STA 250 Final Part2

Chen Zihao 915490404

Problem 2. Extreme classification

In the multi-label classification problem, given the data matrix $X \in \mathbb{R}^{nxd}$ (each row is an input data point) and label matrix $Y \in \mathbb{R}^{nxL}$. L is number of labels, and each row of Y is an L-dimensional 0/1 vector indicating the labels for a data point. We want to predict the label for a given new input data point. In "extreme" multi-label classification, number of labels can be extremely large (e.g. 10,000, or 1 million). Let's develop an algorithm for soving this problem.

We will test our algorithms using the dataset from

http://manikvarma.org/downloads/XC/XMLRepository.html.

We solve the following optimization problem to get the model W,H:

$$\min_{W \in \mathbb{R}^{dxk}, H \in \mathbb{R}^{Lxk}} \frac{1}{2}||Y - XWH^T||_F^2 + \lambda||W||_F^2 + \lambda||H||_F^2$$

Forthis problem we set k=50. After solving the optimization problem, for each testing data $x \in \mathbb{R}^d$, we predict the label $\tilde{y} \in \mathbb{R}^L$ by

$$\tilde{y} = x^T W H^T$$

This is supposed to be close to the true label vector y since we minimize the square loss $||XWH^T - Y||_F^2$ in the objective function.

We will evaluate the results using precision@1 and precision@5, where precision@k is

(number of ture labels in the top-k predictions)/k.

or formally,

$$P@k := \frac{1}{k} \sum_{i \in rank_k(\tilde{y})} y_l,$$

where $rank_k(\tilde{y})$ returns the k largest indices of the predictive label vector $\tilde{y}$, and $y_l = 1$ if it is a correct label.

There will be multiple testing samples, so the P@1,P@5 will be the average among those samples.

1.(10pt) Download the bibtex data. Transform the training data (only samples with training indices) into data matrix X and label matrix Y. Note that in "Bibtex_trSplit.txt" and "Bibtex_tstSplit.txt" there are 10 splits of training and testing data. We will only use the first split (first column in both files) to conduct the experiments.

I have try every mothod i know in R to read the data. But i failed. So i change my strategy.

With the help of windows notepad I replace all the space in the txt file with ";" and then delete all the ":1.000000". Read it in Excel and found it works well, then I save it as the new dataset.

In this method, I got a X and a Y excel.

It is not a elegent way, but i at least work.

```
X1=read.csv("G:\\Bibtex\\X.csv",header = FALSE)
Y1=read.csv("G:\\Bibtex\\Y.csv",header = FALSE)
bibtex_trSplit=read.table("G:\\Bibtex\\bibtex_trSplit.txt")
bibtex_tstSplit=read.table("G:\\Bibtex\\bibtex_tstSplit.txt")
```

```
X=matrix(0,nrow(X1),1836)

for (i in 1:nrow(X1)){
  for (j in X1[i,]){
      X[i,j]=1
  }
}

Y=matrix(0,nrow(Y1),159)
for (i in 1:nrow(Y1)){
  for (j in Y1[i,]){
      Y[i,j]=1
  }
}
```

Choose the training data and test data according to the splits.

```
Xtr=X[bibtex_trSplit[,1],]
Ytr=Y[bibtex_trSplit[,1],]

Xte=X[bibtex_tstSplit[,1],]
Yte=Y[bibtex_tstSplit[,1],]
```

2.(25 pt) Develop an algorithm for solving this problem. Describe your algorithm. What's the time complexity?

Use the coodinate descend to solve this problem.

first fix W to update H, and then Fix H to update W

$$F(W,H) = \frac{1}{2}||Y - XWH^T||_F^2 + \lambda||W||_F^2 + \lambda||H||_F^2$$

first, fix W we get

$$\min_{H \in \mathbb{R}^{Lxk}} \frac{1}{2}||Y - XWH^T||_F^2 + \lambda||H||_F^2$$

$$\nabla_H F(H) = (XWH^T - Y)^T XW + \lambda H$$

similarly, we get

$$\nabla_W F(W) = X^T(XWH^T - Y)H + \lambda W$$

For i=1,2,...,t

$G = \nabla_H F(W^k, H^k)$

$H^{k+1} = H^k - \eta G$

$J = \nabla_W F(W^k, H^{k+1})$

$W^{k+1} = W^k - \eta J$

end