HW4

Scikit-Learn

Decision Tree, Naive Bayes

Due date: 23:59, Dec 3, 2018

Classification

- In this homework, you have to use scikit-learn to implement the classification in 2 ways
  - Naive Bayes
  - Decision Tree
- There are 3 datasets to implement on
  - News
  - Mushroom
  - Income

Dataset

- News
  - Subset of 20 Newsgroups dataset
  - 20 Newsgroups dataset is a collection of approximately 20000 newsgroup documents, partitioned across 20 newsgroups
  - We use the subset which includes 4 newsgroups as our data
  - The features are TF-IDF features (total 23909 dimensions)
  - Predict the text belong to which class
  - Testing label is provided
  - The last column of news_train.csv and news_test.csv are the class label

Dataset

- Mushroom
  - Hypothetical samples corresponding to 23 species of gilled mushrooms
  - 22 categorical attributes
  - Classify into edible or poisonous
  - Testing label is provided
  - The last column of mushroom_train.csv and mushroom_test.csv are the class label, (1 for edible, 0 for poisonous)

Dataset

- Income
  - 14 attributes, including categorical and integer
  - Predict whether the income exceeds $50K/yr
  - Testing label is **not** provided
  - The last column of income_train.csv is the class label (1 for income >50K, 0 for income <= 50K)

4-1 [30%]

- Answer the following questions in report.
- Implement Naive Bayes on News dataset
  - What's the parameters and performance of your best model ? (Baseline: Test accuracy 85%) [10%]
  - Compare different distribution assumption, which is the most suitable for News dataset ? [5%]

- Implement Decision Tree on News dataset
  - What's the parameters and performance of your best model ? (Baseline: Test accuracy 61%) [10%]

- How do you choose the parameters to get the best model ? [5%]

4-2 [40%]

- Your data preprocessing for mushroom dataset [5%]
- Implement Naive Bayes on mushroom dataset
  - What's the parameters and performance of your best model ? (Baseline: Test accuracy 98%)[10%]
  - Compare different distribution assumption, which is the most suitable for mushroom dataset ? [5%]

- Implement Decision Tree on mushroom dataset
  - What's the performance of your best model ? (Baseline: Test accuracy 99%)[10%]
  - Use graphviz tool to plot your decision tree [5%]

4-2 [40%]

- Observe the data properties of News and mushroom dataset. According to the model performance, what kind of dataset is more suitable for naive bayes / decision tree ? [5%]

4-3 [30%]

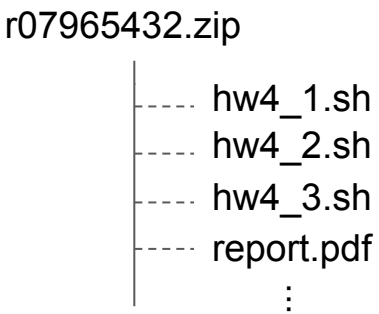
- Implement Naive Bayes and Decision Tree on income dataset
  - How do you preprocess the data ? Missing value ? [10%]
  - Which model gets better performance ? Show the parameters. Surpass the weak baseline (Test accuracy: 80%) [10%]
  - Surpass the strong baseline (Test accuracy: 85%) [10%]

Submission

- python version >= 3.5
- Allowed packages:
  - scikit-learn version: 0.20.0
  - numpy, pandas
- We will run your code using command:
  - ./hw4_1.sh $1 $2 $3 $4
  - ./hw4_2.sh $1 $2 $3 $4
  - ./hw4_3.sh $2 $3 $4
    $1: N or D (N for naive bayes, D for decision tree)
    $2: training data path $3: testing data path
    $4: output file path
    e.g: ./hw4_1.sh N news_train.csv news_test.csv predict.csv

Submission

- Output file contains the same number of rows with test file, each row is the predicted class (see sample_output.csv)
- 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:
  - hw4_1.sh, hw4_2.sh, hw4_3.sh, report.pdf, your code
  - Do not submit the data



Submission

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