Data Mining HW 5

LIBSVM

Due date: 23:59, Dec 24, 2018

TA: Chia-Chin Ho, r06921052@ntu.edu.tw

Classification

- In this homework, you have to use LIBSVM to perform classification.
- There are 4 datasets to implement on
  - Iris
  - News
  - Abalone
  - Income

Dataset - Iris

- Predict iris class based on physical measurements.
- The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant. There are 75 instances for training and 75 instances for testing.
- The 4 attributes are sepal length, sepal width, petal length, and petal width. See iris.names for more detailed information.
- One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.
- The provided files are well-organized for LIBSVM.
- Please use iris.tr and iris.te.

Dataset - News

- Predict news category based on its message.
- We take a subset from News20 dataset.
- This dataset contains 3579 instances and their TF-IDF feature (total 23909 dimensions) for 4 different categories. There are 2149 instances for training and 1340 instances for testing.
- Original TF-IDF feature and well-organized files for LIBSVM are both provided. You can observe how LIBSVM handle sparse data format.
- Please use news.tr and news.te in this part.

- Baseline: test accuracy : 84.3%

Dataset - Abalone

- Predict the age of abalone based on its physical measurements.
- This dataset contains 4177 instances, 8 attributes and 3 classes. The Sex attribute is categorical values and the others are continuous values. Take abalone.names for more detailed information.
- There are 3133 instances for training and 1044 instances for testing.
- Raw data is provided. The last column of abalone_train.csv and abalone_test.csv are the class label.
- Please use abalone_train.csv and abalone_test.csv.

- Baseline : test accuracy = 65.1%
- [Note] You can use checkdata.py to check the data format

Dataset - Income

- Predict whether income exceeds $50K per year based on census data.
- This dataset is more complicated than the datasets above. It consists of 48842 instances and 14 attributes, which may be categorical values, or integer values. Moreover, there are missing values. See income.names for more detailed information.
- Raw data is provided. The last column in income_train.csv is the class label. The testing label is not provided.
- Please use income_train.csv and income_test.csv.

- Baseline : 85%
- [Note] You can use checkdata.py to check the data format

5-1 [15%]

Please use Iris dataset in this part. You should answer the following questions in report.

a. Comparison of performance with and without scaling. [5%]
b. Comparison of different kernel functions. [5%]
c. Parameter set and performance of your best model. (Report training accuracy and testing accuracy) [5%]
d. More discussions is welcome. [Bonus 1%]

5-2 [35%]

Please use News dataset in this part. You should answer the following questions in report.

a. Comparison of performance with and without scaling. [5%]
b. Comparison of 5-1-a and 5-2-a. [5%]
c. Comparison of different kernel functions. [5%]
d. Parameter set and performance of your best model. (Report training accuracy and testing accuracy) [5%, Surpass baseline 5%]
e. We know that the curse of dimensionality causes overfitting. How does it influence Naive Bayesian, Decision Tree and SVM separately? [5%]
f. More discussions is welcome. [Bonus 1%]

5-3 [20%]

Please use Abalone dataset in this part. You should answer the following questions in report.

a. Your data preprocessing and scaling range. Please state clearly. [10%]
b. Comparison of different kernel functions. [5%]
c. Parameter set and performance of your best model. (Report training accuracy and testing accuracy) [5%, Surpass baseline 5%]
d. More discussion is welcome. [Bonus 1%]

5-4 [30%]

Please use Income dataset in this part. You should answer the following questions in report.

a. Your data preprocessing / data cleaning. Please state clearly. [10%]
b. How do you choose parameters set and kernel function ? [5%]
c. Report cross validation accuracy, and training accuracy. [5%]
d. Parameter set of your best model. [Surpass baseline 5%, Top 20% in class: 5%]
e. More discussion or observation are welcome. [Bonus 1%]

Submission

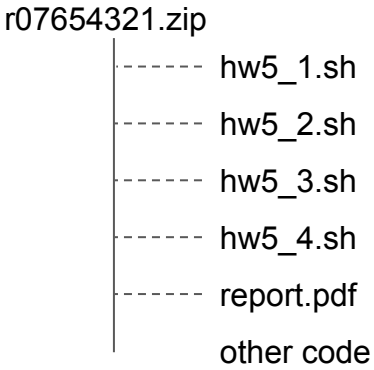
- Use the split datasets offered by TA. We have made some modifications, so DONOT separate training/testing data by yourself.
- You can implement data preprocessing in Python(version>=3.5)
- Allowed packages:
  - numpy, scipy, pandas, scikit-learn 0.20.0+ and Python standard library.
  - If you need to import other packages, please email TA first.
- Please state clearly in your report. The points you will receive based on the completeness of your answers.

Submission

- We will run your code using command:
  - ./hw5_x.sh $1 $2 $3 $4
  - $1: path to svm-scale, svm-train, svm-predict
  - $2: training data
  - $3: testing data
  - $4: prediction output file
  - Eg. for part 5-1 : ./hw5_1.sh ~/libsvm-3.23/ iris.tr irit.te iris.te.prediction
- Output file contains the same number of rows with test file, each row is the predicted class (see sample_output.csv)

Submission

- Report use report template
- Submit a zip file containing your code and report.pdf. Name the zip file to studentID.zip
- The zip file must contain: hw5_1.sh, hw5_2.sh, hw5_3.sh, hw5_4.sh, report.pdf, your code.
- DONOT upload datasets, model files, or LIBSVM related files.



Submission

- No Plagiarism.
- Accept late submission for 2 days after the deadline.
- Late submission penalty is 15 points per day.
- Wrong submitted format will get 10 points penalty.
- It is your responsibility to make sure the submission is completed. Showing an unsubmitted set of homework after the due date will not work.

Reference

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository

Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936)

David Clark, Zoltan Schreter, Anthony Adams "A Quantitative Comparison of Dystal and Backpropagation", submitted to the Australian Conference on Neural Networks (ACNN'96).