FIT5197 2018 S1 Assignment 2

Kelvin Li and Wray Buntine

Due date: Sunday, Week 11, May 20th 2018

Details

Your solution should be submitted as a single R markdown file and a PDF report via the submission page on "My Assessments" by the deadline at the page. The R markdown file needs to run so we can convert it to PDF and view your answers, as well as read your code. For this, assume the assignment input data exists in the same directory.

Submit the following 3 items:

- the **markdown file**, which clearly marks in comments which task and sub-task (A.1-A.6, B.1-B.4, etc.) each but of code is used for;
- the **PDF report**, clearly marking each section with the corresponding task and sub-task, and any major decisions or justifications need to be clearly explained;
- your **modified version** of "auto_mpg_train.csv"

Note that "adult_income_train.csv", if you change it, should be done insitu in the R code, so no modified version should be submitted.

Add in metadata at the top of the R markdown file to give your name and student ID and the submission date, as follows:

```
title: "FIT5197 2018 S1 Assignment 2"
author: "Your Full Name, 202343254"
date: "18 May 2018"
```

This assignment is worth 15% of your final mark for the unit. Each question is worth 20 marks so there is 60 marks in total. It is subject to the standard guidelines regarding academic integrity, late penalties, and special consideration. Some assignments may be selected and the authors interviewed in person to further assess their knowledge of their written submission.

Marking criteria:

- correctness of solution (main)
- readability of explanations (important for partial marks)
- justification of code if interviewed
- clarity of output
- quality of report

Task A: Linear Regression

Build a linear regression model using the specific “auto_mpg_train.csv” provided with the assignment to predict mpg (mile per gallon). The second file “auto_mpg_test.csv”will be used for evaluation.

Data Set Information: ”The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.” (Quinlan, 1993)

1. mpg: continuous, target variable
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

More details of the dataset is available at
`https://archive.ics.uci.edu/ml/datasets/Auto+MPG` .

Tasks

A.1: There are some missing values listed as “?”. Describe your strategy for treating missing values and update (edit by hand) the file accordingly.

A.2: Pair plot mpg vs. the other variables to visualize the relationships and discuss what you see.

A.3: Based on your pair plots, propose an initial set of variables to use for a multiple linear regression model to predict mpg.

A.4: With variables of your choice build the model using the `lm()` routine in R, and then print the summary of the model to get the R diagnostics. Briefly explain the statistics in the summary, e.g. R^2 value, t-value, standard error, p-value (ignoring the F-statistics line). What does this imply about the predictors for your model?

A.5: Test the fitted model using the “auto_mpg_test.csv”, and calculate the MSE on the test set, reporting it. Note the test set has no missing values.

A.6: Can you improve your model with different predictors? Try out some different ratios or products of the better predictor variables. How will you evaluate the different alternative predictors on your existing model (not using the test set)? Evaluate them and suggest which single predictor you would like to add (or none, if it looks like none would improve it). If you suggest adding a single predictor, then add it and repeat step A.5 to evalute it on the test set.

Task B: Logistic Regression

Build a logistic regression model using the specific “adult_income_train.csv” provided with the Assignment to predict the income variable. The second file “adult_income_test.csv”will be used for evaluation.

Data Set Information: Details of the dataset is available at `http://www.cs.toronto.edu/ delve/data/adult/` . Attribute information:

1. age: continuous
2. workclass: multi-valued discrete

3. fnlwgt: continuous
4. education: multi-valued discrete
5. educational-num: continuous
6. marital-status: multi-valued discrete
7. occupation: multi-valued discrete
8. relationship: multi-valued discrete
9. race: multi-valued discrete
10. gender: binary discrete
11. capital-gain: continuous
12. capital-loss: continuous
13. hours-per-week: continuous
14. native-country: multi-valued discrete
15. income: binary discrete

Tasks

B.1: There are some missing values listed as "?". Describe your strategy for treating missing values, but note sometimes it is OK to leave missing value as a separate categorical value (we call this "missing informative"). Note there are too many to edit by hand, so if you wish to modify them, identify them with a Boolean test like

```
data$workclass[id]=='?'
```

and modify the values in a loop.

B.2: With all variables, build a model usingthe

```
glm(income∼.,family=binomial, data=???)
```

routine in R, and then print the summary of the model to get the R diagnostics. Briefly explain the statistics in the summary, e.g. Z-value, standard error, p-value. What does this imply about the predictors for your model? Notice many of the variables are multi-valued categorical, and in most cases only some of the values are significant.

B.3: Test the fitted model using the "adult_income_test.csv", and calculate the confusion matrix on the test set, reporting it. Also, give the precision, accuracy and recall (Lecture 3). Note the test set has no missing values.

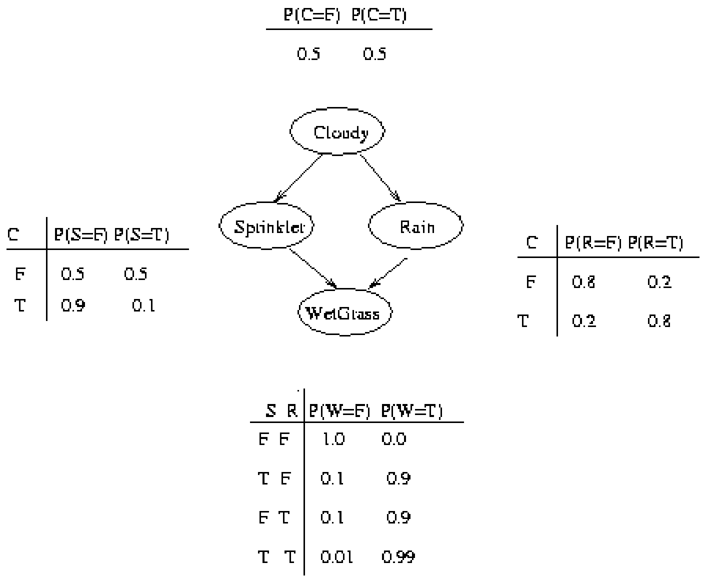

Figure 1: Simple Bayesian Network

B.4: Can you improve your model with different predictors? For instance, you might reconstruct the categorical features to only include significant values and then have an "other" value that groups together all non-significant ones. Perhaps the best way to do this is to create a new data frame with your modified attributes and build the model on that using the R construct "`income~.`" Report the R diagnostics and the confusion matrix and other scores on the test set (as per B.3) for the new model and comment on the difference.

Task C: Sampling

This task will perform two different sampling exercises. For this, consider the Bayesian network defined in

C.1: In R, develop functions to generate samples from the probability distribution defined by

$$p(x|\lambda) = \lambda e^{-\lambda x}$$

for $x > 0$ and the single case when $\lambda = 1.5$. First write a sampling algorithm that uses the rejection method. In your report describe how this was designed. Define this in R, sample 1000 values and histogram them.

C.2: Second write a sampling algorithm that uses the inverse sampling method. In your report describe how this was designed. Define this in R, sample 1000 values and histogram them.

C.3 The simple Bayesian network of Figure 1 has the joint probability distribution

$$p(cloudy)p(rain|cloudy)p(sprinkler|cloudy)p(wetgrass|sprinkler, rain)$$

Use this to write a Gibbs Monte-Carlo sampler for the distribution. Run the sampler for 1000 cycles, throwing away the first 100 samples, and record counts for the two tables of

$$p(wetgrass, cloudy) \qquad\qquad p(sprinkler, rain)$$

Convert the counts to probabilities and report them.