DS 222: ML with Large Datasets

Rahul Chittimalla

1. Local Naive Bayes

Local Naive Bayes is the serial code that is written to classify the DBPedia data. Applying Naive Bayes on DBPedia dataset is a multi-class, multi-label problem which has 50 classes and each document has one or more labels. The accuracies and wall clock times taken for training and testing are given in Table 1 & Table 2

Table 1. Classification accuracies for naive Bayes on DBPedia dataset.

DATA SET	ACCURACY
DEV SET	58.94
TEST SET	60.97

Table 2. Wall clock times for naive Bayes on DBPedia dataset.

DATA SET	TRAINING TIME(SEC)	TESTING TIME(SEC)
DEV SET	29.73	103.194
TEST SET	28.374	65.088

We can observe from Table 2 that the testing times are much higher than training times. This is because training involves just counting the occurrences whereas testing includes calculating the probabilities for classifying among various classes. The number of examples in dev set is 61k and the number of examples in test set is 29k and hence the time is more for testing dev set than test set.

2. Map-Reduce Naive Bayes

In the map-reduce version of naive bayes, we have two stages:

- mapping
- reduction

In the mapping phase of the algorithm we map the words and their counts as key-value pairs where the value is the count of the words. Here is an example:

In Table 3, these are the messages sent by the mapper. Reducer accumulates all the counts and finds the probability distribution for the classification of the multi-class, multi-label classification. The accuracies for the map-reduce version of Naive Bayes are given in Table 4

Table 3. Accumulating counts in a sorted message file.

Y=BUSINESS AND W=ANY, 1
Y=BUSINESS, 1
...
Y=BUSINESS AND W=AAA, 1
...
Y=BUSINESS AND W=ZYNGA, 1
...
Y=SPORTS AND W=HAT, 1
Y=SPORTS AND W=HOCKEY, 1
Y=SPORTS AND W=HOCKEY, 1
Y=SPORTS AND W=HOCKEY, 1
Y=SPORTS AND W=HOE, 1
...
Y=SPORTS, 1
...

Table 4. Classification accuracies for naive Bayes on DBPedia dataset.

DATA SET	ACCURACY
DEV SET	55.6
TEST SET	57.2