巨量資料分析

Homework 1

Cheng-Te Li (李政德)

Dept. of Statistics / Institute of Data Science

National Cheng Kung University

chengte@mail.ncku.edu.tw

HW1: **General Classification Solution**

Deadline: 9/28 (Tue.) 23:59

- Self-check whether you are suitable for this course
- Guidelines for your self learning on machine learning
- Review and practice your Python ML skills
- Get your body warm
- Understand this course is heavy-load
- Road to deal with “Big/Real Data”

- Problem 1: Supervised Learning
- Problem 2: Semi-Supervised Learning
- Problem 3: Summary of insights & difficulties

- Note: You use any packages!

HW1: Datasets

	#Features	#Categorical	#Continuous	Size	#Pos	#Neg	% of Pos
Income	14	8	6	32561	7841	24720	24.0
Arcene	783	0	783	200	88	112	44.0
Bank	16	9	7	45211	5289	39922	11.7
BlastChar	20	17	3	7043	1869	5174	26.5
Shoppers	17	2	15	12330	1908	10442	15.5
Shrutime	11	3	8	10000	2037	7963	20.4

https://www.kaggle.com/lodetomasi1995/income-classification
https://archive.ics.uci.edu/ml/machine-learning-databases/arcene/
https://archive.ics.uci.edu/ml/datasets/bank+marketing
https://www.kaggle.com/blastchar/telco-customer-churn
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
https://www.kaggle.com/shrutimechlearn/churn-modelling

HW1: Models

- **Baseline** models
  - KNN
  - Logistic Regression
  - Random Forest
  - MLP (2-layer NN)
  - SVM
  - XGBoost
  - CatBoost
  - LightGBM

  * NO limitations on the model design (settings and hyperparameters) – all can be chosen by yourself
- **Your Model**: Create your own new model that can consistently outperform all baselines in all datasets

HW1: Settings

- Data splitting
  - Supervised: Train / Validation / Test = **65% / 10% / 25%**
  - Semi-supervised: **50 samples** as labeled data (Training set), the remaining is unlabeled data (Testing set)
- Report the **average** results by **running 5 trials with different seeds of data sampling**
- Evaluation Metric: **AUC**, **Accuracy**
  - 2 tables in total for supervised and semi-supervised learning

	AUC						Accuracy					
	Income	Arcene	Bank	BlastChar	Shopper	Shrutime	Income	Arcene	Bank	BlastChar	Shopper	Shrutime
KNN												
Logistic Reg.												
Random Forest												
MLP												
SVM												
XGBoost												
CatBoost												
LightGBM												
My Model												

HW1: Submission

- This is an **individual** homework
- **Report + Code** submission (in **.zip**) via Moodle
  - Deadline: 9/28 (Tue) 23:59
  - Submit code: **.ipynb** or **.py** (we will check reproducibility)
  - Submit report: **PDF** (no page number limit)
    (* you cannot include code in report)
- Structure of the report
  - **1) Introduction**
  - **2) Methodology**: briefly describe all baseline models, and describe the details of your own model
    - Strategy of data pre-processing
    - Hyperparameter settings of baseline models
    - Your proposed models
  - **3) Experiments**
    - Report the results in 2 tables
    - Report any other methods/effort you have ever tried
    - **Discussion of insights and difficulties (supported by conducting additional experiments)**
  - **4) Conclusions and Thoughts** (心得感想)

HW1: Dealing with Real Data [Discussion]

- How pre-processing affects the performance?
- Which models perform better and why?
- Which dataset is hard to predict and why?
- How does class imbalance affect performance?
- None of models consistently work best? Why?
- Any strategies to improve performance on semi-supervised classification?
- How does your own method work? Can it consistently outperform all baselines across all datasets? Why and why not? When and why does it work better/worse?

HW1: Q&A

- **Q1:** Do I need to have different pre-processing strategies for different datasets?
  - The same strategy of data pre-processing needs to be applied to ALL datasets
- **Q2:** Should I tune hyperparameters of every model to have better results?
  - Not necessary. I depend on you. You can either use default values or do hyperparameter tuning
- **Q3:** What if I cannot create a new method that can beat all baselines over all datasets?
  - Just try your best. Do as many as you can.
  - You can stand on the shoulders of giants!
  - If your method fails, you can justify why it should work and discuss when it succeeds and why it fails