German University in Cairo
Faculty of Media Engineering and Technology
Mervat Abu-Elkheir and Ayman Al-Serafi

CSEN911: Data Mining

Problem Set 1 – Data Preprocessing

Problem 1

Suppose a hospital tested the age and body fat data for 9 randomly selected adults with the following result:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

Given $\sigma_{age} = 11.29$, and $\sigma_{\%fat} = 8.91$

a) Calculate the mean of age and %fat.

b) Normalize the two variables based on the z-score normalization. ($new\ v_i = \frac{v_i - \bar{A}}{\sigma_A}$)

c) Mention whether these two attributes are positively or negatively correlated or not correlated at all, if the correlation coefficient is 0.71

Answer

a) **Age:** mean = 36.22

%fat: mean = 22.74

b) For each value in the table we compute its z-score as follows:

$$new\ v_1 = \frac{v_1 - \bar{A}}{\sigma_A} = \frac{23-36.22}{11.29} = -\ 1.17$$

$$new\ v_3 = \frac{v_3 - \bar{A}}{\sigma_A} = \frac{27-36.22}{11.29} = -\ 0.82$$

And so on for the other values. Computation is similar for the %fat. The z-score are shown for the age, and %fat attributes for verification.

Age	23	23	27	27	39	41	47	49	50
z-age	**-1. 17**	**-1. 17**	**-0. 82**	**-0. 82**	**0.25**	**0.43**	**0.96**	**1.13**	**1.22**
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	**-1.49**	**0.42**	**-1.68**	**-0.55**	**0.97**	**0.35**	**0.52**	**0.5**	**0.95**

c) Age and %fat are **positively** correlated

Problem 2

Jake did a survey of the numbers of brothers and sisters of the children in his class, and summarized the results in the following table:

# Siblings	Frequency
0	91
1	93
2	93
3	98
4	75
5	79
6	90
7	85
8	50
9	5

Draw the boxplot for these observations and identify outliers if present.

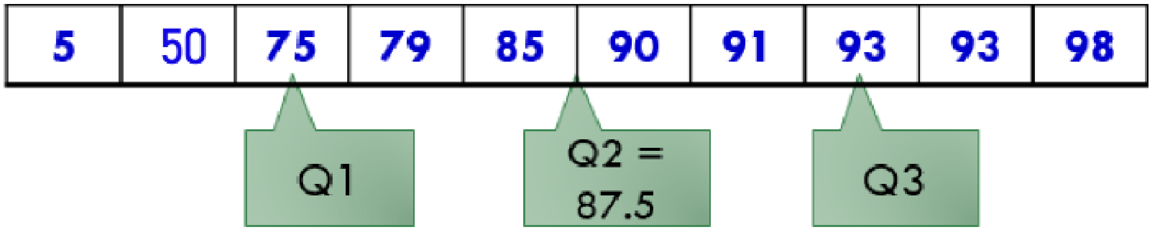
Answer

The steps for drawing the boxplot and identifying the outliers are:

1- Order the observations

5	**50**	**75**	**79**	**85**	**90**	**91**	**93**	**93**	**98**

2 - Divide the observations into quarters

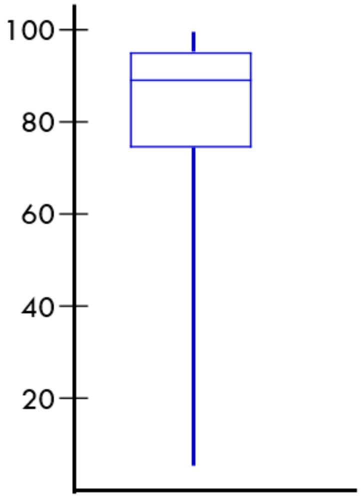


5	50	75	79	85	90	91	93	93	98

3 - Compute the IQR = Q3 – Q1 = 18

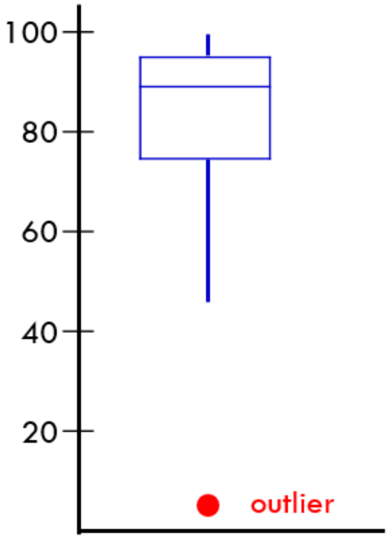
4 - Compute IQR×1.5 = 27

5 - Draw the boxplot



6 - Compute Q1 – (IQR×1.5) = 75 – 27 = 48 & Q3 + (IQR×1.5) = 93 + 27 = 120

7 - Update boxplot and mark outliers □ 50 becomes new min & 5 is outlier



Problem Set 2 – Regression

Problem 3

Most supermarkets use scanners at the checkout counters. The data collected this way can be used to evaluate the effect of price and store's promotional activities on the sales of any product. The promotions at a store change weekly, and are mainly of two types: flyers distributed outside the store and through newspapers (which may or may not include that particular product), and in-store displays at the end of an aisle that call the customers' attention to the product. Weekly data was collected on a particular product brand, including sales (in number of units), price (in dollars), flyer (1 if product appeared that week, 0 if it didn't) and display (1 if a special display of the product was used that week, 0 if it wasn't).

a) As a preliminary analysis, a simple linear regression model was done. The fitted regression equation was: $sales = 2259 - 1418\,price$**.**

1) The response variable is:

i. price.	□
ii. flyer.	□
iii. sales.	■
iv. display.	□

2) According to this model, how many units will be sold, on average, when the price of the product is \$1.10?

i. 3818.8.	□
ii. 699.2.	■
iii. 1066.9.	□
iv. 3902.9.	□

b) Next, the categorical variables flyer and display were added to the model. The fitted regression equation was: $sales = 3829.5 - 5056\,price + 804.12 flyer - 31.49 display$**.**

1) According to this model, what is the average effect of advertising the product on the weekly flyer, after adjusting for price and display?

i. There will be a significant increase in sales of about 804 units.	■
ii. There will be an insignificant increase in sales of about 9 units.	□
iii. There will be a significant decrease in sales of about 804 units.	□
iv. There will be an insignificant decrease in sales of about 5056 units.	□

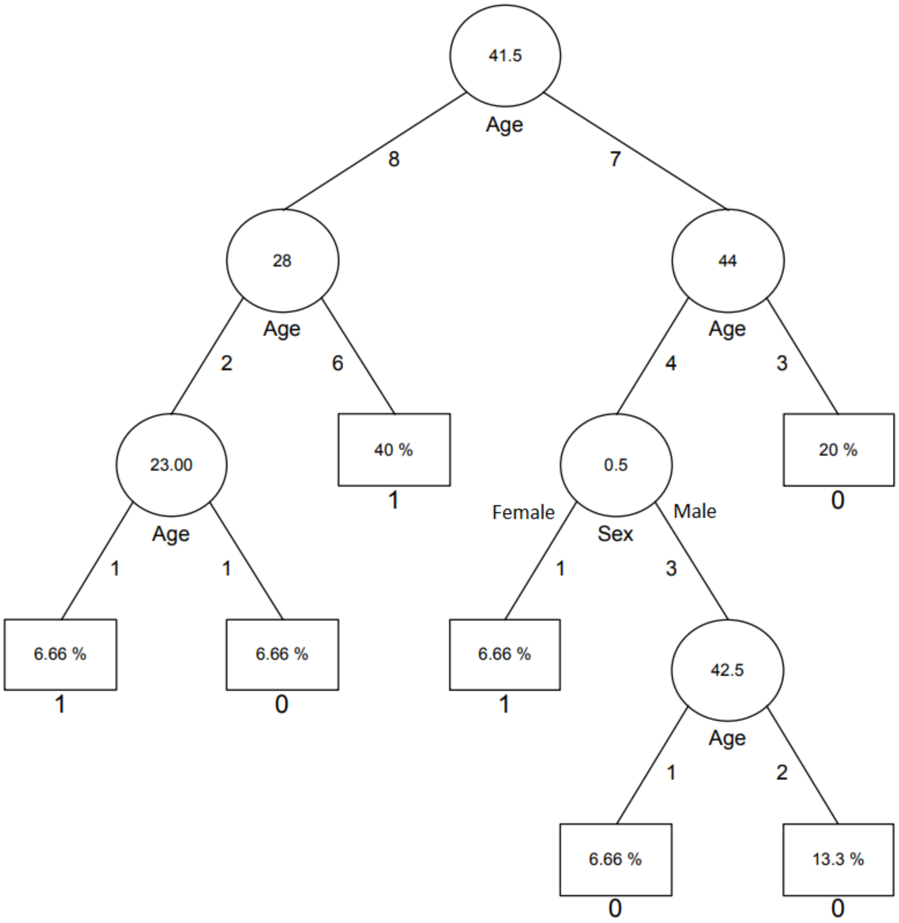
2) Flyers usually advertise products that are on sale that week. This implies that we should add to the model:

i. an interaction between display and flyer.	□
ii. an interaction between price and flyer.	■
iii. an interaction between price and display.	□
iv. an additional term for flyer.	□

Problem Set 3 – Classification

Problem 4

A credit card company has created a classification tree shown below for a promotion consisting of a special offer on credit card insurance. The class **1** corresponds to customers who responded, class **0** consists of those who did not. The company will use the tree to send promotion offers to dormant customers to persuade them to begin to use the card. Age, sex and income were used as input variables for these customers.



The company would like to have a few, simple, English language rules that embody the decisions represented by the tree. Write out succinctly the rules you would suggest.

Answer

Rule 1: If the customer is younger than 23 years old, then they would respond to the promotion.

Rule 2: If the customer is older than 28 years old and younger than 41.5 years old, then they would respond to the promotion.

Rule 3: If the customer is female who is older than 41.5 years old and younger than 44 years old, then they would respond to the promotion.

Rule 4: If the customer is older than 23 years old and younger than 28 years old, then they would not respond to the promotion.

Rule 5: If the customer is male who is older than 41.5 years old and younger than 44 years old, then they would not respond to the promotion.

Rule 6: If the customer is older than 44 years old, then they would not respond to the promotion.

In brief:

Customers who will respond to promotions are younger than 23, or between 28 and 41.5 years old, or females who are between 41.5 and 44 years old.

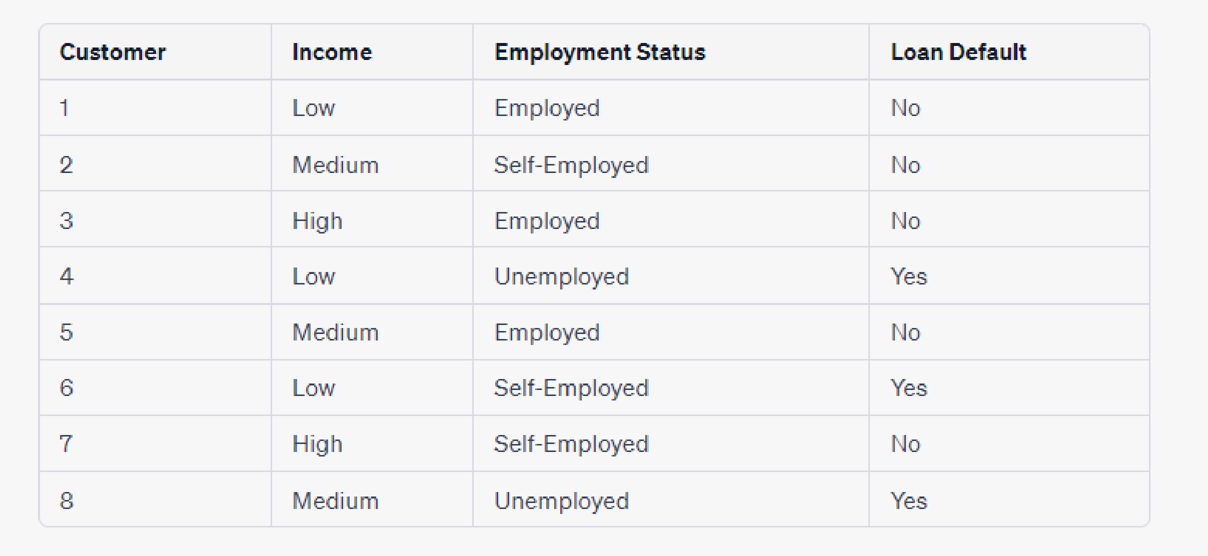
Customers who will not respond to promotions are between 23 and 28 years old, or males between 41.5 and 44 years old, or customers who are 44+ years old.

This can be summarized as follows:

< 23	$23 - 28$	$28 - 41.5$	$41.5 - 44$	$44 +$
Respond (1)	Won’t respond (0)	Respond (1)	Females: Respond (1) Males: Won’t respond (0)	Won’t respond (0)

Problem 5

Use the Naive Bayes algorithm to classify the following customers into one of the two categories: Default (Yes) or No Default (No).



Customer	Income	Employment Status	Loan Default
1	Low	Employed	No
2	Medium	Self-Employed	No
3	High	Employed	No
4	Low	Unemployed	Yes
5	Medium	Employed	No
6	Low	Self-Employed	Yes
7	High	Self-Employed	No
8	Medium	Unemployed	Yes

a) Customer 9: Income = Low, Employment Status = Employed

b) Customer 10: Income = Medium, Employment Status = Self-Employed

Solution:

$$P\ (C|X)\ = \frac{P\ (X|C) * P(C)}{P(X)}$$

- **Neglect the P(X) since it is common.**

- **P(Yes) =** $\frac{3}{8}$

- **P(No) =** $\frac{5}{8}$

a) **P(Income=Low | Yes)** = $\frac{2}{3}$

P(Income=Low | No) = $\frac{1}{5}$

P(Emp.Status= Employed | Yes) = $\frac{0}{3}$

- To prevent a product with 0 leading to 0 overall probability we have to use laplace smoothing (add 1 smoothing).

P(Emp.Status=Employed | No) = $\frac{3}{5}$

- **After applying Laplace smoothing (**Add **1** to the **numerator** and add the number of distinct values for each variable to the **denominator):**

P(Yes| C9) = $\frac{2+1}{3+3} * \frac{0+1}{3+3} * \boldsymbol{P(Yes)} = \frac{3}{6} * \frac{1}{6} * \frac{3}{8}$ **= 0.031**

P(No| C9) = $\frac{1+1}{5+3} * \frac{3+1}{5+3} * \boldsymbol{P(No)} = \frac{2}{8} * \frac{4}{8} * \frac{5}{8}$ **= 0.078**

Since **0.078 > 0.031**

Therefore, **Customer 9** will be classified to **No**

b) **P(Income=Medium | Yes) =** $\frac{1}{3}$

P(Income=Medium | No) = $\frac{2}{5}$

P(Emp.Status= Self_Employed | Yes) = $\frac{1}{3}$

P(Emp.Status= Self_Employed | No) = $\frac{2}{5}$

P(Yes| C10) = $\frac{1}{3} * \frac{1}{3} *$ ***P(Yes)*** $= \frac{1}{3} * \frac{1}{3} * \frac{3}{8}$ **= 0.042**

P(No| C10) = $\frac{2}{5} * \frac{2}{5} *$ ***P(No)*** $= \frac{2}{5} * \frac{2}{5} * \frac{5}{8}$ **= 0.1**

Since **0.1 > 0.042**

Therefore, **Customer 10** will be classified to **No**