Suggested Exercises and Guided Projects

INTRODUCTION TO MACHINE LEARNING

Between Sessions

Session 1 to Session 2

- complete the exercises of session 1
- download the datasets from the website
- read Programming Primer
- install R / RStudio (Posit)
- install the following R packages: dplyr, tidyverse, ggplot2, arules, arulesViz, rpart, rpart.plot, rattle, party, flexclust, e1071, psych

Session 2 to Session 3

- complete the exercises of session 2

Session 3 to Session 4

- complete the exercises of session 3

After Session 4

- complete the exercises of session 4
- attempt the guided projects

Guided Project I

This project uses the Gapminder Tools.

1. In the default configuration, we can identify some potential association rules. Using visual and ballpark estimates, evaluate the performance of the following rules:
   - Income > 8000 → Life Expectancy > 70
   - Income < 8000 AND Life Expectancy < 70 → World Region = Africa
2. Play around with various charts and identify/evaluate 5+ additional AR.
3. Identify groups of similar countries, in 2018 [validate your clusters using various charts]. Were they also similar in 1930? 1970? 2000?
4. In the default configuration, follow the trajectories of Finland, Sweden, Iceland, Norway, and Denmark between 1900 and 2018. Do the countries appear to follow similar trajectories? Are there outliers or anomalous trajectories?
5. Repeat step 4 for Brazil, Paraguay, Uruguay, Venezuela, Colombia, Peru, and Ecuador.
6. Based on your results in steps 4 and 5, would you expect the trajectory for Argentina to be more like those of the Nordic countries or those of the South American countries? Or perhaps neither? Is your answer the same over all time horizons?

Guided Project II

Select a dataset from the list below (or any other set of interest to you):

- GlobalCitiesPBI.csv
- 2016collisionsfinal.csv
- HR_2016_Census_simple.xlsx
- custdata.tsv

For your dataset(s):

1. Perform the appropriate data understanding, data preparation, data cleaning, and data exploration steps to allow you to determine if it is trustworthy and what it could be used for (see Guided Project IV [*Data Science Essentials*] and Guided Project III [*Data Visualization and Dashboards*].
2. Conduct an association rule mining analysis of the datasets, determining 10-20 strong association rules. Visualize them, validate them, and interpret their results.

Guided Project III

Consider the Algae Bloom Dataset (see this example). We try to build a model to predict the presence/absence of algae based on various chemical properties of river water. The data science motivation for such a model is simple: chemical monitoring is cheap and easy to automate, whereas biological analysis of samples is expensive and slow. Another reason is that analyzing the samples for harmful content does not provide a better understanding of algae drivers: it just tells us which samples contain algae.

1. Load the data and summarize/visualize it: you will be tasked with predicting the presence/absence of algae a1 and a2.
2. Clean the data and impute missing values, as needed.
3. Remove 20% of the observations and save them to a validation set.
4. Create a training/testing pair on the remaining 80% of the observations and train 2 decision trees to predict the presence/absence of algae a1 and a2, respectively. Evaluate the performance of each model. Which models performs best on your training/testing pair?

Guided Project III (cont.)

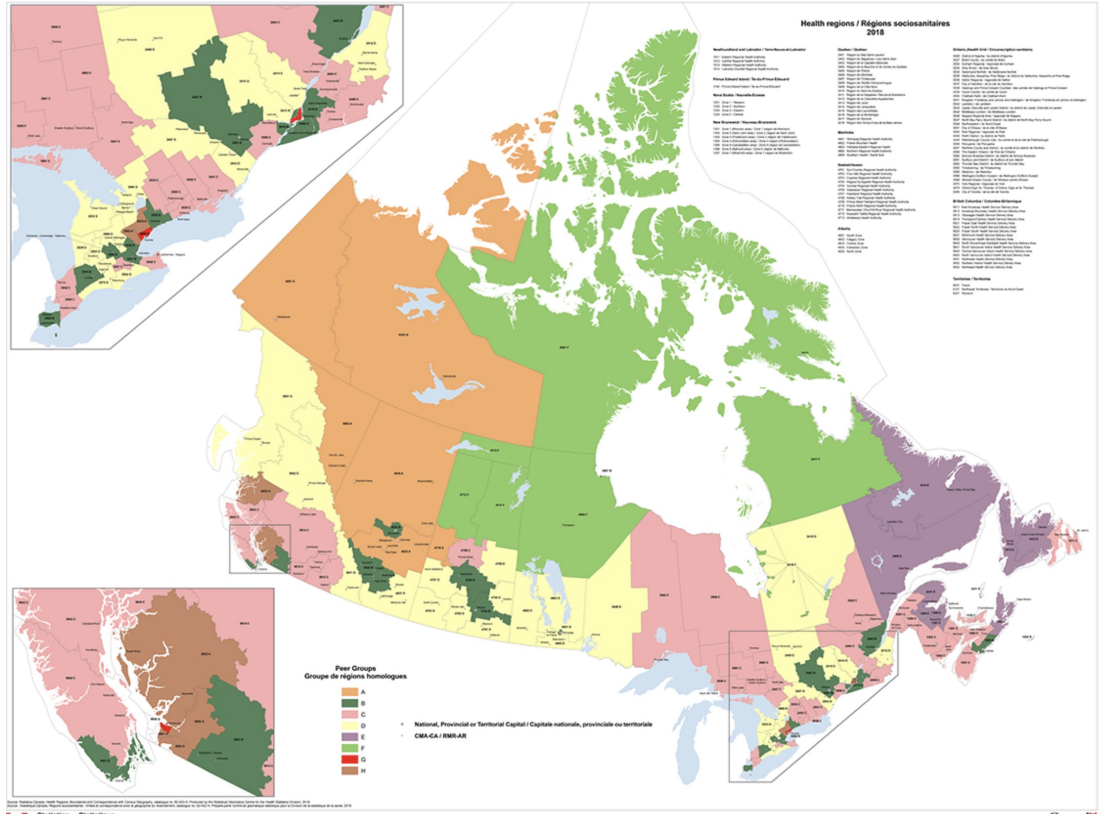
5. Repeat step 4 on at least 20 distinct training/testing pairs. Evaluate the performance of each model, and save them.
6. For each algae, pick the best of the models (how would you determine this) and use it to make predictions for the readings in the validation set. Evaluate these predictions.
7. Instead of picking the best of the 20+ models, find some way to combine the results of the 20 models and to make predictions for the readings in the validation set. Evaluate these predictions.
8. Which of the resulting models of steps 6 or 7 provide the best performance? Which are easier to interpret?
9. Use the same validation set as in step 3. In step 4, use the remaining 80% of the data to build a decision tree (do not split into a training/testing pair first). Use these models to make predictions for the readings in the validation set. Evaluate these predictions. Is there evidence of overfitting?
10. Use the same validation set in step 3. In steps 4 to 7, use decision stumps (decision trees with only 1 branching point) instead of full growth trees. Is there evidence of underfitting?
11. Conduct the analysis steps from 1 to 10 using other classification algorithms. Discuss the results.

Guided Project IV

The population of Canada is divided physically into provincial and territorial areas, most of which are further subdivided into health regions.

The Census information (from 2016) is available for those health regions. The equivalent 2018 dataset has been clustered to produce peer groups: the result is shown here (and on the right).

The data is in HR_2016_Census_simple.xlsx



Guided Project IV (cont.)

1. Load the data and summarize/visualize it (extract the rows with a 4-digit geocode).
2. Clean and scale the data.
3. Run k −means (with Euclidean distance) on the scaled data, using ALL the features, for reasonable value sof k. Use the Davies-Bouldin index and the Within-SS index to determine the optimal number of clusters. Is that clustering scheme plausible?
4. Reduce the dimension of the health region dataset by running a principal component analysis (PCA) and keep the principal components that explain up to 80% of the variability in the data. Repeat step 3. Are the results significantly different than they were?
5. Run k −means on the original health regions data (previous question) and on the reduced data, for the same range of k −values, but replicate the process 30+ times per value of k. What are the optimal k values in the aggregate runs?
6. Save the cluster assignments for each run with the optimal values of k. Two observations A and B have similarity $w(A, B) \in [0,1]$ if A and B lie in the same cluster in $w(A, B)\%$ of the runs. What are some observations with high similarity measurements? With low similarity measurements?