COMP90051 Statistical Machine Learning Project 2 Report

Liang Han (712397), Siqi Liu (669910), Yumeng Shi (664891)

Part A. Hypothesis Driven Research

1. Problem and Hypothesis

Greater Melbourne is a multi-cultural region consisting of more than 300 suburbs, which are divided mainly by the geographic locations. Each suburb differs from others in various aspects. It is reasonable to hypothesize (H_0) that suburbs that close to each other will have similar profiles. To test this hypothesis, this experiment adopts a dataset of 34 suburbs (S_1, S_2, …, S_{34}) and 225 features (X_1, X_2, …, X_{225}).

2. Data Preprocess and Methodologies

a) Data Preprocess: Many of the features are categorical, not confronting the requirement to perform similarity calculation and multidimensional scaling, while there are also many missing values. In order to clean the data, first, we convert all the categorical features into numerical. Two different strategies are applied: a) to handle lone categorical variables, convert the categorical features into binary (dummy) variables; b) to handle categorical variables with corresponding values, make a complete matrix of the numeric values. Second, we replace all the missing fields and '<5' labels with value 0.

b) Methodologies: For the similarity measure defining, we first choose 3 different aspects that people would most care when considering where to live. For each aspect, we consider the weights of each feature, and the way combining them. For the Multidimensional scaling, we first apply the metric multidimensional scaling to map the high dimensional feature space into a 2D plane and examine the goodness of fit. Afterwards, we use centroid based clustering technique (K-means clustering) to put 34 suburbs into several groups and compare with the geographic location segmentations in order to test our hypothesis.

3. Similarities Measures Definition (A1)

a) Cultural Diversity Similarity: In reality, people from same country tend to live close to each other (probably due to post-war immigrant waves), and from another aspect, some suburbs tend to have unitary cultural background, while others will have more residents coming from overseas. The cultural diversity measure considers the ratio of residents born overseas, their LOTE languages in percentage (31 features in total). We ignore the birth country features because it is highly correlated to the language features.

Weight: Our similarity measure uses the percentage figures, therefore, the ratio of residents born overseas feature will have higher weight as it is the sum of residents from all other countries. While the combinations of birth countries and languages will be less weighted, as the figures are mostly close to 0%. Also, this measure put more weights on the larger birth countries and languages, such as Italy, India, China, and less weight to minor ones. This weighting strategy makes sense because the cultural diversity is often 'decided by' people from these countries.

b) Educational Similarity: Educational resources are one of the most important resources, parents now are willing to pay up to 32.8% more to get a property near top schools. On the other hand, suburbs such as Parkville, with two major universities nearby has high percentage of student residents and probably with higher education level. In general, the education similarity measure considers student numbers of different levels, school numbers, number of people working in education and training (37 features in total).

Weight: Before applying the similarity calculation, it is necessary to rescale the data into (0, 1) range. This will make features describing the number of schools and students more important, since more than 1 feature will contribute to these categories.

c) Medical Similarity: Medical resources are often considered as the most important element for a community of elder people. We choose number and size of hospitals, number of aged people and number of workers in Health Care and Social Assistance category in the suburb as features to get the similarity measure (42 features in total). This similarity measure aims to compare the suburb with similar amount of medical requirements and resources.

Weight: We also need to normalize the data, but for the categorical data (hospitals), less weights are allocated, because hospital names do not matter in this case.

Combine Method: We choose to use Euclidean distance as the distance function for all three measures, because our features are differently weighted. Euclidean distance preserves the weights better than correlation since it does not scale and it is more intuitive.

4. Data Analysis (A2, A3)

Figure 1.1 – 1.3 in the Appendix is the 2D plots generated from multidimensional scaling using diversity (Goodness of Fit: 0.9344234), educational and medical similarity measures. In cultural diversity aspect, it is as expected, most suburbs have similar degrees of cultural diversity, since the percentage of born overseas people has the largest weight. However, suburbs such as Footscray and Springvale have over 50% of the residents born overseas, making them outside the majority. In educational aspect, we expected that suburbs within or near CBD will have a relatively similar (higher) performance, while remote suburbs would probably have less advantage in educational resources and student numbers. It is also expected that Melbourne Airport which has very few residential lands would probably be an outlier since there is no students nor schools. However, it is surprising to see Parkville did not stand out even though it has the top university. It shows that this measure cares more about the quantity of resource instead of quality. Finally, in medical aspect, it is unexpected to see the Pascoe Vale South with no hospitals, nor health centres, yields a performance similar to most of the near CBD suburbs. For all of the 3 measures, Nobel Park, Waterways, Glenroy, Springvale and Fawkner are all outliers in the plots.

In this experiment, the location is approximated by Map Reference. As shown in the geographic features, Map Reference classifies the 34 suburbs into 4 groups, each group represents a broad map area, which are CBD area, peninsular area, North West and South East

Area. We hence assume that suburbs with same Map Reference category are close geographically. Furthermore, our hypothesis can be restated as suburbs with same Map Reference will have similar profiles. According to the K-means clustering result shown in Figure 1.4 – 1.6 and Table 1.7, diversity measure has 65% overall purity[1], education measure has 59% purity, while medical measure also has a 65% purity in clustering. Also according to the F_1 measure and Rand Index, we can conclude that the measures we define in the previous section have a statistically significant advantages in mapping the suburbs' geographic locations, **therefore, this clustering result supports the hypothesis H_0 to be true**. In addition, it is obvious that diversity similarity measure has the highest scores among the three, while education the lowest. Overall, all of the three measures can substantiate the hypothesis, yet the diversity and medical measure have strong support to the conclusion than the education measure.

Part B. Predicting Population Change and Feature Selection

In the past few years, population of Melbourne increased faster than any other area in Australia[6]. Our task is to predict population change by non-population features[2]. The exploration has two main topics. a) One is to test if population change can be predicted by other features. b) Another is to find the features that contribute most to the quality of prediction. Three main types of feature selection algorithms are performed: Filter, Wrapper and Embedded algorithms. This part is also based on the pre-processed dataset in part A. Prediction is based on all other 418 features after removing population-related features. Lastly, downloaded 365 suburb data from Department of Health and Human Service[7] works as the test data to verify our experiment results.

1. Population Change Prediction

Classification approaches are used to predict the trends of population change, values are reassigned to 0 and 1, which means decrease and increase respectively. a) Linear Support Vector Machine is chosen to classify, as it is a strong classifier and generally gets good performance. But it just gets 85.71% accuracy on cross validation which is less than 88.24% of baseline ZeroR seen in Table 2.1. b) Linear Regression is used to predict the 'exact value' of population change. Linear regression is performed on least square method. The number of features is greater than number of instances which makes $(X^TX)^{-1}$ ill-posed inverse problem.

Linear SVM can predict population change but not ideally. Linear regression fails as a result of too many features. So feature selection need to be applied to make classification optimized and make regression predictable.

2. Feature Filtering

As categorical features are transformed to binary variables but variables with great proportion of 0s or 1s are of no contribution in prediction. So binary variables with low variance need to be removed. Table 2.2 shows that different thresholds are set to remove variables with low variance. Here, a relatively reasonable threshold 0.9, which means the neither the number of 0s nor 1s should be less than 10% of the sample size, is adopted.

3. Wrapper Feature Selection for Classification

Bayesian feature selection is based on recursive feature elimination method. This experiment is default tuned and cross-validated. Figure 2.3 indicates that the optimal number of features is 11 and the selected features and are shown in Table 2.4. Linear SVM is performed again based on the selected features and the accuracy increases greatly to 91.43% as shown in Table 2.5.

Table 2.4 shows that most of the features belongs to Socio-demographic and Diversity domains. As a result, population change trend can be predicted by features from Socio-demographic and Diversity.

4. Embedded Feature Selection for Regression

Lasso Regularization is applied in linear regression to select features. Here lambda values are chosen from 1 to 100. From Table 2.6, it can be seen that regression model gets optimal R^2 as and the inverse problem is well-posed when lambda is set to 70. The penalty of 70 lambda value leave 33 features with a R^2 of 0.75. Similarly, population change in certain suburb is mainly related to socio-demographic and diversity, also affected by medical factors and population density as seen in Table 2.7.

5. Critical Analysis and Conclusions

Selected features both belong to Socio-demographic and Diversity. Because features in these two contains information about population from various points. So these features are effective in predicting population changes for classification and regression.

Sample size of 34 is too small for classification and regression, which may lead some errors. Thus, test experiments are performed on 365 suburbs dataset to verify the robust of the classification and regression models. For classification, the classifier is trained on selected 11 features and then test on the 365 instances test data, the test accuracy is 80.82%. For linear regression, the selected 33 features are selected from the test data and then calculate the R^2, which drops dramatically to 0.37. Results are seen in Table 2.8

Predicting on the population change has great influences on government policy and personal residential choice. According the analysis above, population change of different suburbs can be predicted by other non-population features. The selected features mainly belong to the socio-demographic and diversity aspects. However, as the number of instances is small, the robustness of classification and regression models based on selected features is not satisfactory.

[1] *Purity*: Each cluster is assigned to the most frequent label in the cluster. Purity is calculated as the number of correctly assigned instances dividing by total number.

[2] Predicted feature is '% change, 2007-2012, total' and Non-population features means all the other domain features apart from the population domain features.

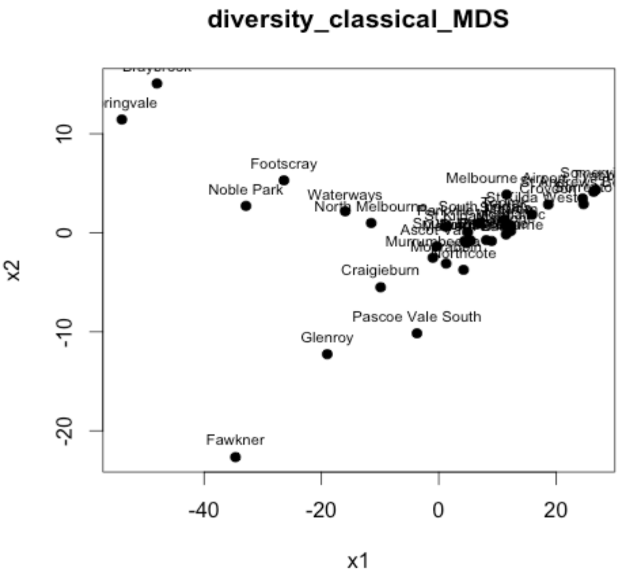

Figure 2.1 MDS Plot: Diversity

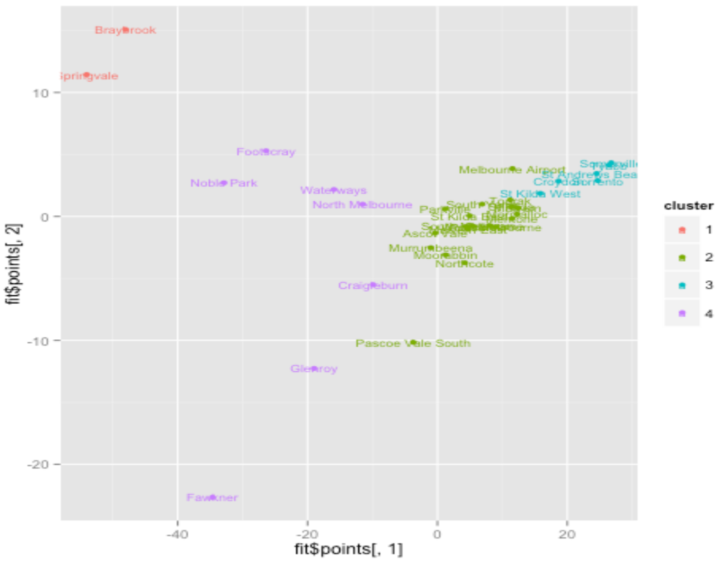

Figure 1.4 Diversity Cluster Result

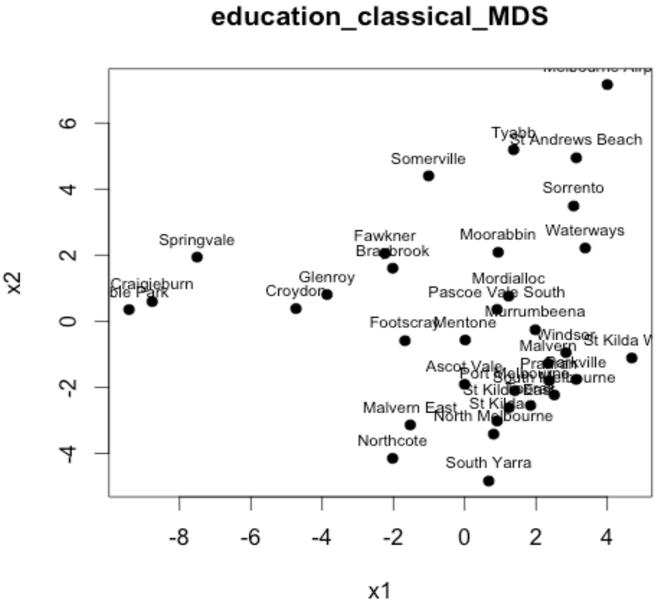

Figure 2.2 MDS Plot: Education

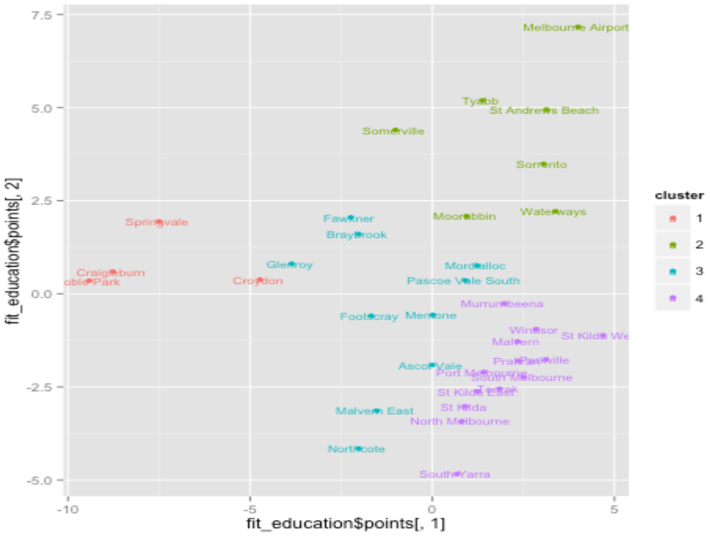

Figure 1.5 Education Cluster Result

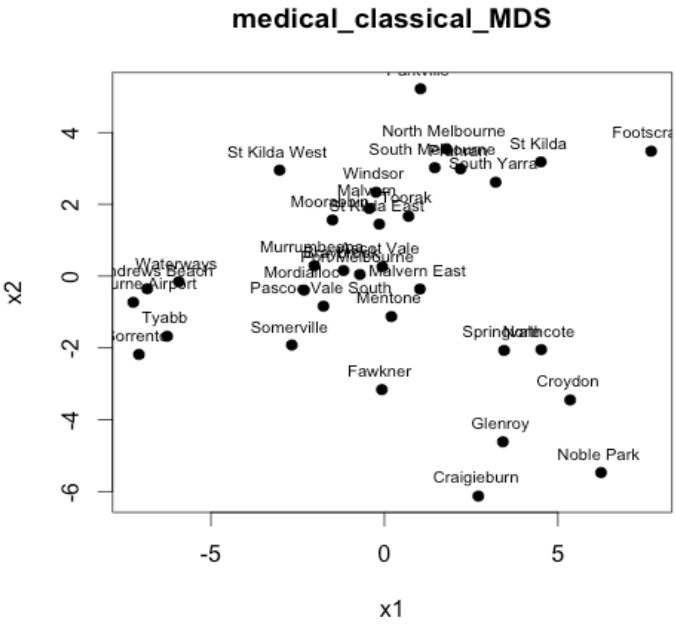

Figure 1.3 MDS Plot: Medical

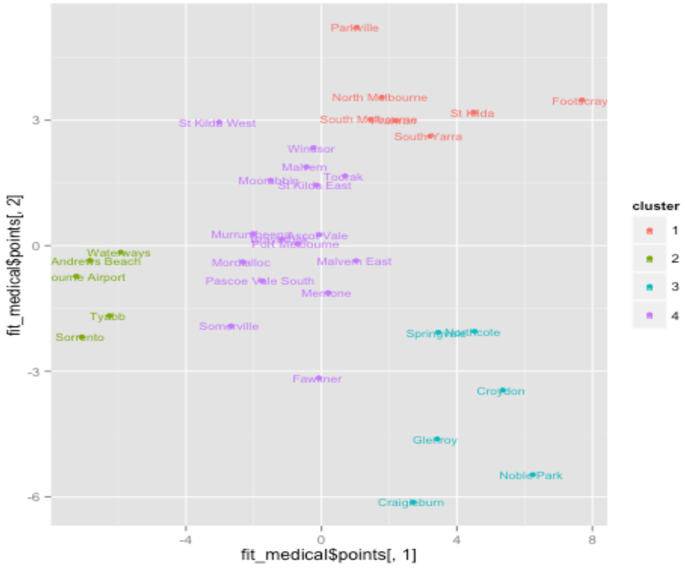

Figure 1.6 Medical Cluster Result

Map		Medical		Education		Diversity	
1	Malvern	1	Ascot	1	Malvern	1	Ascot
	Malvern		Malvern		Mentone		Malvern
	Melbourn		Mentone		Mordiallo		Malvern
	Moorabbi		Mordiallo		Murrumb		Melbourn
	Murrumb		Murrumb		Parkville		Mentone
	Port		North		Pascoe		Moorabbi
	Prahran		Parkville		Port		Mordiallo
	St Kilda		Pascoe		Prahran		Murrumb
	St Kilda		Port		South		Northcot
	St Kilda		Prahran		St Kilda		Parkville
	Toorak		South		Toorak		Pascoe
	Windsor		South		Windsor		Port
	South		St Kilda	2	Braybroo		Prahran
	South		St Kilda		Craigiebu		South
2	Craigiebu		St Kilda		Croydon		South
	Croydon		Toorak		Fawkner		St Kilda
	Mentone		Windsor		Glenroy		St Kilda
	Mordiallo	2	Springval		Noble		Toorak
	Noble		Craigiebu		Springval		Windsor
	Springval	4	Braybroo	4	Ascot	2	Braybroo
	Waterwa		Croydon		Footscray		Springval
4	Braybroo		Fawkner		Malvern	4	Craigiebu
	Ascot		Footscray		North		Fawkner
	Fawkner		Glenroy		Northcot		Footscray
	Footscray		Malvern		South		Glenroy
	Glenroy		Northcot		St Kilda		Noble
	North	5	Melbourn		St Kilda		North
	Northcot		Moorabbi	5	Melbourn		Waterwa
	Parkville		Noble		Moorabbi	5	Croydon
	Pascoe		Somervill		Somervill		Somervill
5	Somervill		Sorrento		Sorrento		Sorrento
	Sorrento		St		St		St
	St		Tyabb		Tyabb		St Kilda
	Tyabb		Waterwa		Waterwa		Tyabb
Purity			0.647059		0.588235		0.647059
F-1 measure			0.450549		0.400000		0.552486
Rand Index			0.682540		0.689840		0.711230

Table 1.7 Cluster Result Validation

Classifier	ZeroR	LinearSVM
Average Accuracy	88.24%	83.21%

Table 2.1 Accuracy comparison of LinearSVM

Threshold	0	0.95	0.9	0.85
Number of remaining features	419	303	196	163

Table 2.2 Number of remaining features with different threshold

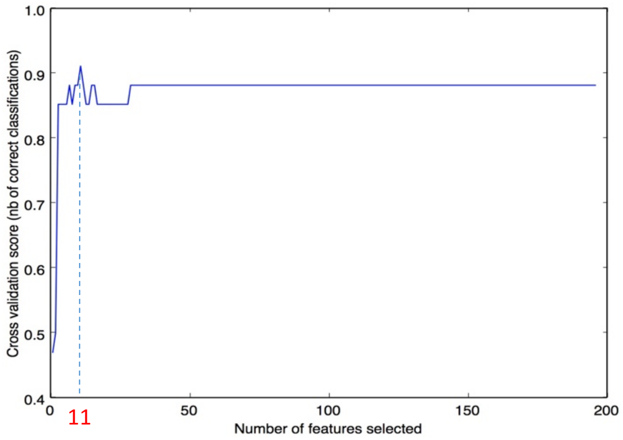

Figure 2.3 Accuracy of different size of feature subset

Feature	Feature Domain
Occupied private dwellings	Socio-demographic
Population in non-private dwellings	
Personal income <$400/week, persons	
Primary school students	
Requires assistance with core activities, persons	
Aged 75+ and lives alone, persons	
Unpaid carer to person with disability, persons	
Born overseas, persons	Diversity
Born in non-English speaking country, persons	
Speaks LOTE at home, persons	
Public hospital separations, 2012-13	Hospital

Table 2.4 Optimal feature subset of RFE

	Before	After
Accuracy of LinearSVM	83.21%	91.43%.

Table 2.5 Accuracy comparison before and after feature selection

Lambda	R^2	Feature Selected
1	0.9992	62
20	0.9619	47
50	0.8424	38
70	0.7504	33
80	0.6987	33
100	0.5955	28

Table 2.6 Lasso estimation with different lambda

Feature Domain	Feature	Coefficient
Geography	Population Density	-1.76688054E-02
Services	Aged Care (High Care)	1.46681695E-01
Socio-demographic	Number of Households	-8.78455962E-04
	Occupied private dwellings	-7.62660611E-03
	Population in non-private dwellings	-1.01158286E-02
	Public Housing Dwellings	1.86628080E-02

	Dwellings with no motor vehicle	8.33213629E-02
	Equivalent household income <$600/week	-2.77970189E-02
	Personal income <$400/week, persons	-1.69824150E-02
	Number of families	-5.76313886E-03
	% residing near PT	6.7503764E-01
	IRSD (max)	1.10097759E-01
	IRSD (avg)	6.76711464E-01
	Secondary school students	1.05130213E-01
	Holds degree or higher, persons	-2.62631065E-03
	Did not complete year 12, persons	3.42221423E-02
	Unemployed, persons	-4.39382435E-02
	Aged 75+ and lives alone, persons	-1.1212599E-01
	Unpaid carer of children, persons	-3.930637E-02
Diversity	Born overseas, persons	3.69493088E-03
	Born in non-English speaking country, persons	-1.60100663E-02
	Speaks LOTE at home, persons	-1.34107667E-03
	Speaks LOTE at home, %	1.02117439E+00
	Poor English proficiency, persons	6.24794402E-02
	2nd top country of birth, persons	1.97796127E-03
	3rd top country of birth, persons	6.09123568E-03
	5th top country of birth, persons	2.43014608E-01
	Top language spoken, persons	4.77791045E-03
	2nd top language spoken, persons	3.76290148E-02
	3rd top language spoken, persons	-1.15527383E-01
	4th top language spoken, persons	-8.65936319E-02
Hospital	Public hospital separations, 2012-13	-6.72059974E-03
	Presentations to emergency departments, 2012-13	4.49280535E-03

Table 2.7 Lasso feature selected with optimal Lambda = 70 (intercept=-840.665315017)

	Training Dataset	Testing Dataset
Accuracy of Linear-SVM	91.43%	80.82%
R^2 of Linear Regression	0.7504	0.37

Table 2.8 Results on testing dataset

References

[1] Power, E. (2015). *Property prices in popular school enrolment zones rocket to top of the class.* Retrieved from http://www.domain.com.au/news/property-prices-in-popular-school-enrolment-zones-rocket-to-top-of-the-class-20150716-gid279/

[2] Wilkinson, L. (2000). Multidimensional Scaling. *Systat 10 Statistics II*, 185 – 214.

[3] Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary reviews.Cognitive Science*, *4*(1), 93–103. http://doi.org/10.1002/wcs.1203

[4] Tarkiainen, L., Martikainen, P., Laaksonen, M., & Leyland, A. H. (2010). Comparing the effects of neighbourhood characteristics on all-cause mortality using two hierarchical areal units in the capital region of Helsinki. *Health & Place*, *16*(2), 409–412. Retrieved from http://doi.org/10.1016/j.healthplace.2009.10.008

[5] Manning, C. D., Raghavan, P. & Schütze, H. (2008). Introduction to Information Retrieval. *Cambridge University Press.*

[6] Milman, O. (2014). Melbourne's population is Australia's fastest growing. Retrieved from http://www.theguardian.com/world/2014/apr/07/melbournes-population-is-australias-fastest-growing-says-abs

[7] http://www.health.vic.gov.au/

[8] Hansen P C, O'Leary D P. The use of the L-curve in the regularization of discrete ill-posed problems[J]. *SIAM Journal on Scientific Computing*, 1993, 14(6): 1487-1503.