UNSW 2020T1 COMP9417 Group Project

Group name: Game of thrones

Group members:

Jianjun Liu(z5226710), Yiting Hu(z5201794),Hangzheng Li(z5140826),Yiyangji(z5216355),Huiyao Zuo(z5196480)

1. Introduction

Project Overview

In this project, we will be processing a dataset containing about 9500 news articles, which contains 10 topics(arts culture entertainment, biographies personalities people, defence, domestic markets, forex markets, health, money markets, science and technology, share listings, sports.) And there are also 48% irrelevant topics.There will be 500 new articles for us to suggest just 10 of them for users. We will be build 6 models to predict the recommend the article given the dataset taken from the articles.Finally, we find the best performance model is SVM.

Aim

The aim of this project is to investigate the performance of different classification / regression models on a dataset and compare their respective features and performance when dealing with different output values.

Models of choice

The models we will be using to classify the dataset are SVM, MultinomialNB, BernoulliNB, DecisionTreeClassifier, Nearestneighbors, Xgboost.

2.Data

Label character

label	number	percent
Arts culture entertainment	120	1.20%
Biographies personalities people	182	1.82%
Defence	271	2.71%
Domestic markets	135	1.35%
Forex markets	893	8.93%
health	197	1.97%
Money markets	1742	17.42%
Science and technology	73	0.73%
Share listings	225	2.25%
sports	1162	11.62%

eature description

2.1 most frequency word

the most frequency word is ‘percent’ , and the frequency is 12114, it is the only most frequency word in the feature.

2.2 least frequency word

there are numerous of word only appearance once, such as ’dolar’, ’takarebank’, ‘hideg’, ‘silverston’, ’bursatil’, ’shool’, ’nether’, ’sinlg’, ’ionport‘, ‘malotan’, ’delarey’etc. There are total 13706 words like this.

2.3 Total number of words

There are totally 36974 words in all the features.

2.4 most frequency topic

Among all the topic, the most frequency of them is irrelevant, the frequency of it is 5000. Except for this topic, the second frequency topic is money markets, which has appeared for 1742 times.

2.5 least frequency topic

The least frequency topic is 'science and technology', which has just showed up for 73 times.

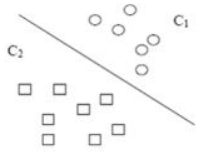
3.Methods

we tried five different methods (SVM , BernoulliNB, MultinomialNB, Decision Tree, NearestNeighbors, Xgboost) with both two feature selections and compare which have a better performance. Here we focus on SVM and MultinomialNB for their better accuracy.

3.1 SVM

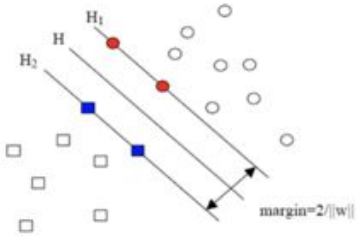
1)method developed

Support vector machines (SVM) is a binary classification model. Its basic model is a linear classifier with the largest interval as the feature. SVM also includes kernel techniques, which makes it a substantially nonlinear classifier. The learning strategy of SVM is to maximize the interval , and is also equivalent to the problem of minimizing the loss function. When the SVM solves problems, it has no limited in dimensions of the sample, even if the sample is tens of thousands of dimensions, which makes the SVM dramatically suitable for solving text classification problems. this is due to the kernel functions in it.



C1 and C2 are the two categories, and their samples in the two-dimensional plane are shown above. The straight line in the middle is a classification function, which can completely separate the two types of samples. In general, if a linear function can

completely separate the samples, the data is said to be linearly separable, otherwise it is called nonlinearly separable.in the graph, infinite lines can choose to separate the two part. the SVM will choose the best one which has the largest margin between the two types of samples.



when the two part is nonlinearly separable, those part can transfer to high dimension and be linearly separable, and this can be done by the kernel functions in SVM.

2)feature selection

all models are use text classification to predict the most relevant news articles for each of the 10 topics. Total use two different sklearn.feature_extraction.text tool , CountVectorizer and TfidfVectorizer.

CountVectorizer is to count the frequency of each word in training text, then form a feature matrix, and each line represents the word frequency statistics of a training text. The idea is that according to all training texts, regardless of the order in which they appear, only count each vocabulary appears in the training text as a feature separately, then forming a vocabulary list. This method is also called the bag of words method.

TfidfVectorizer is a statistical method used to evaluate the importance of a word in a document set. The importance of a word increases proportionally with the number of times it appears in the document, but at the same time decreases inversely with the frequency of its appearance in the corpus. If a word or phrase appears frequently in an article with a high TF and rarely appears in other articles, it is considered that the word or phrase has a good class distinguishing ability and is suitable for classification. TF-IDF is actually: TF * IDF.

Term Frequency (TF) refers to the frequency which a given word appears in the file. It is the number of occurrences of word w in document d count (w, d) divided the total number of words in document d size (d).

Inverse Document Frequency (IDF) is a measure of the general importance of words. The IDF of a specific word can be obtained by dividing the total number of files and the

number of files containing the word, and then taking the logarithm of the obtained number. That is the logarithm of the divided of the total number of documents n and the number of files docs (w, D) appearing in the word w.

By comparing the two results, when use the TfidfVectorizer, the model has better performance.

3)evaluation metrics

all model use same evaluation metrics. In a binary problem, assume that yi = 1 corresponds to positive samples and yi = 0 corresponds to negative samples. Suppose established a classification model H, and a predicted value H (xi) will be output for each input sample xi, then comparing the predicted value H (xi) with the actual value yi, we will get the following four cases : H(xi)=1,yi=1, H(xi)=1,yi=0, H(xi)=0,yi=1, H(xi)=0,yi=0.

In the first case, the prediction is positive, and the output is also positive, so it is true positive (TP), in the second case, the prediction is positive, the output is negative, it is false positive (FP), in the third case, the prediction is negative, the output is positive, called false negative (FN),in the last case, the prediction is negative, the output is also negative, called true negative (TN).

In a test set, can get:

$$N_{pre}=TP+TN$$

$$N_{total}=TP+TN+FP+FN$$

If define a test set where the number of positive samples is P and the number of negative samples is N, then : P = TP + FN, N = TN + FP

Therefore, the acc is actually equal to

$$Acc=TP+TN/TP+TN+FP+FN=TP+TN/P+N$$

The recall is equal to :

$$Recall=TP/TP+FN=TP/P$$

The precision is equal to :

$$Precision=TP/TP+FP$$

F1-score is equal to :

$$F1=2TP/2TP+FN+FP=2 \cdot Precision \cdot Recall/Precision+Recall$$

The formula shows that recall reflects the classification model's ability to recognize positive samples. The higher the recall, the stronger the model's ability to recognize positive samples. Precision reflects the model's ability to distinguish negative samples. F1-score is a combination of the two. The higher the F1-score, the classification model is more robust.

The micro method refers to counting all the classes together. Specifically, to precision, it is to add the TP of all the classes and divide by the sum of the TP and FN of all the classes.

The macro method is to first calculate the precision of each class separately and then count average.

3.2 MultinomialNB

1)method developed

1. conditional probability :

The probability of another event happen when an event happen, such as the probability of event A occurring under the condition of event B:

$$P(A|B)=P(A\cap B)/P(B)$$

multiplication rule of probability :

$$P(A\cap B)=P(A)P(B|A) \text{ or } P(A\cap B)=P(B)P(A|B)$$

2. Bayes's Rule :

If there are k mutually exclusive events, B1, B2 ···, Bk, and P (B1) + P (B2) + ·· + P (Bk) = 1 and an observable event A.then:

$$P(Bi|A)=P(Bi\cap A)/P(A)=P(Bi)P(A|Bi)/P(B1)P(A|B1)+P(B2)P(A|B2)+\cdots+P(Bk)P(A|Bk)$$

Base on above formula.

When only have two class:

If p1 (x, y)> p2 (x, y), then it is classified into class 1

If p1 (x, y)< p2 (x, y), then it is classified into class 2

Combine with Bayes's Rule :

x, y represents the characteristic variable, ci represents the classification, p (ci | x, y) represents the probability of being classified into the class ci under the x, y. Therefore, combining conditional probability and Bayes' theorem,then:

If p (c1 | x, y)> p (c2 | x, y), then the class belongs to c1

If p (c1 | x, y)< p (c2 | x, y), then the class belongs to c2

3. Then in the text classification :

The vocabulary appearance is represented by word vector ω, which is composed of multiple numerical values, and the number of numerical values is the same as the number of vocabulary in the training set.

Therefore, the above Bayesian conditional probability formula can be expressed as:

$$p(ci|\omega)=p(\omega|ci)p(ci)/p(\omega)$$

4. The difference with BernoulliNB :

MultinomialNB is use the number of occurrences as the num of matrix, and BernoulliNB is use 1 and 0 as word used or not as the num of matrix which means ignore the number of occurrences.

5. Apply MultinomialNB in python

```
from sklearn.naive_bayes import MultinomialNB clf = BernoulliNB()
clf = MultinomialNB()
clf.fit(trainX, trainY)
result = clf.predict(testX)
```

2)Hyperparameter tuning

alpha : Floating-point type, optional, default 1.0, add Laplace repair / Lidstone smoothing parameters

fit_prior : Boolean, optional, default True, indicates whether to learn a priori probability,Parameter False means that all class labels have the same prior probability

set fit_prior= False , alpha change from 0.1 to 2 ,the trend of total acc change is increase then decrease. And the peak point is show in alpha=0.6.

set fit_prior=True , alpha change from 0.1 to 2 ,the trend of total acc change is decrease. And the peak point is show in alpha=0.1, and this model perform better then set fit_prior= False.

So the best performance parameter set is alpha = 0.1, fit_prior=True.

4.Results

Following the requirements of application, we will predict the most relevant news articles for each of the 10 topics, when some topics within test set have less than 10 articles, the

application will not want to suggest 10 articles if they are unlikely to be relevant. We would use accuracy score, precision score, recall, F1 macro score and recall macro score to measure effectiveness. Macro score have a better present on small class, when we want to classify some topics that have a small number of articles , this indicator will be more helpful. After implementing those methods on last part, here are result evaluated in four metrics. Every methods would implement with two different word preprocessing methods, TF-IDF and bag of words(BoW). We will processing with vectorized words.(The results of BernoulliNB, Decision Tree, NearestNeighbors, Xgboost are put in Appendix)

MultinomialNB

MultinomialNB	**Models**	**Accuracy**	**Precision**	**Recall**	**F1**
			macro	**macro**	**macro**
TF-IDF	Default	0.680	0.196	0.205	0.197
	Laplace smooth	0.662	0.337	0.367	0.258
	Lidstone smooth	0.708	0.198	0.232	0.213
	Without prior probabilities	0.716	0.381	0.274	0.273
BOW	Default	0.736	0.543	0.573	0.545
	Laplace smooth	0.736	0.543	0.573	0.545
	Lidstone smooth	0.700	0.534	0.684	0.571
	Without prior probabilities	0.730	0.535	0.586	0.548

The table above are consequences of MultinomialNB model predicting topics. we can see from accuracy score that BoW basically have higher accuracy on predicting topics. From the precision score and recall score, the default MultinomialNB model also working on well with BoW method, which means the predicted true positive proportion is higher than it in TF-IDF method.

It can be seen from the chart that, in the BoW method, the accuracy and precision metics shows an decrease trend on smooth statistical results, but have a slightly increase on recall score, this means the predicted true positive cases increased, but the rate of incorrect cases are increased.

Overall, the feature selection process generally not increase the accuracy in BoW method, and TF-IDF method have a lower performance on MultinomialNB model.

SVM

SVM	Models	Accuracy	Precision	Recall	F1
			macro	macro	macro
TF-IDF	Default	0.742	0.473	0.330	0.354
BOW	Default	0.738	0.464	0.324	0.351

From the result form, we can see that the SVM model reaches a highest accuracy by using TF-IDF method, at 0.742. The final selected method is SVM with TF-IDF method. Here are the detailed metric score chart for each topics below:

Topic name	Precision	Recall	F1
ARTS CULTURE ENTERTAINMENT	0	0	0
BIOGRAPHIES	0	0	0

PERSONALITIES PEOPLE			
DEFENCE	1	0.462	0.316
DOMESTIC MARKETS	0	0	0
FOREX MARKETS	0.600	0.125	0.103
HEALTH	0.5	0.214	0.150
MONEY MARKETS	0.600	0.101	0.087
SCIENCE AND TECHNOLOGY	0	0	0
SHARE LISTINGS	1	1	1
SPORTS	1	0.6	0.375

We can see that some minority topics have 0 precision and recall score, this indicate that SVM model cannot find correct article of this kind of topics, that lead to the lower performance on precision and recall score on the final result. Below are the final article recommodation chart:

Topic name	**Suggested articles**	**Precision**	**Recall**	**F1**
ARTS CULTURE ENTERTAINMENT	N/A	0	0	0
BIOGRAPHIES PERSONALITIES PEOPLE	N/A	0	0	0
DEFENCE	9559, 9576,9616, 9670, 9773, 9842	1	0.462	0.316
DOMESTIC MARKETS	9989	0	0	0
FOREX MARKETS	9530, 9551, 9588, 9671, 9682, 9718, 9772, 9798, 9977, 9986	0.600	0.125	0.103

HEALTH	9609, 9661, 9807, 9873, 9929, 9937	0.5	0.214	0.150
MONEY MARKETS	9516, 9534, 9583 , 9602, 9618, 9755, 9761, 9769, 9871, 9998	0.600	0.101	0.087
SCIENCE AND TECHNOLOGY	N/A	0	0	0
SHARE LISTINGS	9518	1	1	1
SPORTS	9568, 9574, 9596, 9752 ,9760, 9774, 9848, 9857, 9886, 9922	1	0.6	0.375

5.discussion

Although we use the same input dataset, we choose different features to train our models in order to generate the best score.

model	Accuracy	F1 score
SVM	0.742	0.354
MultinomialNB	0.68	0.197

we can find the SVM fit the best of all the models.

Findings: We have found that the best models here is SVM, share three common features, which are money markets, forex markets and sports. Meaning the three kind of topics are most popular and should be recommended.

SVM advantages: The optimal hyperplane of feature space division is the goal of SVM, and the idea of maximizing the classification margin is the core of SVM method.

SVM disvantages: SVM algorithm is difficult to implement for large-scale training samples

Improvement：SVM have a lot of disadvantages. For example, it is difficult to choose kernel function parameters for different applications, the classification accuracy of complex

problems is not very high, and the training time for large-scale classification problems is long. And we could improve it through many ways:

(1) It is a meaningful work to improve the learning and training speed, reduce the number of support vectors and simplify support vectors.

(2) How to extend the good binary classification processing ability of SVM to the multi-classification problem effectively is an important aspect of the current research, which is required by the practical application field of SVM expansion.

6 Conclusion

In conclusion, we generate some data which can present the popular article topics which users may interested by using different features. We choose SVM, MultinomialNB ,BernoulliNB, DecisionTreeClassifier, NearestNeighbors and Xgboost models to predict the articles that users may interested. Each of the six models has its advantages, and each model has a suitable features group. Finally, we choose SVM because its accuracy and F1 score are highest.

7 Reference

https://ieeexplore.ieee.org/abstract/document/1007516

https://link.springer.com/chapter/10.1007/978-981-10-5041-1_57

https://dl.acm.org/doi/abs/10.1145/3152494.3152520

https://ieeexplore.ieee.org/abstract/document/97458

https://lear.inrialpes.fr/~douze/enseignement/2014-2015/presentation_papers/muja_flann.pdf

https://dl.acm.org/doi/abs/10.1145/2939672.2939785

https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/neighbors.html#neighbors

https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Appendix

BernoulliNB

BernoulliNB	Models	Accuracy	Precision	Recall	F1
			macro	macro	macro
TF-IDF	Default	0.662	0.337	0.267	0.258
	Laplace smooth	0.662	0.337	0.267	0.258
	Lidstone smooth	0.624	0.420	0.578	0.454
BOW	Default	0.662	0.337	0.267	0.258
	Laplace smooth	0.662	0.337	0.267	0.258
	Lidstone smooth	0.624	0.419	0.578	0.454

From the chart above, we can intuitively see that BoW and TF-IDF methods have no influence on predicting results in BernoulliNB model. Besides, the laplace smooth for data also takes no effort on improving the accuracy score, and the lidstone smooth even get a worser result.

DecisionTreeClassifier

DecisionTree Classifier	Models	Accuracy	Precision	Recall	F1
			macro	macro	macro
TF-IDF	Default	0.680	0.423	0.440	0.423
BOW	Default	0.698	0.500	0.503	0.484
	criterion =’gini’	0.718	0.537	0.524	0.508
	Entropy	0.690	0.459	0.488	0.453
	splitter =’random’	0.674	0.479	0.482	0.464
	max_depth = 20	0.712	0.614	0.465	0.492
	max_samples_split = 4	0.700	0.569	0.458	0.469
	min_samples_leaf = 7	0.722	0.687	0.466	0.504
	max_features = ‘sqrt’	0.590	0.268	0.165	0.170

It can be seen from the chart above that, in BoW method, the default accuracy on DecisionTreeClassifier classify topics are higher. By adjust hyper-argument, the best accuracy score happens at when in BoW vectorize, criterion=’entropy’, splitter=’best , max_depth=20 and min_sample_leaf=7, the accuracy reaches 0.722, but when add an argument of max_features=’sqrt’, the accuracy falls down quickly to 0.590.

when max_features='sqrt', it means that we use the square root of total features to predict, the result is worse than just using all features we get, because all topic message influence the final assign result, especially for the minority articles.

NearestNeighbors & Xgboost

NearestNeighbors	Models	Accuracy	Precision	Recall	F1
			macro	macro	macro
TF-IDF	Default	0.706	0.492	0.461	0.470
BOW	Default	0.656	0.473	0.323	0.328

Xgboost	Models	Accuracy	Precision	Recall	F1
			macro	macro	macro
TF-IDF	Default	0.694	0.484	0.397	0.394
BOW	Default	0.704	0.451	0.416	0.414

NearestNeighbors and Xgboost model have similar performance on this dataset. TF-IDF works on well with NearestNeighbors model on this dataset, the other three metrics also greater than it in BoW method. Xgboost has achieved a higher accuracy by using the BoW, than by using the TF-IDF method.

Accuracy and F1 score of other models

BernoulliNB	0.662	0.258
DecisionTreeClassifier	0.68	0.423
NearestNeighbors	0.706	0.470
Xgboost	0.694	0.394