Programming HW 2

Web Retrieval and Mining spring 2016

Introduction

- In this homework, you are asked to implement **Naive Bayes Classifier** and **EM Algorithm** with **labeled** and **unlabeled data** on **text categorization**.

- We will give you many articles in the training set, and some documents are with topic labels. Your task is to predict topic for unlabeled documents in the testing set by the **supervised** method and **semi-supervised** method.

Probabilistic model

- Goal : Given documents, we want to predict the topic.
- From the perspective of Probabilistic Model, we want to estimate $\boldsymbol{P}(\boldsymbol{C_k}|\boldsymbol{d_i};\boldsymbol{\theta})$, where d_i is document i and C_k is topic.
- Simple method: **Naive Bayes Classifier**

Naive **Bayes** Classifier

- Recall **Bayes' Theorem**:

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}$$

- $\boldsymbol{P(C_k|x_i;\theta) = \frac{P(x_i|C_k;\theta)P(C_k|\theta)}{P(x_i|\theta)}}$, where x_i is data instance i and C_k is label k.

- The task is to estimate $\boldsymbol{P(x_i|C_k;\theta)}$ and $\boldsymbol{P(C_k|\theta)}$

Naive Bayes Classifier

- Naive Bayes Classifier have **naive independence** assumptions between features.
- $\boldsymbol{P(x_i|C_k;\theta)P(C_k|\theta) = P(C_k|\theta)\prod P(x_{ij}|C_k;\theta)}$
- You can estimate $\boldsymbol{P(C_k|x_i;\theta)}$ from the formula**.**
- For a document, features are **words,** and you can consider the case for text categorization**.**
- Please recall **unigram model** you have learned.

Considering Unlabeled data

- Because labeling is cost, it is important to incorporate unlabeled data for the estimation.
- For unlabeled data, you don’t know the topic, so you cannot directly estimate parameter of classifier from them.
- Solution: **EM algorithm**

EM Algorithm

- Maximize likelihood $\boldsymbol{Lc}(\boldsymbol{D}|\boldsymbol{\theta})$ by E-step & M-step
- E-step : get expectation for current parameter.
- M-step : adjust parameter to fit current prediction.
- By iterative method, you will get better estimation of parameter for the model.

EM Algorithm

- How to derive $\boldsymbol{Lc(D|\theta)}$?

- Hint :

  Consider a document is generated from the mixture of topic, and refer the professor's example in slides.

- Run your algorithm until likelihood converges.

Analysis on EM Algorithm's Result

- You should note that incorporation of unlabeled data may not improve performance even though likelihood is optimized when labeled data is sufficient.

- However, when labeled data is not sufficient for the task, unlabeled data generally helps.

- You need to do experiment on different size of labeled data, and make comparison with Naïve Bayes Classifier. (Ex: 100, 5 or only 1 labeled document in each topic)

20 Newsgroups Dataset

- The data is organized into 20 different newsgroups, each corresponding to a different topic.

- comp.windows.x, rec.motorcycles, sci.electronics……

- We divide it into 3 part, Train, Unlabeled, and Test.
  - **Train** contains labeled document you can use, and the documents are put in the corresponding subdirectory.
  - **Unlabeled** contains documents you can also use, but they are not with labels.
  - **Test** contains documents you need to predict, but we still give the answer for you.

20 Newsgroups Dataset

- Train
  - comp.windows.x
    - 00001
    - 00002
    ……
  - rec.motorcycles
    …….
  - sci.electronics
    ……

20 Newsgroups Dataset

- Unlabeled
  - 00000
  - 00010
  - 00020

  ……
- Test
  - 00000
  - 00010
  - 00020

  ……

Technique of Implementation

- Vocabulary building:
  Only alphabet? Or alphabet and digit?
  Low term frequency?
  High document frequency?
  Stopwords?
  Other condition?

- Smoothing for unseen words and its parameter
  Ex: additive smoothing and its constant

Report

- Please write your report in a **Report.pdf** and put it into the submission zip. The report should contain the following content:
  - Describe your Naïve Bayes Classifier (ex: parameter's meaning in text categorization, and how to estimate them)
  - Describe your EM algorithm (ex: How do you derive likelihood, E-step and M-step)
  - Results of Experiments
    - Result of 2 methods.
    - Analysis on data's size and performance.
  - Some techniques in implementation and their impact.
  - Any other observations in the experiment.

Program IO

- Your program is required to support input of the path of dataset directory containing Training, Unlabeled and Test, and output a result of Test.
- There is no restriction to the programming language you use, but make sure your program is **executable on R217 workstation**.
- Using the third party tools directly for Naïve Bayes Classifier or EM is prohibited.

Result Format

- Each line contains document id and predicted topic.
- First column: **document_id, as filename**.
- Second column: **class_name**, as subdirectory name.
- The two columns should be separated by a space.
- The format is as the same as **ans.test**

Program Execution Details

- You should write scripts according to how you implement this assignment.

- When testing your program, we will execute the following commands **on R217 workstation**, please make sure your program is executable on the workstation.
  - $./compile.sh
  - $./naivebayes.sh -option1 value1 -option2 value2...
  - $./EM.sh -option1 value1 -option2 value2...

Program Execution Details(con't)

Here are the required options that must be supported by your program. If labeled-
data size are not specified, please use all data.

```
SYNOPSIS:
    naivebayes.sh -i data-directory -o outputfile [-n labeled-data size]

OPTIONS:

    -i data-directory
        path of data containing 3 directory, train, unlabeled and test

    -o outputfile
        output file name

    [-n labeled-data size]
        the number of labeled documents in each class the program uses
```

Program Execution Details(con't)

Here are the required options that must be supported by your program. If labeled-
data size are not specified, please use all data.

```
SYNOPSIS:
    EM.sh -i data-directory -o outputfile [-n labeled-data size]

OPTIONS:

    -i data-directory
        path of data containing 3 directory, train, unlabeled and test

    -o outputfile
        output file name

    [-n labeled-data size]
        the number of labeled documents in each class the program uses
```

Submission

- Please put report ,scripts and code into the directory named your **student ID**. Package this folder into a zip file and submit it to CEIBA, following is the structure and content of the zip:
- For example: R04922XXX.zip

+---R04922XXX(directory)

+---**REPORT.pdf**

+---**compile.sh**

+---**naivebayes.sh**

+---**EM.sh**

+---All the other files and source code required by your program

Scoring(15 points)

- 3% for Naïve Bayes Classifier.
- 5% for EM.
- 7% for your report.

Link

- Dataset
  http://www.csie.ntu.edu.tw/~r03922056/20news.tar.gz
- answer of Test
  http://www.csie.ntu.edu.tw/~r03922056/ans.test

Questions?

- Deadline: 2016/06/05 22:00:00
- Late policy: 10% per day

- If you have any question, feel free to ask it on the discussion board at **WebMining@ptt2.cc**.

- Or email to TAs: **irlab.ntu@gmail.com**