Lab 3

1 Preface

This lab focuses on examining automated system (AS) attributes in order to classify them according to their business type. Identifying the types of organizations who control ASes is important from a research perspective in order to better understand the structure and topology of the internet. AS classification also has many practical business and economic implications; allowing us to better understand the economic interests that drive the structure and reliability of the internet.

There are 19 questions in the lab. Submit your written lab report, including your program code, via the CLE Sakai site by November 25, 2019.

2 Data

PeeringDB.com is a database of networks maintained by volunteers, researchers, and practitioners in the field. The Center for Applied Internet Data Analysis (CAIDA) is an organization that conducts network research (http://www.caida.org/home/). CAIDA collected and archived AS snapshots from 2013 to 2016 (http://www.caida.org/data/peeringdb/) to support AS classification research and testing. The original snapshots were archived in SQL and SQLITE. Your professor downloaded the most recent AS snapshot and converted to a CSV to facilitate your assignment.

2.1 Digging into the data

We would like to better understand our dataset. Download the AS data. Write a Python program to answer the following questions.

1. How many features are in the data?
2. How many unique ASes are in the data?

3. As a result of the SQL join used to construct his data, some of column headers in the CSV are not intuitive. Take a look at the column titled ‘‘`info_traffic`’’ and ‘‘`info_ratio`’’. In a few sentences, explain what you think these columns identify for each AS.

4. Each AS has an associated unique identifier called an AS number (ASN). Let's take a closer look at AS 286. How many prefixes does AS 286 provide? What is its business type? What is its city, state, and zip code?

5. How current (up-to-date) is our data? Do some informal research on AS 286 and see if its current data matches what you found in question 4.

6. For this lab we will focus on classifying ASes according to their business type. The AS business type can be found in ‘‘`info_type`’’. How many unique ASes are there for each business type?

For the purposes of this lab, we will focus on the following three business types: A, B, and C. You will need to remove all records categorized as different business types from the data.

7. How many unique ASes are left in the data?

8. Plot a bar graph of the number of IP prefixes versus type of peer participant for the three remaining labels.

9. Remove column ‘‘`info_traffic`’’ and ‘‘`info_prefixes`, what are your results now? Explain why?

3 Feature selection

A critical part of data preparation and machine learning is feature selection. As you should have observed, our data is full of extraneous detail as it relates to business type classification.

10. Based on your intuition, which features can we safely remove from our data? Explain why. A good place to start is by identifying the feature columns that are free-form text and non-standard (mixed data type) inputs.

11. Again, based on your intuition, which features do you think are valuable (features we want to keep) for a model in classifying an ASes business type? Explain why.

12. “info traffic” is certainly a good feature to use for AS classification. Write a python program to standardize these values for each AS and remove any ASes that do not provide this data.

Before continuing, remove the features you identified in question 10. Write a python program to remove cases with blank entries. Lastly, write a python program that (for all features) maps each unique feature value to an integer. For example, for ‘‘`info_ratio_peer_participants`’’, Balanced = 0, Mostly Inbound = 1, Mostly Outbound = 2.

13. After removing the unimportant features and the cases with blank entries in your remaining features, how many cases are in the data?

Feature selection and dimensionality reduction can be done in a more scientific manner using variance, recursive methods, and decision trees, for example. Check out scikit-learn's feature selection module at (`https://scikit-learn.org/stable/modules/feature_selection.html`).

14. Eliminate all features from the data whose variance is lower than 0.9. Try it with a variance threshold of 0.75 and 0.5. Which features were removed for each variance threshold? Explain your results.

15. Tree-based estimators can also be used to compute feature importance and remove unnecessary features from a dataset. Using a tree-based estimator, rank the importance of the data features from most important to least important. Explain your results.

16. Use a tree-based estimator to remove unnecessary features from the data. Which features were removed. Explain your results.

Take some time to experiment with feature selection and elimination. Create a personalized version of our data with at least four important features.

4 Machine learning

Let's see if how well we can predict an AS type based on the features we have selected. First, you should normalize your data matrix. This means looking at the range of values for a given column and scaling the values of that column such that they lie between [0, 1]. Also, shuffle your data before proceeding.

17. Using scikit-learn, write a program that uses Gaussian Naïve Bayes (GNB) to build a model. Split the data and train on 80% and test on 20%. What accuracy, precision, and recall do you achieve with GNB?

18. Using scikit-learn, write a program that uses a decision tree to build a model. Again, split the data and train on 80% and test on 20%. What accuracy, precision, and recall do you achieve with the decision tree?

19. Using scikit-learn, write a program that uses a random forest to build a model. Again, split the data and train on 80% and test on 20%. What accuracy, precision, and recall do you achieve with the random forest?