COMP1942 (Spring 2024)
Data Mining Project: Income Prediction

Last Update: 23 Jan, 2024

Due Date for Phase 1: 21 Feb, 2024 9:00am (via Canvas (softcopy))
Due Date for Phase 2: 17 April, 2024 9:00am (in class (hardcopy))
Due Date for Phase 3: 29 April, 2024 9:00am (in class (hardcopy) and via Canvas (softcopy))

1. Introduction

You are given two real datasets which contain individual information in US. In the first real dataset, each individual is associated with 13 attributes and 1 additional Boolean attribute called "income" indicating whether the individual had an income > 50K (per year) or not. The second real dataset is the same as the first dataset but the second one contains only the first 13 attributes but no attribute "income". The objective of this project is to predict whether each individual in the second dataset has an income > 50K or not.

There are three phases in this project – Phase 1, Phase 2 and Phase 3. In Phase 1, you are required to generate an Excel file from two raw files together with attribute names. In Phase 2, you are required to write a design report for this project. In Phase 3, you are required to follow the design report in Phase 2, generate the predicted attribute files for the second real dataset and write a final report.

Project should be completed in form of groups. For each group, only one copy is required.

2. Milestones

1. Phase 1
   1. You are given two real datasets in TEXT format, "training.txt" and "test.txt".
   2. File "training.txt" contains 13 attributes and 1 additional Boolean attribute.
   3. File "test.txt" contains 13 attributes only.

iv. Open these two TEXT files with MS Excel

v. Save them in one MS Excel file where the content of "training.txt" is included in "Sheet 1" and the content of "test.txt" is included in "Sheet 2". Please re-name "Sheet 1" as "training" and re-name "Sheet 2" as "test" in MS Excel.

vi. Insert a row at the beginning of each of the two sheets of the Excel file where this row gives the attribute names specified in Section 3

2. Phase 2

   i. In Phase 2, you are required to write a design report for this project.

   ii. You should list 5 possible data mining models you want to try

   iii. Note that a possible data mining model can be "Decision Tree Classifier" with a set of parameters and another possible data mining model can be "Decision Tree Classifier" with another set of parameters. Obviously, one of the possible data mining models can be "Nearest Neighbor Classifier" with a set of parameters.

3. Phase 3

   i. In Phase 3, you are required to follow the design report in Phase 2, use the XLMiner software to predict attribute "income" for the second real dataset and write a final report.

   ii. The final report should include the following.

      a) All materials in your design report written in Phase 2 (i.e., the 5 possible data mining models)

      b) Description of the XLMiner results for each of 5 possible data mining models

      c) Two examples illustrating what attributes determine an individual to have an income > 50K or not for each of 5 possible data mining models.

      d) Conclusions drawn from each of 5 possible data mining models and an overall conclusion

   iii. In addition to the final report, you are required to generate 5 predicted attribute files for the second real dataset in TEXT file format. Note that each predicted attribute file corresponds to the output of a possible data mining model you proposed in Phase 2. The file format is described in Section 4.

3. Data Specifications

There are 13 attributes in the first dataset and 1 additional Boolean attribute called “income”.

No.	Attribute Name	Attribute Content
1	age	continuous.
2	workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3	education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
4	education-num	continuous.
5	marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
6	occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
7	relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
8	race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
9	sex	Female, Male.
10	capital-gain	continuous.
11	capital-loss	continuous.
12	hours-per-week	continuous.
13	native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
14	income	>50K, <=50K.

4. File format of Predicted Attribute File

In Phase 3, you are required to submit 5 predicted attribute files for the second dataset. The files should be named as “predicted1.txt”, “predicted2.txt”, “predicted3.txt”, “predicted4.txt” and “predicted5.txt” where “predicted1.txt” corresponds to the output of the first data mining model proposed in Phase 2 and the other files have a similar meaning. The file format of each file is shown as follows.

<1st row: 1 or 0 where 1 corresponds to that the first individual in the second dataset has an income > 50K and 0 corresponds to that s/he does not>
<2nd row: 1 or 0 where 1 corresponds to that the second individual in the second dataset has an income > 50K and 0 corresponds to that s/he does not >
…

Here is a sample file.

```
1
0
1
0
0
0
0
0
…
```

We have an answer file for the predicted attribute file. Among 5 files given by you, we will select the one with the highest accuracy as the final file for marking.

5. Grading Policy

Phase 1

- **Excel file (10%)**

Phase 2

- **Design Report (30%)**

Phase 3

- **Final Report (40%)**
- **Predicted Attribute Files for the Second Real Dataset in TEXT file format (20%)**

Mark Deduction

- **Phase 1 and Phase 2**

  No late submission is allowed for Phase 1 and 2. All submissions after the deadline of each of these phases will *not* be accepted.

- **Phase 3**

Number of Days Late	**Deduction (out of 100 marks)**
1	10
2	30
3	70
4 or above	100

6. Turn-in

Note:

a. Only one member in your group is required to submit the project materials.

b. Your group could use one coupon to obtain full scores of the 5 predicted attribute files in Phase 3. Coupon could not be used in other parts of the project (e.g., the final report in Phase 3).

1. **Phase 1**
   - **i. A Soft Copy of One Excel file (Any Excel Version) (via Canvas)**
     (Please name your file as "<group no>.xlsx" or "<group no>.xls" (e.g., "14.xlsx"))
2. **Phase 2**
   - **i. A Hard Copy of Design Report (in class)**
3. **Phase 3**
   - **i. A Hard Copy of Final Report (in class)**
     (Note: If you plan to use your coupon for the 5 predicted attribute files, your coupon could be stapled to this final report. We will mark the final report and the 5 predicted attribute files together).
   - **ii. A *single* zipped file (in ".zip" file format) containing the following (via Canvas).**
     (Please name your file as "<group no>.zip" (e.g., "14.zip"))
     1. **A Soft Copy of 5 Predicted Attribute Files for the Second Real Dataset (in TEXT file format)**
        (Please name your files as "predicted1.txt", "predicted2.txt", "predicted3.txt", "predicted4.txt" and "predicted5.txt" as described before.)
        (Note: Even if you use the coupon (included in the final report for submission), you could still keep submitting your 5 files to know the original scores of these 5 files (though you could obtain full scores of these 5 files).)