Applications of Machine Learning for Networking

Lab 2 (Clustering)

Outlines

- Dataset & Task
- Requirement
- Report Requirement
- Report Hints
- Supplement

Dataset & Task

- Dataset
  - header_type.txt: Explanations of all features
  - header.csv: A list of features (86 features)
  - raw_data.csv: Traffic flows gathered from the real world (4317 instances)
  - cluster.csv: The clusters of instances (There are 4 clusters in the dataset)
- Build a model to cluster the traffic flows
- You can follow the recommended steps in ML-based solution shown in Lab 1.

Requirement

1. Report (.pdf)
   - Explain what you did in this lab (Data preprocessing, Feature engineering, Model learning, etc.)
   - Show, illustrate, and explain your results
2. Source code (.py or .ipynb)

Note: Please zip all your files into a **.zip** extension file and name it with your **student ID** (eg. 0123456.zip). You can discuss the problem with your classmates, but **Plagiarism is forbidden**.

Report Requirement (at least 10 pages)

1. Describe how you performed data processing (10%)
2. Visualize data (10%)
3. Describe what feature engineering skills you used (10%)
4. Use at least 3 different clustering algorithms (10%)
5. Use different parameters for your model (10%)
6. Visualize clusters (10%)
7. Measure performance (10%)
   (Using **metrics.adjusted_mutual_info_score** to measure performance.
   The better the performance, the higher the score.)
8. Try to cluster the traffic flows by using your domain knowledge (15%)
9. Discussions and conclusion (15%)

Report Hints

1. Describe how you performed data processing: **as Lab1**
2. Visualize data: **Ref**
3. Describe what feature engineering skill you used: **as Lab1**
4. Use at least 3 different clustering algorithms: **Ref**
5. Use different parameters for your model: explain why and how to tune
6. Visualize clusters: **Ref**
7. Measure performance
   (Using **metrics.adjusted_mutual_info_score** to measure performance.
   The better the performance, the higher the score.)
8. Try to cluster traffic flows by using your domain knowledge: **tcp or udp, src/dst port ...**
9. Discussions and Conclusion

Supplement

- https://github.com/abhat222/Data-Science--Cheat-Sheet
- https://github.com/Yorko/mlcourse.ai
- https://developers.google.com/machine-learning/crash-course/ml-intro
- https://www.kaggle.com/learn/overview
- https://scikit-learn.org/stable/modules/clustering.html#

Enjoy Lab