ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

2Η ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ

Ευάγγελος Μπενέτος, ΑΜ:1072628

Α. Υποθέτοντας πως έχετε κάνει διαχωρισμό με βάση την ιδιότητα Outlook, υπολογίστε την εντροπία E (Sail | Outlook) για την κλάση του προβλήματος.

Έχουμε όπως μας δίνεται από την άσκηση τον παρακάτω πίνακα.

	Sail	Outlook	Company	Sailboat
1	yes	Rainy	big	big
2	yes	Rainy	big	small
3	no	rainy	med	big
4	no	rainy	med	small
5	yes	sunny	big	big
6	yes	sunny	big	small
7	yes	sunny	med	big
8	yes	sunny	med	big
9	yes	sunny	med	small
10	yes	sunny	no	small
11	no	sunny	no	big
12	no	rainy	med	big
13	no	rainy	no	big
14	no	rainy	no	big
15	no	rainy	no	small
16	no	rainy	no	small
17	yes	sunny	big	big
18	no	sunny	big	small
19	no	sunny	med	big
20	no	sunny	med	big

Απάντηση Α ερωτήματος:

Για να υπολογίσουμε την εντροπία στο συγκεκριμένο πρόβλημα αυτό που μας ενδιαφέρει από τα στοιχεία που μας δίνονται στο πρόβλημα είναι η στήλη sail στην οποία πρέπει να δούμε πως έχουμε 9 yes και 11 no τα οποία θεωρητικά σύμφωνα με την άσκηση αντιστοιχούν στα σκάφη του προβλήματος μας.

Οπότε για να υπολογίσουμε την εντροπία θα κάνουμε τα εξής:

ENTROPY (Sail | Outlook)=P(Outlook= rainy)*entropy(2+,7-) + P(Outlook= sunny) *entropy(7+,4-)=9/20*(-2/9*$\log_2$9/2 – 7/9*$\log_2$7/9) + 11/20(-7/11*$\log_2$7/11 – 4/11*$\log_2$4/11)=0,864.

Β. Υπολογίστε το κέρδος πληροφορίας (*IG*) αν ο διαχωρισμός στην κορυφή γίνει με βάση την ιδιότητα Outlook.

Εφόσον ο διαχωρισμός στη κορυφή γίνεται με βάση την ιδιότητα outlook θα έχουμε τα εξής:

Ssail(9+,11-):

ENTROPY (Sail)=-11/20* $\log_2$11/20 – 9/20* $\log_2$9/20=0,992

Srainy(2+,7-):

ENTROPY (Srainy)=-2/9* $\log_2$2/9 – 7/9* $\log_2$7/9=0.764

Ssunny (7+,4-):

ENTROPY (Ssunny)=-7/11* $\log_2$7/11 – 4/11* $\log_2$4/11=0.95

Οπότε σύμφωνα με τα προηγούμενα θα υπολογίσουμε το κέρδος:

Gain(sail,outlook)= ENTROPY (Sail) - 9/20* ENTROPY S(rainy) – 11/20* ENTROPY (Ssunny)=0.992 - 0.343 – 0.52=0.128

Γ. Χρησιμοποιώντας το κέρδος πληροφορίας ως μετρική, ποια ιδιότητα θα χρησιμοποιηθεί για διαχωρισμό στη ρίζα του δέντρου απόφασης; Δείξτε τους υπολογισμούς.

Σαν πρώτη κίνηση αυτό που έχουμε να κάνουμε είναι να υπολογίσουμε την εντροπία για ολόκληρο το dataset μας δηλαδή τα δεδομένα μας και αυτό θα γίνει στην δική μας περίπτωση υπολογίζοντας την εντροπία του sail όπου αν κοιτάξουμε για το sail έχουμε 9 θετικά στοιχεία και 11 αρνητικά αρα:

ENTROPY (sail)=9/20* $\log_2$(9/20) – 11/20* $\log_2$ (11/20)=0.992

Τώρα αυτό που θα κάνουμε είναι για κάθε κατηγορία στον πίνακα (Outlook,Company,Sailboat) να πάμε να βρούμε την εντροπία των περιπτώσεων σε κάθε κατηγορία του πίνακα.

Α) Για το Outlook έχουμε περιπτώσεις rainy και sunny:

- Για την περίπτωση rainy έχουμε 2 yes και 7 no άρα έχουμε ότι,

ENTROPY (Srainy) =-2/9* $\log_2$(2/9) – 7/9* $\log_2$ (7/9)=0.764

- Για την περίπτωση sunny έχουμε 7 yes και 4 no άρα έχουμε ότι,

ENTROPY (Ssunny) =- 7/11* $\log_2$(7/11) – 4/11* $\log_2$ (4/11)=0.945

Τώρα αυτό που πρέπει να γίνει είναι να χρησιμοποιήσουμε τον τύπο για το κέρδος πληροφορίας για την κατηγορία Outlook οπότε θα έχουμε:

GAIN(S,Outlook) = ENTROPY (Sail) – $\Sigma_{(v=yes,no)}$(|Sv|/|S|)* ENTROPY (Sv) και σύμφωνα με τους προηγούμενους υπολογισμούς θα βρούμε την τιμή του GAIN=0.992 -9/20* ENTROPY (Srainy) – 11/20* ENTROPY (Ssunny) =0,1287 (***ΣΗΜΕΙΩΣΗ ΔΕΝ ΔΕΙΧΝΩ ΥΠΟΛΟΓΙΣΜΟΥΣ ΔΙΟΤΙ ΤΟΥΣ ΕΧΩ ΚΑΝΕΙ ΜΕ ΑΡΙΘΜΟΜΗΧΑΝΗ)

Θα ακολουθήσω την ίδια διαδικασία και για τα υπόλοιπα ώστε να δω ποια ιδιότητα θα χρησιμοποιηθεί για διαχωρισμό στη ρίζα του δέντρου απόφασης.

Β) Για το Company έχουμε περιπτώσεις big, med καιno:

- Για την περίπτωση med έχουμε 5 yes και 1 no άρα έχουμε ότι,

ENTROPY (Smed) =-5/6* $\log_2$(5/6) – 1/6* $\log_2$ (1/6)=0.65

- Για την περίπτωση big έχουμε 7 yes και 4 no άρα έχουμε ότι,

ENTROPY (Sbig) =- 7/11* $\log_2$(7/11) – 4/11* $\log_2$ (4/11)=0.945

- Για την περίπτωση no έχουμε 1 yes και 5 no άρα έχουμε ότι,

ENTROPY (Sno) =-1/6* $\log_2$(1/6) – 5/6* $\log_2$ (5/6)=0.4308

Αρα: GAIN(S,Company) = ENTROPY (Sail) – $\Sigma_{(v=yes,no)}$(|Sv|/|S|)* ENTROPY (Sv) και σύμφωνα με τους προηγούμενους υπολογισμούς θα βρούμε την τιμή του GAIN=0.992 -8/20* ENTROPY (Smed) –6/20* ENTROPY (Sbig) -6/20* ENTROPY (Sno) =0,286

Γ) Για το Sailbot έχουμε περιπτώσεις big και small:

- Για την περίπτωση big έχουμε 5 yes και 7 no άρα έχουμε ότι,

ENTROPY (Sbig) =-5/12* $\log_2$(5/12) – 7/12* $\log_2$ (7/12)=0.979

- Για την περίπτωση small έχουμε 4 yes και 4 no άρα έχουμε ότι,

ENTROPY (Ssmall) =-4/8* $\log_2$(4/8) – 4/8* $\log_2$ (4/8)=1

Αρα: GAIN(S,Sailbot) = ENTROPY (Sail) – $\Sigma_{(v=yes,no)}$(|Sv|/|S|)* ENTROPY (Sv) και σύμφωνα με τους προηγούμενους υπολογισμούς θα βρούμε την τιμή του GAIN=0.992 -12/20* ENTROPY (Sbig) –8/20* ENTROPY (Ssmall)=0.048

Έχοντας βρει τα κέρδη και για τις 3 περιπτώσεις αυτή που θα χρησιμοποιηθεί για διαχωρισμό στη ρίζα του δέντρου απόφασης είναι αυτή με το μεγαλύτερο κέρδος πληροφορίας άρα η Company.

Δ. Το παραπάνω σύνολο δεδομένων θα αλλάξει, αν διαγράψουμε ορισμένα αντίγραφα από κάθε μοναδική εγγραφή. Για το αλλαγμένο σύνολο δεδομένων, που φαίνεται παρακάτω, ποια ιδιότητα θα επιλεγόταν στην κορυφή του δέντρου απόφασης;

	Sail	Outlook	Company	Sailboat
1	yes	rainy	big	big
2	yes	rainy	big	small
3	no	rainy	med	big
4	no	rainy	med	small
5	yes	sunny	big	big
6	yes	sunny	big	small
7	yes	sunny	med	big
8	yes	sunny	med	big
9	yes	sunny	med	small
10	yes	sunny	no	small
11	no	sunny	no	big
~~12~~	~~no~~	~~rainy~~	~~med~~	~~big~~
13	no	rainy	no	big
~~14~~	~~no~~	~~rainy~~	~~no~~	~~big~~
15	no	rainy	no	small
~~16~~	~~no~~	~~rainy~~	~~no~~	~~small~~
~~17~~	~~yes~~	~~sunny~~	~~big~~	~~big~~
18	no	sunny	big	small
19	no	sunny	med	big
~~20~~	~~no~~	~~sunny~~	~~med~~	~~big~~

Εφόσον έχουμε να βρούμε το ίδιο ακριβώς με το προηγούμενο ερώτημα γ) θα ακολουθήσουμε την ίδια διαδικασία με πριν απλά προφανώς θα έχουμε άλλα αποτελέσματα.

Οπότε σαν πρώτη κίνηση αυτό που έχουμε να κάνουμε είναι να υπολογίσουμε την εντροπία για ολόκληρο το dataset δηλαδή με λίγα λόγια για το sail,για το οποίο έχουμε 8 yes και 7 no:

ENTROPY (sail)=8/15* $\log_2$(8/15) – 7/15* $\log_2$ (7/15)=0.996

Τώρα αυτό όπως και πριν αυτό που θα κάνουμε είναι για κάθε κατηγορία στον πίνακα (Outlook,Company,Sailboat) να πάμε να βρούμε την εντροπία των περιπτώσεων σε κάθε κατηγορία του πίνακα.

Α) Για το Outlook έχουμε περιπτώσεις rainy και sunny:

- Για την περίπτωση rainy έχουμε 2 yes και 4 no άρα έχουμε ότι,

ENTROPY (Srainy) =-2/6* $\log_2$(2/6) – 4/6* $\log_2$ (4/6)=0.92

- Για την περίπτωση sunny έχουμε 6 yes και 3 no άρα έχουμε ότι,

ENTROPY (Ssunny) =- 6/9* $\log_2$(6/9) – 3/9* $\log_2$ (3/9)=0.91

Αρα:GAIN(S,Outlook) = ENTROPY (Sail) – $\Sigma_{(v=yes,no)}$(|Sv|/|S|)* ENTROPY (Sv) και σύμφωνα με τους προηγούμενους υπολογισμούς θα βρούμε την τιμή του GAIN=0.992 -6/15* ENTROPY (Srainy) – 9/15* ENTROPY (Ssunny) =0,079

Β) Για το Company έχουμε περιπτώσεις big, med καιno:

- Για την περίπτωση med έχουμε 3 yes και 3 no άρα έχουμε ότι,

ENTROPY (Smed) =- 3/6* $\log_2$(3/6) – 3/6* $\log_2$ (3/6)=1

- Για την περίπτωση big έχουμε 4 yes και 1 no άρα έχουμε ότι,

ENTROPY (Sbig) =- 4/5* $\log_2$(4/5) – 1/5* $\log_2$ (1/5)=0.72

- Για την περίπτωση no έχουμε 1 yes και 3 no άρα έχουμε ότι,

ENTROPY (Sno) =-1/4* $\log_2$(1/4) – 3/4* $\log_2$ (3/4)=0.8

Αρα: GAIN(S,Company) = ENTROPY (Sail) – $\Sigma_{(v=yes,no)}$(|Sv|/|S|)* ENTROPY (Sv) και σύμφωνα με τους προηγούμενους υπολογισμούς θα βρούμε την τιμή του GAIN=0.992 -6/15* ENTROPY (Smed) –5/15* ENTROPY (Sbig) -4/15* ENTROPY (Sno) =0,14

Γ) Για το Sailbot έχουμε περιπτώσεις big και small:

- Για την περίπτωση big έχουμε 4 yes και 4 no άρα έχουμε ότι,

ENTROPY (Sbig) =-4/8* $\log_2$(4/8) – 4/8* $\log_2$ (4/8)=1

- Για την περίπτωση small έχουμε 4 yes και 3 no άρα έχουμε ότι,

ENTROPY (Ssmall) =-4/7* $\log_2$(4/7) – 3/7* $\log_2$ (3/7)=0.99

Αρα: GAIN(S,Sailbot) = ENTROPY (Sail) – $\Sigma_{(v=yes,no)}$(|Sv|/|S|)* ENTROPY (Sv) και σύμφωνα με τους προηγούμενους υπολογισμούς θα βρούμε την τιμή του GAIN=0.992 -8/15* ENTROPY (Sbig) –7/15* ENTROPY (Ssmall)=0.0035

Όπως συμπεραίνουμε με τα αποτελέσματα πάλι η κατηγορία Company με το μεγαλύτερο κέρδος είναι αυτή που θα επιλεγόταν στην κορυφή του δέντρου απόφασης.