meeting minutes

Statistical Machine Learning – Assignment 1

Group 51: Lang (Ron) Chen

Un Leng Kam

Di Wu

# 29 march

* Brainstorm:
  + Address the data imbalance problem
  + Trying to train NLP large language models
  + Address the problem that data comes from two domains
* Road map:
  + Week 0 (holiday)
    - Ron set up the framework
    - Experiments on traditional ML
    - Finds and reads paper that addresses domain problems.
  + Week 1
    - Di to complete basic machine learning experiments.
    - Un Leng to complete feature-balancing research and experiments.
    - Ron to experiment NLP tools.
  + Week 2
    - Begin compound experiments.
    - Begin to think about report structure.
  + Week 3
    - Report writing.

# 9 april

* Progress update:
  + LSTM and Transformer Encoder work normally; the model is reproducible; weighted binary cross entropy loss is introduced; embedding training on off is introduced.
  + The experiment of training only with words with frequency > 40, where other words are set as UNK, has satisfactory results on LSTM.
  + The gradient reversal layer of DANN should be working as expected.
* Done:
  + Group agreement signed.
  + Formed Kaggle team.
  + Set up a table to record the performance of each experiment.
* To do:
  + Di to perform feature engineering and add burstiness, perplexity, and length to the model.
  + Di to experiment with different oversample techniques, including simple oversample, ADASYN, SMOTE on the full dataset, single domain dataset, on LGBM.
  + Un Leng to finish feature balancing research and experiments, including Word2Vec, Bert (with sentence cohesion), DANN, and Transformer validation.
  + Ron to experiment NLP tools, including Bert, DANN, and transformer validation.

# 16 april

* Progress update:
  + Not to use residual learning for Transformer, poor performance at MLP layer.
  + Use weighted cross entropy as loss.
  + Domain can be confidently predicted with 99% testing accuracy.
* Ideas on existing models:
  + Optimize on a single domain.
  + Ensemble (global or local).
  + Add the engineered features to deep learning.
  + Use 2 classifier heads.
  + Use 2 losses.
* To do:
  + Di to fix the bug on accuracy reporting.
  + Ron to start pretraining for Bert.
  + Ron to enhance the deep learning model for optimization.
  + Un Leng to set up Latex.
  + Un Leng to set up a pipeline for ensemble learning, and model tuning.
  + Report writing job split.

# 23 april

* Progress update:
  + The report is almost done, needs to be cut off a bit to fit into 3 pages.
* Done:
  + Model finalized.
* To do:
  + Everyone proofreads the report.
  + Fit the report into 3 pages.