COMP90049
Classifying the Geolocation of Tweets
Report

Anonymous

1 Introduction

With the development of microblogging services such as Twitter, the importance of tweets has caught more and more people's attention. As Cheng (2010) said, *"Mining this people-centric sensor data promises new personalized information services".* In this regard, it is valuable to infer the author's location through Twitter data analysis.

This report illustrates how machine learning can use information about a tweet to predict the **geographic location** of the author. Furthermore, in the end of the report, there will have a divergent discussion and **critical analysis** of machine learning in this area.

The data set is derived from the resource published in Eisenstein et al. (2010) and was divided into three data sets: training set, development set, and testing set. Also, the feature engineering will be applied to the raw tweets. In the report, 'count' and 'tfdif' data set will be used.

2 Literature review

The report based on the contents of Cheng et al. (2010), Rahimi et al. (2018) and Eisenstein et al. (2010), these are the literature that discusses user location through network information.

2.1 Cheng et al. (2010)

This paper designs a **probabilistic framework** to judge the user's city location through tweets.

It describes in detail the process from the proposed to the implementation of the probability framework, this report got an important inspiration from the paper.

2.2 Rahimi et al. (2018)

In this paper, a **geolocation model** called GCN is proposed, which is based on a neural network algorithm, and the performance of GCN is evaluate with two baselines in the literature.

2.3 Eisenstein et al. (2010)

In this paper, a multi-level **generation model** is proposed, which recovers topics and regions by the content information of microblogs, and predicts the location of users based on the raw text.

3 Preproccess

3.1 Baseline choose: frequency based

The amount of tweets post in different regions is obviously different, and the regions with more tweets amount will have a higher possibility to become source of a new tweets. The reasons for the different amount of tweets sent in different regions are as follows: firstly, the regional economy will affect the amount of tweets released in the region. In economically developed regions, people have more advanced information equipment, which provides the basis for the sending of tweets; Secondly, the population of the region is also an important factor in the amount of tweets. More people provide more subjects to send tweets. Finally, the amount of tweets sent will also be related to the regional culture. When more people around use Twitter, individuals will be more likely to use Twitter. There are differences in the number of tweets released by different regions and A raw tweet is more likely to come from a region where there are more tweets. Thus, for the baseline, **frequency** is used in this report as the baseline for the predict.

Score on dev = 0.37429193899782137

3.2 Data Process: Transfer to Matrix

Since the format of the given data set is like this

User_id,[(word ID,count)...]

Therefore, it is impossible to use it directly into library functions, so it is necessary to **preprocess** the input data. In this report, the input data is converted to a matrix format. The matrix is generated by numpy. zeros(), the number of rows is the total number of all

words in vocabulary, and the number of columns is the total number of imported tweets. Thus, m[x] [y] can be expressed as the number of times the Y word appears in the vocabulary in the X twitter.

Then, "count" data is iterated through and imported into matrix form for use.

3.3 Model Selection

In the study of IML, various models are discussed in lecture. Here choose the basic model of these model for simple testing.

Model	Score
Naive Bayes	0.454
Decision tree	0.360
KNN	The computation time is too long
Logistic Regression	0.458
Multi-layer Perceptron	The computation time is too long

After the simple test of those models, considering with the score and the computation speed, the final choice of **Logistic Regression** and **Naive Bayes** model for discussion.

4 Logistic Regression

In this section, the logistic regression model will be used to predict the location of tweets posted. Then optimize the performance of the model by trying to compare the result with the given data.

4.1 Attempt of Default Model

At the beginning, try the most basic logistic regression model:

LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)

In this basis, the performance is shown as follow

Data set	Score
train	0.497103778168093
dev	0.45786492374727666

Through observation, we can find that the score of the default model performs on the training set and the development set is about 50%, which is higher than the baseline. This proves that the logistic regression model can predict the location of tweets to a certain extent. But at the same time, the score of 50% is still slightly insufficient. How can we get a higher score?

4.2 Iter Optimization Attempt

When the default model runs, python issues the following warning:

STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

It can be found that the maximum number of cycles is insufficient in the model. Thus, will higher number of cycles lead to better results?

Max_iter	Score on dev set
100	0.45786492374727666
1000	0.45751633986928103

The result is surprising, higher max iter leads to lower score of development set, and the score of training set is almost the same (0.4978 vs. 0.4971). This may be caused by iter **overfit**, more iter improve the accuracy of the training set, but overfit makes the model perform poorly in the development set.Therefore, choosing 100 as the max iter of logical regression in this report would be a better choose.

4.3 Slover Optimization Attempt

As the basis of logistic regression, the solver has a great influence on the result of logistic regression. Will replacing the solver make the model perform better?

Solver	Score on dev set
lbfgs	0.45786492374727666
sag	0.4594335511982571

It can be found that different solvers bring little difference in scores. The better performance of sag solver may be due to the better performance of sag on large data sets. Therefore, logistic regression will try to use sag solver for furtheranalysis.

4.4 Feature Optimization Attempt

Back to the source data, the complexity and variety of the source data is undoubtedly one of the reasons why the function is always difficult to converge. Therefore, it may be possible to improve the efficiency and possibility of model fitting through the processing of source data. So, since there

have the ready-made data set, first compare the score of "**count**" data set in logistic regression with that of "**tfidf**" data set:

Data set	Score on dev set
count	0.4594335511982571
tfidf	0.444880174291939

Obviously, count dataset has better operation efficiency,.Finish processing the source dataset, next the data features are processed.

Combined with the characteristics of language, it can be thought that words often used in language may carry smaller regional features. For example, words such as "I" and "those" will be used in various regions. If these common words are deleted from the data, the efficiency of the model may be improved. Thus, try to remove the top 1 or 5 **commonly** used words and comparing the score.

#of removed words	Score on dev set
0	0.4594335511982571
1	0.45342047930283225
5	0.4523747276688453

In this way, the high-frequency words carry the characteristics of the regional language. Then, try to delete the **lowest frequency** words from the data set, and observe whether there will be better results. Because the words use is less, choose to remove more words for comparison.

#of removed words	Score on dev set
0	0.4594335511982571
10	0.44688453159041397

Therefore, removing words will only reduce the score of the model, probably because it reduces the number of samples to some extent.

4.5 Calibration Attempt

The calibration is carried out by **sigmoid** and **isotonic** methods, and the results are as follows：

Method	Score on dev set
Null	0.4594335511982571
sigmoid	0.464400871459695
isotonic	0.4639651416122004

After **sigmoid** calibration, the model has the best performance.

4.6 Final Score

After modification, the final score of logistic regression was 0.464. The following are the precision_score, recall_score and f1_score on each label:

Score	NORTH EAST	MIDW EST	SOUTH	WEST
precision	0.54131409	0.19607843	0.42332855	0.15147059
recall	0.50640279	0.06738544	0.62189405	0.07202797
f1	0.5232768	0.1003009	0.50375012	0.09763033

5 Naive Bayes

In this section, the Naive Bayes model will be used.

5.1 Attempt of Default Model

There are many kinds of Naive Bayes models, but due to the **discretization** and non **Boolean** characteristics of source data, Multinomial Naive Bayes classifier is used for data prediction in this report.

MultinomialNB(alpha=1.0,fit_prior=True, class_prior=None)

In the beginning, we used the default Bayes model for scoring.

Data set	Score
train	0.49113195560372214
dev	0.4541176470588235

It can be found that the score of naive Bayes is similar to that of logistic regression, so what can be improved here?

5.2 Feature Optimization Attempt

Firstly, try to choose the source data type with comparing the performance of Multinomial Bayes on count and tfidf data sets:

Data set	Score on dev set
count	0.4541176470588235
tfidf	0.46352941176470586

It can be found that the model performs better on tfidf dataset.

5.3 Calibration Attempt

Try to calibrate the Naive Bayes model by sigmoid and isotonic methods:

Method	Score on dev set
Null	0.46352941176470586
sigmoid	0.46309368191721134
isotonic	0.46265795206971677

It can be found that calibration has little or even negative effect on Naive Bayes model.

5.4 Final Score

After modification, the final score of Naive Bayes was 0.464.The following are the precision_score, recall_score and f1_score on each label:

Score	NORTH EAST	MIDW EST	SOUTH	WEST
precision	0.49962756	0.	0.43264163	0.28125
recall	0.62467986	0.	0.61579934	0.00629371
f1	0.55519917	0.	0.50822209	0.0123119

6 Discussion

6.1 Thinking about Results

Through the observation of the results of Logistic Regression and Naive Bayes model, we can find that both models have good performance in the prediction of Northeast and South, but poor performance in the prediction of Midwest and est. This may be caused by two reasons: first, through the **statistics of labels** in y_train, it can be find that Northeast (52582) and South (49901) are the two most frequent labels in y_train, while Midwest (15084) and West (16228) are significantly less frequent than the former two, which may lead to higher proportion of Northeast and South in the process of model learning and have more impact on the model results. Secondly, it may be related to the **language habits** of different regions. The language style of North and South may be more characteristic than that of Midwest and West, which makes the model perform better in judging the location of the two regions.

6.2 More Thinking

As Kerry(2020) said, "*As artificial intelligence evolves, it magnifies the ability to use personal information in ways that can intrude on privacy interests by raising analysis of personal information to new levels of power and speed*"In this era, it seems so easy to get other people's personal information- machine can predict people's current location through a microblog posted by them. On the one hand, as Cheng(2010) said, mining user information can bring more **convenience** to users, but at the same time, the user's **privacy** is also invisible. In the future, when personal information becomes more and more transparent, how to protect personal privacy will become a problem that worth thinking about. In addition, in the study of the model, it is also found that the potential discrimination of the model to the data is obvious. In the prediction of Naive Bayes model, only one of more than 10000 tweets is predicted to come from Midwest (ironically, this predict is wrong), which is obviously more prominent than the prejudice between people. In this context, bias in machine learning is also worth to discussing in the future.

7 Conclusion

This report discusses and optimizes two models(Logistic Regression and Naive Bayes) used to predict the location of tweets posted, and the final scores of models are significantly higher than the 0-R classifier based on frequency. However, in the process of optimization, it can be found that many "thought" model optimization had a negative impact on the performance of the model, which also requires researchers do more testing and consideration when optimizing the model. At the end of the report, report discusses the causes of model performance and think about the potential problems of machine learning.
(2066words)

References

Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P. (2010) A latent variable model for geographic lexical variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), pages 1277–1287. Cambridge, USA.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 759–768.

Rahimi, A., Cohn, T., and Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2009–2019.

Kerry, C. F. (2020, February 10). Protecting privacy in an AI-driven world. Brookings. https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/