COMP90051 Statistical Machine Learning

Semester 2, 2015 – Project 2 Spec

Due date: Tuesday, the 16th of October 2015 at 5:00 PM

Weight: 20% or 25% depending on mid-semester test performance (whichever is higher)

Note: This project forms a part of your hurdle requirements

1. Overview

Australian suburbs are geographic subdivisions mainly used for addressing purposes. Greater Melbourne is subdivided into several hundreds of suburbs, home to 4.4 million people spread over 10,000 km^2. There is a large diversity across suburbs that originates from geographic and historical reasons. Suburbs differ in a number of aspects ranging from land area to population demographics. Understanding this diversity, learning from it and using it to the advantage of community presents a major challenge for the government.

The first step in addressing this challenge is collecting statistical information about suburbs, and this has been performed on a regular basis. Moreover, Victorian government supports the policy of open data, whereby accumulated statistical profiles are available to public free of charge. In particular, Department of Health and Human Services (DHHS) has released statistical profiles for suburbs. The task of understanding similarities and differences between the profiles now becomes a data analytic challenge.

In this project, you will take on the challenge, addressing the problem in two parts. The first part will be driven by a specific research question, while the second part will be an exploratory open-ended analysis. You will perform unsupervised data analysis using statistical suburb profiles available from DHHS. You will work in a team of 3 people and, as a team, you will submit a report about your findings.

2. Dataset

We have downloaded randomly chosen 34 suburb profiles, and providing these data as a single archive distributed from the LMS (Assessment tab → Project 2 → link to data). Note that statistical profiles is open access data available from DHHS website, but for the purposes of this assignment you should use the provided archive. This is to ensure that everyone has the same dataset as a starting point.

Statistical profiles are contained in Excel files, with a separate file for each suburb. Each profile comprises over 200 measurements (features) that describe a suburb from a range of different angles. For example, some features show land usage structure, while others show population age structure and number of available health care facilities. Note that there is a definitions tab in each Excel file (these definitions are the same in every file). This tab provides additional information to help with feature interpretations. Reading all detailed definitions is not required, rather these can be used if the meaning of a particular feature is not clear.

Also note that each Excel file contains a hidden tab titled "what is this". This tab should be ignored.

Data Attribution

Copyright: Department of Health and Human Services

License: Creative Commons Attribution 3.0 Australia

Source: The data was downloaded from http://www.health.vic.gov.au/ in mid-2015, and used without modification

3. Part A: Hypothesis driven research

Identifying similarities between profiles has a range of applications. For example, local governments may introduce new or adjust existing policies based on the experience of other similar suburbs that already have the policy in place. It appears reasonable to assume that suburbs that are geographically close to each other have similar profiles. However, this intuition might not be supported by data. In this part of the project, you will interrogate data to find evidence for or against the hypothesis that geographically close suburbs are similar. As it is often the case with data analysis, there is not necessarily a single correct answer! The important point is that whichever conclusion you reach should be carefully justified by findings from data.

A1: Define 3 similarity measures that characterize how similar two suburbs are. Each measure should focus on similarity from a particular perspective (e.g., demographics). There should be little overlap between the measures. Each measure does not have to use all of the features, but should make an adequate use of available data. Introduce each measure, its aim, and justify the choice of features and the method for combining them.

A2: Use a multidimensional scaling method to generate 2D scatter plots with profiles using each of the defined similarity measures. Describe which aspects of the plots were expected, and which were surprising. Identify which suburbs, if any, remain similar under each measure.

A3: Address the hypothesis that geographically close suburbs are similar. Use [Geography/Location] feature from the data to estimate geographic proximity. Describe the method for proximity estimation. Argue whether the data supports or contradicts this hypothesis, and to what extent the conclusion depends on the choice of a similarity measure.

4. Part B: Exploratory Analysis

From Part A results you will see how the data can provide evidence for or against a specific hypothesis. However, the dataset potentially contains a lot of other interesting and useful information. In this part you will perform an open-ended exploratory analysis of the data. A typical result of such an analysis is a set of observations accompanied by putative explanations and a reflection on how each particular observation can be exploited for public benefit. You may also speculate on social, historical or geographical reasons for your observations, but you should avoid trivial “discoveries”. Examples of trivial “discoveries” to avoid are “suburb A is larger than suburb B”; “Tullamarine is close to the airport, that’s why it is different”.

B1: Report trends, frequent patterns or clusters discovered in the data. You will also need to identify a suitable way to present the observations (e.g., table, scatterplot, histogram, etc.).

B2: Suggest an explanation for reported discoveries, possibly, relate facts to each other.

B3: Reflect on why this knowledge required analysis and was not apparent from just eyeballing the data, and reflect/speculate on how this knowledge can be used for community good.

In part B, you may perform any type of analysis that you like, including clustering, association rules, and supervised learning. This flexibility is both an opportunity and a challenge very common in real-world data analytics applications. In order to help you to get started we are providing several example scenarios of how you might structure your analysis. Remember that you do not have to follow any of these scenarios.

Example 1: elaborate on findings from A1 and A2. Identify outliers and figure out which measurements make certain suburbs so distinct. Explore whether a suburb that looks like an outlier in one of the plots (using one of the measures) remains an outlier in all plots. Propose more similarity measures (in addition to A1), and investigate whether they produce unexpected clustering.

Example 2: explore the relation between the features. Choose a particular variable (e.g., proportion of unemployed population) and figure out if you can predict this variable from the other features. Identify, which of the other features are largest contributors to the quality of predictions.

Example 3: focus on a particular suburb, and centre your analysis on that suburb. Explore relations between this suburb and other suburbs. Identify in which regards this suburb is similar, and in which regards it is different from others.

5. Submission Format

The submission comprises a report in a PDF format, accompanied by source code. The report is up to 5 pages in length. The text must fit in the first 2 pages. Tables, figures and references must fit in the remaining 3 pages. Tables are optional. Figures should be numbered, and each figure should be referenced from at least one place in the text. Text labels within figures must be legible (avoid tiny fonts or blurry text).

The accompanying source code should have instructions on how to run the code in order to reproduce each of the figures and/or tables. These instructions should be in a `readme.txt` file in the code folder, and not be a part of the report.

You may use any additional external source of data as long as it is properly acknowledged. A justified use of additional data will attract a bonus mark. However, you are not required to use external data. Note that the analysis should be mainly based on the provided dataset (distributed from LMS), and any external data (if used) should be restricted to providing supplementary information.

Plagiarism policy: You are reminded that all submitted project work in this subject is to be your own individual work. Automated similarity checking software will be used to compare submissions. It is University policy that cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student(s) concerned.

For more details, please see the policy at http://academichonesty.unimelb.edu.au/policy.html.

6. Assessment

The project will be marked out of 20. The project’s weight towards your final mark for the subject could be either 20% or 25% depending on your mid-semester performance. Late submissions of reports will incur a deduction of 4 marks per day or part of thereof.

The total project mark is the sum (capped to 20) of the components listed below. Part A is scored out of 8 using the following marked scheme.

	Maximum mark	**Assessment Criteria**
A1	2	The measures address different aspects of suburbs (rather than largely repeating each other). Selection of features and the method for combining them is well motivated for each measure.
A2	3	Conclusions are justified and supported by data
A3	3	Method for estimation of geographic proximity is justified. Conclusions are justified and supported by data
Total Part A	8	

Part B is scored out of 8 using the following marking scheme.

	Maximum mark	**Assessment Criteria**
B1	3	Observations are non-trivial (e.g., cannot be discovered by a quick glance at Excel tables). The method for presenting the data is justified.
B2	3	Explanations make sense
B3	2	Conclusions are justified and supported by data
Total Part B	8	

Report presentation is scored out of 4 based on clarity of descriptions, coherence of arguments, and clarity of figures.

A **bonus mark** (maximum of 1) is given for a justified use of external data.

Based on our experimentation with the project task and the design of the marking scheme below, we expect that all reasonable efforts at the project will achieve a passing grade or higher. So relax and have fun!

END OF PROJECT 2 SPEC