3*Log2[1/3]*= 0.918296 *bit*
*I(Cc)= I(Cx) = -1/2*Log2[1/2] – 1/2*Log2[1/2]*= 1 *bit*
*E(F1)= 2/5*1+3/5**0.918296= *0.950978 bit*
*Gain(F1) =Gain(F4)=* 1.92193-*0.950978=0.970952 bit*

*F2 Ca=(n,t) Cb=(t,m,f)*
*I(Ca)= -1/2*Log2[1/2] – 1/2*Log2[1/2]*= 1 *bit*
*I(Cb)= -3*1/3* Log2[1/3]= 1.58496 bit*

*E(F2)= 2/5*1+3/5*1.58496=1.35098 bit*
*Gain(F2)=* 1.92193-*1.35098=0.57095 bit*

*F3 Cb=(n,t,f) Cc=(t,m)*
*I(Cb)= -3*1/3*Log2[1/3] =Log2[3]=* 1.58496 *bit*
*I(Cc)= -2*1/2Log2[1/2]=Log2[2]=1 bit*
*E(F3)= 3/5*1.58496+2/5*1=1.3509 bit*
*Gain(F3)= 1.92193-1.3509=0.57103 bit*

*F4 Ca=(t,t) Cx=(n,m) Cc=(f)*
*I(Ca)=I(Cc)=0*
*I(Cx)= -2*1/2Log2[1/2]=Log2[2]=1 bit*
*E(F4)= 2/5*1+2/5*0+1/5*0=**0.4** bit*
*Gain(F4)= 1.92193-0.4=**1.52193** bit*

*We chose **F4** as the root*

(b) (2 pts)
Determine the decision tree using the ID3 algorithm with the target "Output". Indicate the calculation and draw your decision tree.

Solution:

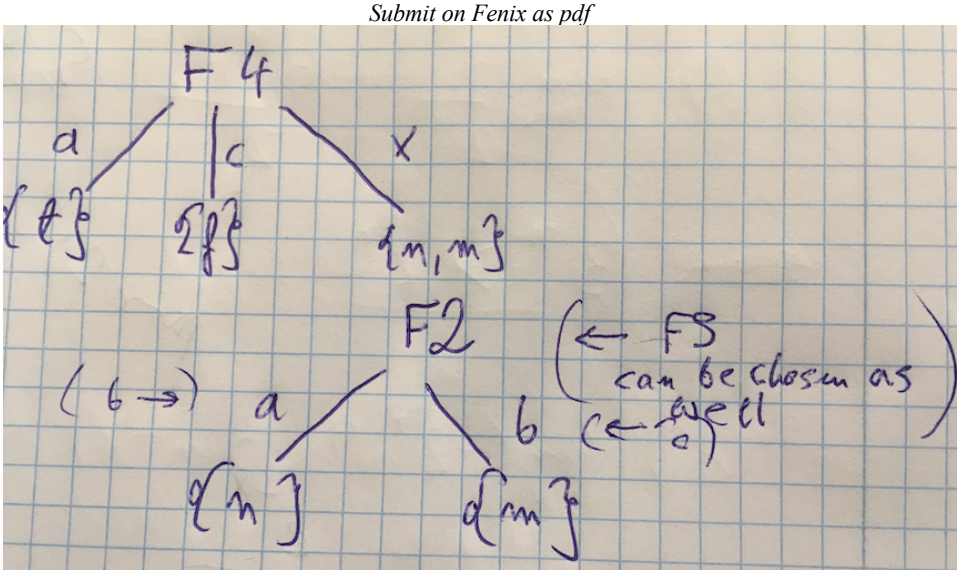
*The remaining table Cx:*

| *F1* | *F2* | *F3* | *Output* |
|---|---|---|---|
| *c* | *a* | *b* | *n* |
| *c* | *b* | *c* | *m* |

*F2 and F3 are equal, they give the same gain. We chose for the tree F2*

TÉCNICO LISBOA

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework I
Deadline 29/9/2024 20:00
*Submit on Fenix as pdf*




(c) (1 pts)
Draw the training confusion matrix for the learnt decision tree.

| | TRUE | | | | |
|---|---|---|---|---|---|
| | | n | t | m | f |
| | n | 1 | 0 | 0 | 0 |
| PREDICTED | t | 0 | 2 | 0 | 0 |
| | m | 0 | 0 | 1 | 0 |
| | f | 0 | 0 | 0 | 1 |