Assignment
Due: 12th Nov 22@1pm in your github account

- Must answer all questions thoroughly in writing in your words.
- Typed assignment will not be accepted.
- No grades will be given if any malpractice found

Problem 1 Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time

Cond. to next slide...

Air-Traffic Data

Cond. from previous slide…

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Air-Traffic Data

- In this database, there are four attributes

  A = [Day, Season, Fog, Rain]

  with 20 tuples.
- The categories of classes are:

  C= [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, we are to find most likely classification for any other unseen instance, for example:

Week Day	Winter	High	None	???

- Classification technique eventually to map this tuple into an accurate class.

Problem 2: Statistical Learning

- Suppose that a group of 1,500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their preferred type of reading material was fiction or nonfiction. Thus, we have two attributes, gender and preferred reading. The observed frequency (or count) of each possible joint event is summarized in the contingency table shown below, where the numbers in parentheses are the expected frequencies

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

- Provide conclusion whether to reject or accept hypothesis where gender and preferred reading –

Vaibhav Raina 2018110040 DA Assignments.

Q1

Ans Attributes.

1) Day	Ontime	late	Very late	Cancelled
Weekday	9/14	1/2	3/3	0/1
Sat	2/14	1/2	0/3	1/1
Sun	1/14	0/2	0/3	0/1
Holiday	2/14	1/2	0/3	0/1

2) Season	Ontime	late	Very late	Cancelled.
Spring	4/14	0/2	0/3	1/1
Summer	6/14	0/2	0/3	0/1
Autumn	2/14	0/2	1/3	0/1
Winter	2/14	2/2	2/3	0/1

3) Fog	Ontime	late	Very late	Cancelled
None	5/14	0/2	0/9	0/1
High	4/14	1/2	1/8	1/1
Normal	5/14	1/2	2/9	0/1

4) Rain	Ontime	late	Very late	Cancelled.
None	6/14	1/2	1/3	0/1
Slight	6/14	1/2	0/3	0/1
Heavy	2/14	0/2	2/3	1/1

Prior probability

Ontime	late	Very late	Cancel
14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Given Instance

Weekday, winter, High, Heavy, ?

① Case ①
On time = 0.0013

② Case ②
late = 0.0

③ Very late = 0.0222

④ Cancelled = 0.000

Case ③ is strongest

∴ The correct classification for given Instance is "very late"

∴ Weekday, winter, High, Heavy, very late.

Q2 The expected frequency

$$e_{11} = \frac{\text{count(male)} \times \text{count(friction)}}{n} = \frac{300 \times 450}{1500}$$

$$= 90.$$

$$e_{12} = C(\text{female}) \times C(\text{friction} = \frac{1200 \times 450}{1500}$$

$$= 360$$

$$e_{13} = 210 \qquad e_{14} = 840$$

In any row Sum of expected freq must equal total / observed freq. for that row and Sum of expected freq. in any column must also. total observed freq for that column.

$$\therefore \chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\therefore \chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360}$$

$$+ \frac{(1000-840)^2}{840} = 507.93$$

$$DOF = (2-1)(2-1) = 1.$$

For DOF = 1, the χ^2 value needed to reject hypothesis at 0.001 significance level is 10.828.

∴ Our value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that two attributes are strongly correlated for given group of people.