Data Modelling and Analysis With Feedback

Coursework: 2020/2021

Dr Mercedes Torres Torres

Contents

Feedback Overview

This document presents general feedback on the Coursework Submission. I will cover: a) general performance in the coursework and b) common mistakes.

Common general errors have included:

- Disorganised report.
- Tables or Figures not captioned and/or not cited in the text.
- Raw code or implementation details included in the text.
- Screenshots or code dumps included in the text.

There were 174 submissions. The maximum mark is 99, and the minimum is 6. The mean is 58.55, and the median is 60.66. Q1 is 50.36 and Q3 is 68.88.

Instructions

This coursework is organized into three parts, each one focusing on a different and important aspect of Data Analysis and Pre-processing, Data Mining, and Data Classification, respectively. All parts involve the use of the same dataset.

The first part focuses on describing and visualizing the data and preparing it for subsequent treatment. The second part focuses on studying the effects of combining different pre-processing methods and different clustering approaches. Finally, the third part focuses on classification and prediction.

The ultimate goal of this coursework is to give you first-hand experience on working with a relatively large and real data set, from the earliest stages of data description to the later stages of knowledge extraction and prediction.

Data Set

The data set is a slightly modified version of a real-world dataset: the **Sloan Digital Sky Survey (SDSS) DR14 dataset**.

The dataset consists of over 10,000 observations of space taken by the SDSS. Every observation (rows) is described by 21 features (columns). Features combine photometric and spectral information. Finally, there is one class column, which identifies each sample as a *star*, *galaxy* or *quasar*.

Feature Description

- *objid*: Object Identifier
- *dia*: approximated diameter of identified object.
- *rereun*: Rerun Number. Specifies how the image was processed.
- *ra*: J2000 Right Ascension (r-band). The angular distance measured eastward along the celestial equator from the Sun at the March equinox to the hour circle of the point above the earth in question.
- *dec*: J2000 Declination (r-band). Declination. When combined with *ra*, these astronomical coordinates specify the direction of a point on the celestial sphere (i.e. the sky) in the equatorial coordinate system.
- *u*: magnitude (ultraviolet)
- *g*: magnitude (green)
- *r*: magnitude (red)
- *i*: magnitue (infrared)
- *z*: magnitude (infrared). The Thuan-Gunn astronomic magnitude system. *u*, *g*, *r*, *i*, *z* represent the response of the 5 bands of the telescope.
- *run*: Run Number. Identifies the specific scan (i.e. image) from which these measurements were extracted.
- *m_unt*: Uncertainty associated with *dia* measurement.
- *flux*: spectra information.
- *native*: True if instance was collected in the first round of data collection. False if otherwise.
- *camcol*: Camera column. A number from 1 to 6, identifying the scanline within the run.

- *field*: Field number. Part of the image from which these measurements were extracted.
- *redshift*: Final Redshift. Redshift happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum.
- *plate*: plate number
- *mjd*: MJD of observation. The date that a given piece of SDSS data (image or spectrum) was taken.
- *fiberid*: fiber ID. Optical fibers used to direct the light at the focal plane from individual objects to the slithead
- *class*: object class (galaxy, star or quasar object)

The problem presented is a *grey system*, in which you have been given *some* information about the data and classes, but not all of it.

Software

You are required to only use R and Weka, as indicated in the details below.

Data Report Deliverable

You will need to submit a written report describing all the analysis conducted. The *length for the report should be a maximum of 2500 words (excluding tables and figures) and twenty sides of A4 (excluding the cover, contents page and appendix, but including all tables and figures)*. These limits are not flexible.

Number all of your pages and make sure to include your name and student ID on the front page.

The minimum font size allowed is 11pt (a full page of text in a similar style to this document would contain about 500 words, so the majority of the 20 sides will be tables and figures). The report should clearly explain what you did with the data, how you did it and why you did it, and it should be well structured and illustrated.

Your report should contain three sections in total as marked below (*Analysis and Preprocessing*, *Clustering*, *Classification*). You cannot include any code, or raw output (e.g. the output of R commands, screen-captures of the results, etc.) in the main body of your report. Include a copy of your code in an Appendix. Note that appendices will not contribute to the word count and are not explicitly marked: they are for reference only.

Marks and Assessment Criteria

Part 1 carries 40 marks, while Part 2 and Part 3 carry 30 marks each. In total, this coursework aggregates to 100 marks and accounts for 75% of the module. Marks will only be awarded for the first twenty pages of the main body of your report.

The main assessment criteria for the report are:

- *Correctness*: that is, do you apply techniques correctly; do you make correct assumptions; do you interpret the results in an appropriate manner; etc.?
- *Completeness*: that is, do you apply a technique only to small subsets of the data; do you apply only one technique, when there are multiple alternatives; do you consider all options; etc.?
- *Originality*: that is, do you combine techniques in new and interesting ways; do you make any new and/or interesting findings with the data?
- *Argumentation*: that is, do you explain and justify all of your choices?

Deadline and submission procedure

- The submission deadline is on the *24th of May (Monday) at 15:00*.
- Name your report *COMP4030-Cwk-XXXXXX.pdf*, where *XXXXXX* should be replaced with your student ID number (e.g. DMA-Cwk-4078181.pdf)
- Submit the single PDF document via Moodle (see Moodle page for details).
- Number all of your pages
- **Your full name and student ID have to be shown in the first page of your report.**

Plagiarism and Collusion vs. Group Discussions

As you should know, plagiarism and collusion are completely unacceptable and will be dealt with according to the University's standard policies. Having said this, we do encourage students to have general discussions regarding the coursework with each other in order to promote the generation of new ideas and to enhance the learning experience. Please be very careful not to cross the boundary into plagiarism. The important part is that when you sit down to actually do the data analysis/mining and write about it, you do it individually. Do NOT, under any circumstances, share code, share figures, graphs tables, results or charts, etc.

1 ANALYSIS AND PRE-PROCESSING [R ONLY, 40 MARKS]

The overall goal of this section is to get you to start realising the problems within the dataset and how these problems will affect its use in clustering and classification.

1. Explore the data [6]
   i. Provide a table for all the input features of the dataset including measures of centrality, dispersion, and how many missing values each attribute has.
   ii. Analyse the class variable using appropriate statistics and visualisations.
   iii. Produce histograms for each input attribute and characterise all the distributions according to shape. Provide details on how you created the histograms. You may also use descriptive statistics to help you characterise the shape of the distribution.

Important information to be extracted from this exercise includes: the fact that there are two variables with the same value duplicated. There are some variables with a high number of missing values. Additionally, the problem is skewed, as one class appears 50% of the time, while another one appears only around 8%.

Common errors have included:

- Incomplete analysis (i.e. only reporting mean and standard deviation)
- Incorrect metrics (i.e. trying to calculate metrics from clearly nominal attributes)

2. Explore the relationships between the attributes, and between the class and the attributes [8]
   i. Calculate the correlation and produce a scatterplot for the variables: r and g. What does this correlation tell you about the relationships between these variables?
   ii. Calculate the correlation and produce a scatterplot for the variables: *mjd* and r. What does this correlation tell you about the relationships between these variables?
   iii. Produce scatterplots between the class variable and u, z, and *redshift*. What do these three scatterplots tell you about the relationships between these variables and the class?
   iv. Produce boxplots for all of the appropriate attributes in the dataset grouping each variable according to the class attribute.

Important information to be extracted from this exercise includes: the fact that there are very high and very low correlations in the dataset. The importance of *redshift*, which will be useful for detecting QSO. The fact that the dataset is skewed.

Common errors have included:

- Not reporting the correlation coefficients.
- Not producing scatterplots.
- Not reporting the nature of the relationship (a) high positive correlation, and b) no correlation whatsoever.
- Incorrect scatterplots (i.e. trying to plot scatterplots between the variables, instead of between the class and each variable.)
- Not producing the boxplots according to the class.

3. General Conclusions [5]

Take into considerations all the descriptive statistics, visualisations, and correlations you produced previously and comment on the importance of the attributes. Which of the attributes seem to hold significant information and which you can regard as insignificant? Provide an explanation for your choice.

A complete and comprehensive list on the importance of each variable is required. Different ways of showing the information (table, text, etc) are ok, as long as the list is comprehensive and based on evidence.

Common errors have included:

- Not reporting on *dia*
- Not reporting on *redshift*
- Not reporting on *objid*
- Not reporting on all variables
- Not linking your classification to evidence produced in the previous problems.

4. Dealing with missing values in R [6]

Replace missing values in the dataset using three strategies: replacement with 0, mean and median. Define, compare and contrast these approaches, and explain their effects on the data. For mean and median replacement, take the class of the instances into consideration.

In this task, you are expected to address all three elements of the question: definition, comparison and contrat between all three approaches. Explanations on their effects using the data is also required.

Common errors have included:

- Not comparing and contrasting the effects of the techniques.
- Not showing the effects on the data.

5. Attribute transformation [6]

Using the three datasets generated in 1.4, explore the use of three transformation techniques (mean centering, normalisation and standardisation) to scale the attributes. Define, compare and contrast these approaches, and explain their effects on the data.

In this task, you are expected to address all three elements of the question: definition, comparison and contrat between all three approaches. Explanations on their effects using the data is also required.

Common errors have included:

- Not comparing and contrasting the effects of the techniques.
- Not showing the complete effects on the data.
- Using formulas, but not defining the elements of the formulas.

6. Attribute / instance selection [4]

    i. Starting again from the raw data, consider attribute and instance deletion strategies to deal with missing and duplicated values. Choose a number of missing values per instance or per attribute and delete instances or attributes accordingly. Explain your choices and its effects on the dataset.

ii. Start from the raw data, use correlations between attributes to reduce the number of attributes. Try to reduce the dataset to contain only uncorrelated attributes and no missing values. Explain your choices and its effects on the dataset.

In this task, you are expected to create two new versions of the datasets focused on eliminating empty values. You will be driven by two different methods: a) deletion and b) correlation.

Common errors have included:

- Not showing the effects on the dataset.
- Not explaning choices.

7. Attribute transformation / reduction [5]

Starting from the raw data, perform appropriate pre-processing steps first, and then use Principal Component Analysis. Explain your process, along with the results obtained.

i. Compare the effects of Principal Component Analysis when looking at PCA as a transformation technique (i.e. considering all PCs) and as a dimensionality reduction technique in which the data will be reduced to 12 dimensions (i.e: PC1-PC12).

ii. How many PCs should be used to obtain a cumulative variance of at least 90%?

Common errors have included:

- Not explaning which dataset was used in the process.
- Not comparing PCA as a transformation and as a dimensionality reduction technique.
- Not shwoing evidence of PCA as a transformation and as a dimensionality reduction technique on the dataset chosen.
- Not showing the Cumulative Proportion of the PCs and/or any extra information providing evidence on the performance of PCA.

As a result, you will end up with several different sets of data to be used in Sections 2 & 3. Give each set of data a clear and distinct name, so that you can easily refer to again in the later stages.

2 CLUSTERING [R ONLY, 30 MARKS]

Using only R, explore the use of clustering techniques to find natural groupings in the data, without using the class variable – i.e. use only the appropriate input attributes to perform the clustering. Once the data is clustered, you may use the class variable to evaluate or interpret the results (how do the new clusters compare to the original classes?).

1. Choose an appropriate dataset and use HCA, k-means, and PAM as clustering algorithms to create groupings of three clusters and write the results. Which dataset have you used? Use a combination of internal and external metrics to evaluate which algorithm produces better results. Describe the metrics and how you calculated them [10].

Common errors have included:

- Not explaning which dataset was used in the process and/or why.
- Using an incorrect dataset (i.e. the dataset with the class).
- Not defining the internal or external metrics
- Showing incomplete results.
- Not interpreting the results, including per-class/per-cluster behaviour.
- The use of incorrect/not-useful visualisations in the reporting of the results.

2. Using the dataset from the previous task, optimise each clustering method according to two parameters or more. Which parameters did you choose? Define them. Using the same metrics as in the previous exercise, which parameters produced the best results for each clustering algorithm? Provide the reasoning for the techniques you used to find the optimal parameters [10].

Common errors have included:

- Not explaning which dataset was used in the process and/or why.
- Using an incorrect dataset (i.e. the dataset with the class).
- Choosing incorrect parameters. For example, choosing to optimise k is incorrect, as you are being asked to consider the same metrics as in the previous exercise (including external and internal).
- Not defining the parameters.
- Not carrying out comprehensive optimisation/evaluations.
- Choosing different metrics from the previous exercise.
- The use of incorrect/not-useful visualisations in the reporting of the results.
- Not interpreting the results, including per-class/per-cluster behaviour.

3. Choose one clustering algorithm of the above and a combination of internal and external metrics, then perform clustering on the following alternative datasets which you have produced in Part 1 [10]:

   i. The transformed dataset featuring all Principal Components
   ii. The reduced dataset featuring 12 Principal Components.
   iii. The dataset after deletion of instances and attributes.
   iv. All three versions of the mean-centred data
   v. Which of these datasets had a positive impact on the quality of the clustering? Provide explanations using the results for each clustering of the alternative data set.

Common errors have included:

- Not carrying out comprehensive evaluations.
- Choosing only internal or external metrics.
- The use of incorrect/not-useful visualisations in the reporting of the results.
- Not interpreting the results, including per-class/per-cluster behaviour.

3 CLASSIFICATION [WEKA ONLY, 30 MARKS]

You must use Weka to perform the classification, but you may use R to present results. Use Weka classification techniques to create models that predict the given class from the input attributes.

1. Choose an appropriate dataset to obtain predictions using the following classifiers: ZeroR, OneR, NaïveBayes, IBk with 5 neighbours (5-NN) and J48 (C4.5). Which dataset have you used and why? Choose an evaluation protocol that would prevent any possible overfitting problems. Which evaluation protocol did you use? Use a combination of metrics to justify your reasoning. Which algorithm produces the best results? [10].

Common errors have included:

- Not explaning which dataset was used in the process and/or why.
- Using an incorrect dataset (i.e. the dataset with the class).
- Not explaning which method was used to avoid overfitting. Or, saying that you have chosen k-fold, but not defining how many folds were used and why.
- Showing only overview results.
- Not interpreting the results, including per-class behaviour.

2. Choose the same dataset as in the previous task and experiment with IBk (k-NN) with the goal of optimising it according to: a) train/test split percentage, b) number of neighbours, and c) distance metrics. Describe each parameter. Show all results obtained in the experiment according to accuracy, confusion matrix, precision, and recall. Which combination of parameters produced the best result? [10]

Common errors have included:

- Not carrying out comprehensive optimisation
- The use of incorrect/not-useful visualisations in the reporting of the results.
- Not interpreting the results, including per-class behaviour.

3. Apply J48 to the datasets listed below using 10-fold cross-validation. Provide explanations on the performance of the datasets using a combination of metrics [10].

    i. The transformed dataset featuring all Principal Components
    ii. The reduced dataset featuring 12 Principal Components.
    iii. The dataset after deletion of instances and attributes.
    iv. All three versions of the normalised data
    v. Which of the datasets had a good impact on the predictive ability of the algorithm? Provide explanations using the results for each classifier of the alternative data set.

Common errors have included:

- Not evaluation all datasets
- Not interpreting the results, including per-class behaviour.