على بالك - عين سر!

I, Define the business Problem: Define the Problem that the company wants to solve using data mining. ,1

II, Build the data mining data base: Collecting & integrating data from variety of sources to build a data mining database.

III, Explore the data: using variety of data analysis techniques & visualizations

IV, Prepare the data for modeling (PreProcess): Mostly feature engineering, cleaning data, removing NaN data types, etc.

V, Build a model: using a variety of machine-learning models & algorithms to identify Patterns in the data that can be used to make Prediction

VI, Evaluate model: Evaluating the model's Performance on a test set.

VII, Act on the results: once the evaluation step's results are satisfying, it can be deployed to Production.

Noise: unwanted signal that can interfere with the analysis of a dataset. ,2

outlier: A data point the deviates significantly from the rest of the data.

a, Benefits of finding outliers: { Improved data quality: By removing outlier, we can improve the quality of ~~error~~ data
New discoveries: By discovering them, we can gain better understanding of the data

b, Disadvantages of Noise & outlier: { Reduced accuracy of data mining models.
Misleading results
Increased computational time

---

a, Ratio: { weight, Height. The difference between 30^{kg} & 100^{kg} is the same with 100^{kg} & 130^{kg} ,3

Interval: { Temperature: Date & time: same as Ratio, but, there's no true Zero Point of the Celsius scale, or Date and time scale has no Zero Point.

Ordinal { Likert scale: disagree, neither agree nor disagree, agree
Military rank: Private, Corporal, Sergeant, ...

Nominal { Eye color:
Zip Code: } There is no natural order to the categories.

b, Std Number: → Tracking Student Progress over time
→ Matching students to records

Gender: → Identifying gender disparities
→ Studying gender differences

c, Median { Cannot be defined, Nominal
Can be defined, ordinal, Interval, Ratio

Mode { Cannot - - , —
Can - - : all 4 categories

Average { Cannot - - : Nominal, ordinal
Can - - : Interval, Ratio (mean)

a, one-hot encoding: method of converting categorical var into binary vectors. For each category in the var, a new column is created in the dataset. The value of each column is 1 if the data point belongs to that category, and 0 otherwise.

Label encoding: method of converting categorical var into numeric values. Each category is assigned a unique integer value.

b, one-hot,

ID	Color: Blue	Green	Red
1	1	0	0
2	0	0	1
3	1	0	0
4	0	1	0

Label:

ID	Color	Size	weight	Shape
1	1	10	0.3	1
2	2	8	0.3	1
3	1	8	0.7	2
4	3	12	0.3	3

↑ integer encoding.

c, Both have their own advantages & disadvantages. But generally one-hot encoding is more suitable method for converting categorical into numerical values.

mean vector: $(v + (-v) + 2v + (-2v))/4 = [2.5, 0, -7.5]$,5

$X = t \times$ Projection vector $= t \times [2.5, 0, -7.5]$ → The line on which the data is is depicted can be represented by this equation.

↙ scalar

?

$\beta_2 = \bar{y} - \beta_1 \bar{x}$, $\beta_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$,6

↳ $\beta_2 = \bar{y} - \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \bar{x}$

$\Rightarrow Cov(\beta_1, \beta_2) = E[(\beta_1 - E(\beta_1))(\beta_2 - E(\beta_2))]$

$= E\left\{ \left(\frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} - E\left(\frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) \right) \times \cdots \right.$

$\left. \left(\bar{y} - \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \bar{x} - E\left(\bar{y} - \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \bar{x} \right) \right) \right\}$

$= \frac{E\left[\sum (y_i - \bar{y})(x_i - \bar{x}) - E\left[\sum (y_i - \bar{y})(x_i - \bar{x}) \right] \right]}{\sum (x_i - \bar{x})^2} \times \cdots$

$\left(\bar{y} - \frac{E\left[\sum (y_i - \bar{y})(x_i - \bar{x}) \right]}{\sum (x_i - \bar{x})^2} \cdot \bar{x} - \bar{y} \right)$

$= \frac{Cov(x, y)}{\sum (x_i - \bar{x})^2} \cdot (-\bar{y}) = \frac{-Cov(x, y)\, \bar{y}}{\sum (x_i - \bar{x})^2}$ → if x, y independent $= 0$

---

a, ,7

col 1: 4 | 5 8 | 9 12 | 13

Q_2 = median = 8

$Q_1 = 5$, $Q_3 = 12$

IQR = 7, Lower bound $= 5 - \frac{3}{2}7 = -5.5$

Upper bound $= 12 + \frac{3}{2}7 = 22.5$

col 2: 0.5 | 1 1.2 | 3 3.2 | 4

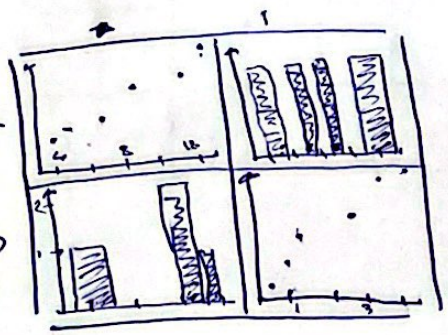
Q_2 = median = 2.1

$Q_1 = 1$, $Q_3 = 3.2$, IQR = 2.2

Lower $= 1 - \frac{3}{2} \cdot 2.2 = -2.3$

Upper $= 3.2 + \frac{3}{2} 2.2 = 6.5$

b,



SMC = $\frac{\text{No. of matching attributes}}{\text{Total No. of attributes}}$, $J(A,B) = \left|\frac{A \cap B}{A \cup B}\right|$

for 1st & 2nd row → SMC = $\frac{1}{3}$, $J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = 0$

b) Cosine similarity = $\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} = \frac{0.1+1.1+0.1}{\sqrt{1+1+0}\sqrt{1+1+1}} = \frac{2}{\sqrt{2}\sqrt{3}} = 0.816$

Bhattacharyya dist. $= -\log_e\left(\sum\sqrt{x_i y_i}\right) = -\log_e\left(\sqrt{0.1}+\sqrt{1.1}+\sqrt{0.1}\right)$

$= -\log_e(2) = -\ln(2) = 0.693$

c) Correlation coefficient $= \frac{n\sum x_i y_i - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

$= \frac{4[0+0+1+1]-3.2}{\sqrt{(4\cdot3-9)(4\cdot2-4)}} = \frac{8-6}{\sqrt{3\cdot4}} = \frac{2}{2\sqrt{3}} = \frac{\sqrt{3}}{3} = 0.577$

but because it's binary data type, Pearson correlation coefficient is not appropriate for this situation Phi correlation coefficient is used.

$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, where $\begin{cases} a = \text{No. of times } c1 \text{ \& } c2 \text{ are T} = 2 \\ b = \text{No. of times } c1 = T,\ c2 = f = 1 \\ c = \text{No. of times } c1 = f,\ c2 = T = 0 \\ d = \text{No. of times } c1 \text{ \& } c2 = f = 1 \end{cases}$

$\Rightarrow \phi = \frac{2\cdot1-1\cdot0}{\sqrt{3\cdot1\cdot2\cdot2}} = \frac{2}{\sqrt{12}} = 0.577$ Ans.

d) $H(\text{Class}) = -\left(\underbrace{\frac{2}{4}\log_2\frac{2}{4}}_{A} + \underbrace{\frac{2}{4}\log_2\frac{2}{4}}_{B}\right) = 1$, $H(C1) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right)$

$= 0.56$

$H(C2) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$, $H(C3) = 1$

e) $P(x,y) = \begin{cases} e^{-(x+y)} & x>0,\ y<0 \\ 0 & \end{cases} \Rightarrow \begin{cases} P(x) = \int_{-\infty}^{0} P(x,y)\,dy = \int_{-\infty}^{0} e^{-(x+y)}\,dy = e^{-x} \\ P(y) = \int_{-\infty}^{\infty} P(x,y)\,dx = \int_{0}^{\infty} e^{-(x+y)}\,dx = e^{-y} \end{cases}$

$I(x,y) = \int_{0}^{\infty}\int_{-\infty}^{0} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)dy\,dx = \int_{0}^{\infty}\int_{-\infty}^{0} e^{-(x+y)}\log\left(\frac{e^{-(x+y)}}{e^{-x}e^{-y}}\right)dy\,dx = 0$