Description

This assignment asks you to implement the **ensemble learning algorithms** and test your implementation on the given dataset.

The dataset used in our experiment is crawled from Taobao. It contains more than 50,000 reviews with sentiment labels (positive:1, neutral: 0, negative:-1).

Task

Compare different ensemble learning algorithms with different base classifiers. Two ensemble learning algorithms are required (Bagging and AdaBoost.M1); and two base classifiers are required (SVM and Decision Tree). Thus, you should at least compare 4 combinations: Bagging + DTree, Bagging + SVM, AdaBoost.M1 + DTree, AdaBoost.M1 + SVM.

You should design, extract and select the features by yourself.

You are allowed to use existing classifier implementations in the experiment, but you need to implement the ensemble learning algorithms by yourself.

Optional Tasks

- Try other base classifier (such as K-NN, Naive Bayes...).
- Analyze the effect of different (kinds of) features..
- Tune the parameters of ensemble learning algorithms, and analyse their effect on performance.
- Any orther methods you'd like to try.

Submission (learn.tsinghua)

Source code

With necessary comments. No restriction on programming languages, but make sure that TA can run your code easily.

README

A text file that briefly describes how to run your code and produce the reported results. Please also have your name, your student number, your name on Kaggle and your contact information included in it.

Report

A pdf file that includes the following information:

- Your experimental design.
- The experimental results: the results of 4 required combinations.

- You should report the performance of different methods on Kaggle's evaluation set and the rank on the leaderboard.
- Your analysis and discussion: For example: why do the algorithms mentioned above perform differently or similarly on the dataset? What is the difference between Bagging and AdaBoost? Which combination is the best one and why?

Deadline & Other Information

DEADLINE: Thursday May 17 23:59, 2018 (UTC+8)

Upload the packed file (ZIP format is preferred) with your name and student number in filename to learn.tsinghua.edu.cn. Late submissions **WILL NOT BE ACCEPTED**.

Feel free to contact the TA for further information.

shisy13@outlook.com 18505555325

Some Existing SVM and DTree Implementations

SVM

- LibSVM: https://www.csie.ntu.edu.tw/~cjlin/libsvm/
- SVM-light: http://svmlight.joachims.org/

DTree

- C4.5: http://www.rulequest.com/Personal/
- C5.0: http://www.rulequest.com/see5-info.html

Other classifiers

- Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/
- scikit-learn: machine learning in Python http://scikit-learn.org/stable/
- Matlab also has lots of packages for machine learning.

Please note that even if the package provides ensemble learning tools, you SHOULD NOT use them. The implementation of the ensemble learning algorithms (Bagging and AdaBoost.M1) must be done by yourself.

Evaluation

The **root-mean-square error (RMSE)**.

Assume that there are n target samples. Let y denote the labels and p denote the predicted values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n}}$$

Submission Format

For every review in the dataset, submission files should contain two columns: *id* and *label*.

The file should contain a header and have the following format:

```
id,label
1,-0.9
8,0
9,1
10,0.8
etc.
```

File descriptions

- **exp2.train.csv** - the training set
- **exp2.validation_review.csv** - the test set (you should upload your results on reviews in this file)

Data fields

- **id** - an id unique to a given review
- **review** - the text (cut) of a review
- **label** - the sentiment label to a given review (positive:1, neutral: 0, negative:-1)