Sai Mahesh Mudavat

1152 7206

Data Mining

Section - 002

Q1) Data mining

→ Data mining is process of automatically discovering useful information in large data repositories. It provides tools to discover Knowledge from data.

In nutshell datamining turns a large collection of data into Knowledge.

→ ~~So it is more than Simple transformation technology~~

(b) Data mining is definetly more than a simple transformation of technology which evolved into Machine learning, databases, Statstics because huge data when started to accumulate to better manage those it first evolved into data management Systems, to retrieve the require data we started to use SQL. As decades passed data started to accumulate into accumulate. to advance database of so to mining Complex data we later used statistics and Machine learning to get right answer of data to solve a problem.

(2) Ambulance Medical care data

Features: gender, age, duration of visit, diagnosis, demography, symptoms, medication.

Task: Classification

Row: Patient ID (or) Patient name

Column: Symptoms, medication, visit, age

Target: To find whether patient has disease (or) not

Clustering

Task Row: Patient demography (or) patient visits age

(Columns) Symptoms, age, demography etc

Target Clustering can be done by having age group of people in with no symptoms and other set of age, with symptoms and predict whether this patient has chance of getting disease

Patients also be clustered based on demographic Conditions and predict whether there is there a possibility to predict this set of people of certain demography have chance of getting disease or not

Associate Rule mining

Row: Patient visit (or) patient id

Column: Symptoms, age, diagonsis of medical Condition

Target: We identify if Symptoms and medical Conditions to exist

We can also see demograph age and medical Conditions Correlation between them

Anamoly detection

Row: Patient visit

Column: Symptoms, demography, age, medication

Target: Young age people with Sudden ~~ill~~ illness Certain demography showing Certain common disease Symptoms (eg: diabetes)

(4A) Average number of hours user spent on Internet in week

Continuous, quantitative, ratio

Continuous because it is specific has range of values and real numbers

~~Quantitative~~

Ratio: As it average hours if it can also mean of ~~all~~ hours of usage across the week

(b) GPA of Student

→ Continuous, Qualitative, Ordinal

GPA: Values are 4.1, 4.2, 7.0, 8.3,

So they are Continuous

Ordinal because: Values are in certain order

(c) Credit Card number

Discrete, Qualitative, Nominal

Discrete due its distinctness

~~Qualitative~~ Nominal: There are distinguish and formed with certain permutation &c.

(d) Salary above median Salary of all employees

Continuous, Quantitative, Interval

Continuous: Salary figures are always in growing order

Interval: Salary are within this interval range

Hence it is also Quantitative

5A) Null means not applicable, if we replace with zero will be unable to distinguish non sales, Sales people, as there is possibility of Sales person with 0.

This is a disadvantage of replacing null with 0 in ~~putting~~ Sales Commission attribute

→ Mary will be more Similar to Bob and Lisa

6A)

$$p = (22, 1, 42, 10)$$

$$q = (20, 0, 36, 8)$$

Distance $d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$

$$d = \frac{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}{\sqrt{(2)^2 + (1)^2 + (6)^2 + (2)^2}}$$

$$\sqrt{4 + 1 + 36 + 4}$$

$$= 6.708$$

7A

$p+ = 5 \quad 0.5$

$p- = 5$

Entropy $= -0.5\log_2(0.5) - (0.5)\log(0.5)$

$= 1$

Case

Tobacco Smoking Yes - Lung Cancer Yes

4

Tobacco Smoking No - Lung Cancer No

1

$E(4,1)$

$= -\left(4/5 \log_2 4/5\right) - \left(1/5 \log_2 1/5\right)$

$= 0.72$

$E(1,4) \rightarrow$ (Tobacco Smoking Yes - Lung Cancer NO
Tobacco Smoking NO - Lung Cancer NO)

$= 0.72$

Information gain (Entropy (parent - Σ Entropy (child)

$1 - 0.72$

$= 0.278$

Radon Exposure

Radon Yes - Lung Cancer Yes
Lung Cancer - Yes

Radon No -

(2, 3)

$= -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})$

$= 0.52 - (-0.44)$

$= 0.96$

Radon Yes — L.C - Yes
Radon No — L.C - No

(0, 5)

$= 0$

Information gain

$1 - 0.76 = 0.23$

Chronic Cough → Yes — Lung Cancer Yes

Chronic Cough No — Lung Cancer No

$E(Yes) = (4,1)$

$-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$

≈ 0.72

Yes, Yes

Cough No, No

&

L.C

$(3,2)$

$-\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.96$

Information gain:

$1 - 0.96 = 0.04$

Weight loss - yes L.C : yes
Weight loss NO L.C : Yes

(3, 2)

$-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.97$

Weight loss yes L.C - NO
Weight loss: NO L.C; NO

(2, 3)

$-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.97$

Information gain

$1 - 0.971 = 0.029$

Tobacco smoking has highest Information gain

Tobacco smoking Yes

Lung Cancer Yes
Lung Cancer Yes

Radon Exposure Yes

Radon Exposure No

(1, 3)

$\left(-\frac{1}{5}\log_2\frac{1}{5}\right) - \left(\frac{3}{5}\log_2\frac{3}{5}\right)$

$= 0.8113$

$(-0.46) - (-0.44)$

R.Expo (Yes, No / No, No) → 0

Entropy total: 0.649

Info gain: 0.072

Chronic Cough

(Yes, Yes)
No, Yes
(4, 0)

(Yes, Yes / No, No)
(0, 1)

0

→ Info gain: 0.72

T.S → Yes

Weight loss

Yes Yes

No Yes

(1, 0) = 0

T.S: Yes

(Yes, No
No, No)

(2, 2) = 0.918

Information gain = 0.170

We Choose Chronic Cough for Tobacco

Smoking = Yes

Tobacco Smoking :
= NO

Radon - Yes, L.C : Yes
NO, L.C : Yes

(1, 0)

= 0

Radon YNO L.C : Yes
Radon NO : L.C = NO

(0, 4)

= 0

Information gain : 0.72

Chronic Cough, Lung Cancer, T.S = No

(Yes, Yes)
No, Yes

(0, 1) = 0

(3, 1) = 0.40

Information gain : 0.32

Weight loss

T.S : NO

Yes L.C : Yes

Weight loss NO, Yes

(1, 0) = 0

Weight loss

(2, 2)

= 0.91

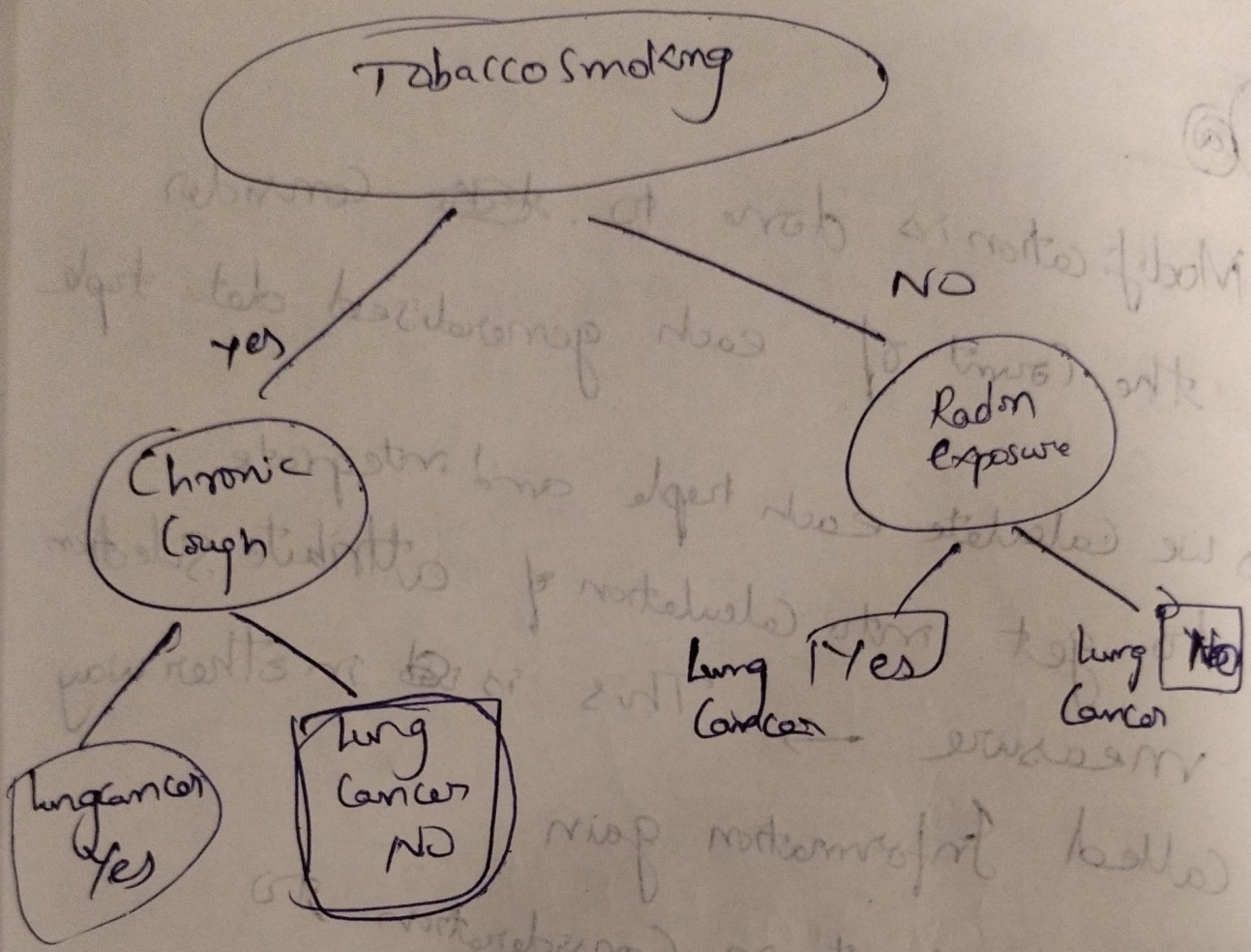
Entropy = 0.55

Info gain: 0.170

Here best split for Tobacco smoking = NO is Radon exposure

If Radon exposure = yes, L.C = yes

" " = no L.C = no


Tobacco smoking
Yes
NO
Chronic Cough
Radon Exposure
Lungcancer Yes
Lung Cancer NO
Lung Cancer
Yes
Lung Cancer
No

8 @

Modification is done to consider the count of each generalised data tuple

→ We calculate each tuple and integrate to get into calculation of attribute selection measure → This is in other way called Information gain

→ We also count in consideration to determine common class among the tuples.