The Practical Application of Bayes' Law in Predicting English Grades

Zhengyang Liang

Cantwell-Sacred Heart of Mary High School, USA

Keywords: Bayes' law, Grade prediction, Data conversion.

Abstract: Aiming at the requirement of information technology for English learning and the advantage of deep learning algorithm, this paper proposes a prediction model of English score based on Naive Bayes. In order to verify the effectiveness of the model in predicting English scores, this paper first analyzes the principle of Naive Bayes theorem, and then uses the probability of naive Bayes model to predict students' English scores under different attributes based on the CET-4 scores of a vocational college from 2016 to 2018 as the basic data source. The results show that the prediction results of Naive Bayes classification are basically consistent with the practice.

1. Introduction

In order to meet the current information needs of English learning, as well as the advantages of deep learning statistics and its algorithm, this paper designs an English score prediction method based on Naive Bayes. In order to verify the effectiveness of this method in predicting English scores, the principle of Naive Bayes' theorem is firstly analyzed, and then the results of CET-4 in a certain university are taken as the basic data source to predict students' English scores under different attributes using Bayes' law, and then the results are compared with the actual results.

2. Bayes's Theory

For actually using Bayes' law to predict English grades I'm going to use naive Bayesian classification to solve these kinds of problems. Naive Bayes classification is a very simple classification algorithm, call it naive Bayesian classification because of the thought of this method is really simple, simple Bayesian ideological basis is this: to give to the classification, solving each category in the conditions of probability, which is the largest, thought that this classification belongs to which category.

In layman's terms, if you see someone on the street wearing a pair of basketball shoes, and I ask you to guess if that person is playing basketball, you'll probably guess that person is playing basketball. Why is that? Because there's a high percentage of people who wear sneakers to play, and of course that person might just wear them because they look good, but in the absence of other information, we'll choose the category with the highest conditional probability, and that's the basis of Naive Bayes.

The application of Naive Bayes classification assumes that attribute values affect a given class independently of other attribute values. This hypothetical process is also known as class conditional independence. Because the assumed process simplifies the required calculations, the Bayesian classification is labeled "Naive." In essence, Bayesian classification is a classification algorithm based on Bayes' theorem.

Suppose the sample of data with unknown class id is X. Set some assumption to H. For example, the data sample X belongs to a particular class C. For such classification problems, it is necessary to first determine the given observation data sample X, that is, $P(H|X)$, and assume the probability that H can be established. $P(H|X)$ is a posterior probability, or the posterior probability of H under the condition X.

For example, if the components of a data sample are assumed to be fruit, different colors and shapes can be used to describe the data sample. If red and circle are represented by X; Assuming that

X is an apple and H is represented, then *P* (*H*|*X*) represents the determination of the probability that X is an apple when people see that X is red and round.*P* (*H, X*) is the prior probability. In this respect, *P* (*H, X*) represents the probability that any given data sample is an apple, without assuming the color or shape of the data sample.

Compared with the posterior probability *P* (*H*|*X*), the prior probability P (H,X) is independent of X and has less color and shape information than *P* (*H*|*X*). *P* (*H*|*X*) refers to the a posterior probability of X under the condition of H. Simply to say, *P* (*H*|*X*) is refers to under the condition of known X is apple, probability of X in red and the shape is circular.

Bayes' theorem provides a corresponding calculation method for the calculation of a posterior probability. The specific calculation formula is:

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \tag{1}$$

3. Naive Bayesian classification

The specific way of naive Bayes classification is

1. Suppose that the data sample has n attributes (*A*1, *A*2..., *An*), these data samples can be viewed as a point in the n-dimensional space *X* = (*x*1, *x*2..., *xn*).

2. Suppose there are m different categories, *C*1, *C*2... The *Cm*. *X* is a sample of some unknown category. Naive Bayes will classify *X* into *Ci*. When

$$P(Ci|X) > P(CiP|X) \tag{2}$$

That was established for all j (1 jm, ji), to the maximum.

3. It can be obtained through Bayes' theorem:

$$P(Ci|X) = \frac{P(Ci|X)P(Ci)}{P(X)} \tag{3}$$

When all classes are constants, only the largest of the constants needs to be evaluated.

4. After the classification of the unknown sample X, the Ci of each type is analyzed

Make detailed and accurate calculations. Naive Bayes classification has the following characteristics: (1) In the practical application of Bayesian classification, an object will not be assigned to a class absolutely.

Instead, you assign an object by calculating probabilities. (2) Generally, all attributes included in Bayesian classification have potential effects, and all attributes will participate in the subsequent classification work.

(3) There is no absolute requirement for the properties of the categorized object, which can be either offline, continuous, or mixed, depending on the situation.

4. Data model (main source of data and its transformation)

For many college students in China, CET-4 is a necessary test. In this regard, it is an important way to improve the passing rate of CET-4 score to predict the individual's CET-4 score and to intervene the students and English teachers in time according to the predicted results. The training data samples used in this paper are from the CET4 score data of a school from 2016 to 2018, which I found on the Internet. There are 2168 records in total. For data research, high quality data is the basis of accurate and effective prediction of English scores, this paper uses Visualfoxpro 6.0 to data cleaning and data transformation treatment, to improve the quality of data. Since some students did not actually take the exam, there is some data with 0 in the original data. Therefore, you first need to clear the missing parts of the original data from the database tables through data cleansing. In addition, data cleansing involves selecting the fields contained in the database tables. The final fields of this paper include student number, total score, listening score, reading score, writing score and comprehensive score, etc. At the same time, after data cleaning and other pre-processing, 1813 effective performance records were obtained.

After the data is cleaned, the types of the data need to be converted to make it more analytical. Due to the composition of CET4 score: total score 710, listening 249, reading 249, writing 142, comprehensive 70, in order to facilitate research, the CET4 score needs to be converted to 100 system. Then, based on the converted score data of 100 system, the score data other than the total score are classified according to the grades of excellent, pass and fail. The corresponding results of each grade are shown in Table 1. According to the students' comprehensive score of CET4, the scores are counted and classified into three levels. Table 1 is divided into two sections, level and score. Moreover, there are three grades of excellent, pass and fail respectively, and the three grades correspond to scores greater than 85, 60 to 85, and below 60 respectively.

In order to facilitate statistics and records, the classification of the results of the data for a conversion. Convert fail to '1', pass to '2', and excellent to '3', as shown in the Table 2.(Data conversion Table)

Table 1. Performance grade comparison table

Level	Score
excellent	more than 85
pass	between 60 and 85
fail	less than 60

5. Prediction of English grades based on Naive Bayes' theorem

In this paper, the pre-processed CET4 score table data shown in Table 2 is taken as the experimental sample, and the naive Bayesian classification algorithm is used to calculate it, so as to achieve the performance prediction.

Table 2. Data conversion table

Student ID	Listening score	Reading score	writing score	Comprehensive score	Total score
201877701	1	1	1	1	0
201877702	1	1	1	1	0
201877703	1	2	1	1	0
201877704	2	1	2	2	1
201877705	2	1	1	1	0
201877706	1	1	2	1	0
....	....	....	....	....	....

As can be seen from Table 2, the experimental data samples contain four attributes: listening score, reading score, writing score and comprehensive score, and the total score has two values of "0" and "1". Given that the total score of the class =1 is C1, the total score of the class =0 is C2, and the given sample.

$$X = (Listeningscore = 2, \mathrm{Re}\,adingscore = 1, Writingscore = 1, Comprehensivescore = 3) \tag{4}$$

The given sample X must be maximized first to calculate, I = 1, 2. According to the training samples, all kinds of prior probabilities are calculated, and the following results are obtained:

$$(Totalscore = 1) = \frac{92}{1813} = 0.05, P(Totalscore = 0) = \frac{1722}{1813} = 0.95 \tag{5}$$

To facilitate the above calculation, the conditional probability is calculated as follows:

$$P(Listeningscore = 2 | Totalscore = 1) = \frac{62}{91} = 0.67 \quad (6)$$

$$P(Listeningscore = 2 | Totalscore = 0) = \frac{41}{1721} = 0.02 \quad (7)$$

$$P(\mathrm{Re}\,adingscore = 1 | Totalscore = 1) = \frac{17}{91} = 0.18 \quad (8)$$

$$P(\mathrm{Re}\,adingscore = 1 | Totalscore = 0) = \frac{1530}{1721} = 0.89 \quad (9)$$

$$P(Writingscore = 1 | Totalscore = 1) = \frac{50}{91} = 0.16 \quad (10)$$

$$P(Writingscore = 1 | Totalscore = 0) = \frac{1578}{1721} = 0.92 \quad (11)$$

$$P(Comprehensivescore = 1 | Totalscore = 1) = \frac{15}{91} = 0.16 \quad (12)$$

$$P(Comprehensivescore = 1 | Totalscore = 0) = \frac{12}{1721} = 0.008 \quad (13)$$

Thus, according to the above calculation, it can be concluded: P ($Totalscore$ = 1) = P ($Listeningscore$ = 2|$Totalscore$ = 1) × P ($Readingscore$ = 1|$Totalscore$ = 1) × P ($Writingscore$ = 1|$Totalscore$ =1) × P ($Comprehensivescore$ = 1|$Totalscore$ = 1) = 0.01

By analogy, we can get the calculation result of P ($totalScore$ = 1):

$$P(Totalscore = 1) = 0.0002 \quad (14)$$

To get the maximum class, calculate:

$$P(X | Totalscore = 1) \times P(Totalscore = 1) = 0.006 \quad (15)$$

$$P(X | Totalscore = 0) \times P(Totalscore = 0) = 0.00014 \quad (16)$$

The computed results show, P (X|$Totalscore$ = 1) × P ($Totalscore$ = 1) is greater than the result of the P (X|$score$ = 0) × P ($totalscore$ = 0) results, so for the sample X = ($Listeningscore$ = 2, $Readingscore$ = 1, $Writingscore$ = 1, $Comprehensivescore$ = 3), Naive Bayesian classification results to estimate, when the total score is 1, the student can pass CET-4. Comparing the results with the sample data, it is found that the predicted results are basically consistent with the reality.

6. Conclusion

Through the above data research, it can be seen that Naive Bayes' theorem plays a positive role in the classification of CET-4 scores. Through this classification, the scores of CET-4 can be predicted, that is, it can know under what conditions the students' English scores can pass CET-4. However, the above classification is only a simple analysis of English scores. In the future further research, more accurate results should be achieved through in-depth analysis of the effects on English learners in addition to grades. With the continuous development of the current application of information in the teaching system, it has become a trend to strengthen the application of information in learning. As an important subject in the current Chinese university curriculum, English teaching has been widely valued by teachers and students with the help of in- formation technology. In this regard, people have gradually begun to apply statistical algorithms to learning. Through the experiments including the above, it can be concluded that the depth algorithm can be widely used in English English research. In fact, not only is Bayes' law not only applied to predict English scores, but also various algorithms including Bayes' law can solve various problems in people's practical life.

References

[1] Wang Xiaoli, Yuan Junhong. Performance Prediction Model based on Weighted Naive Bayes Classification. Electronic Technology and Software Program, 2013

[2] Jiang Liangxiao. Research on Simple Bayes Classifier and Its Improved Algorithm. China University of Geosciences, 2009.

[3] Li Honglin. Research on improving student Learning Index Prediction using Naive Bayesian Algorithm. Intelligence Exploration, 2015