
Use this template to begin typing the first page (summary page) of your electronic report. This template uses a 12-point Times New Roman font. Submit your paper as an Adobe PDF electronic file (e.g. 1111111.pdf), typed in English, with a readable font of at least 12-point type.

Do not include the name of your school, advisor, or team members on this or any page.

Papers must be within the 25 page limit.

Be sure to change the control number and problem choice above.
You may delete these instructions as you begin to type your report here.

Follow us @COMAPMath on Twitter or COMAPCHINAOFFICIAL on Weibo for the most up to date contest information.

Contents

1	Introduction.....	1
1.1	Background	2
1.2	Restatement of the Problem	2
2	General Assumptions	4
3	Model Preparation.....	4
3.1	Notations	5
3.2	Data Cleaning.....	5
4	Model I: Reported Results Number Prediction Model	4
4.1	Overview	5
4.2	Autoregressive Integrated Moving Average (ARIMA) model	5
4.3	Model Building	5
4.4	Results	5
5	Model II: Relativity Between Word Attributes and Hard Mode Percentage	4
5.1	Overview	5
5.2	Single Letter Frequency	5
5.3	Double Letter Frequency.....	5
5.4	Word Frequency and Letter Frequency	5
5.5	Results	5
6	Model III: Percentage Distribution Prediction Model	4
6.1	Overview	5
6.2	Back Propagation Neural Network Model	5
6.3	Model Building	5
6.4	Results	5
7	Model IV: Word Difficulty Classification Model	4
7.1	K-means Model	5
7.2	Model Building	5
7.3	Results	5
8	Some Interesting Features of the Data Set	4
9	Strengths and Possible Improvements	4
	References	4
	The Letter to the Puzzle Editor	4

1 Introduction

1.1 Background

In 2022, when the novel coronavirus pandemic was rampant, a puzzle game called Wordle, which was run on the platform of Twitter, attracted a lot of attention in a short time and then became popular all over the world. By January 2023, more than 300,000 people were taking part daily, according to the New York Times.

The rules of the game are pretty simple: You try to solve a puzzle by guessing a five-letter word six or fewer times, and you get feedback for each guess: if you have the right letter in the right spot, it shows up green. A correct letter in the wrong spot shows up yellow. A letter that isn't in the word in any spot shows up gray. For this version, each guess must be an actual English word, and guesses that are not recognized as words by the contest are not allowed.

For the New York Times staff overseeing the game, they desperately needed a way to do statistical analysis of the data from past puzzles, and a way to assess the difficulty of the puzzles chosen that day.

1.2 Restatement of the Problem

To optimize The Times' evaluation system, MCM produced a daily results file from Jan. 7 to Dec. 31, 2022. We will complete the following tasks according to the given data:

- Develop a model to account for daily variations in the number of reported results and explain whether word attributes affect the percentage of reported scores.
- Build a model to predict the relevant percentage of one day in the future (1, 2, 3, 4, 5, 6, X), and then evaluate the model by analyzing the performance of the training set.
- Establish a model that can distinguish the difficulty of words and analyze the accuracy of the model.
- Perform descriptive analysis or other statistical analysis on the data set.

2 General Assumptions

3 Model Preparation

3.1 Notations

3.2 Data Cleaning

A. Number

We made a line chart of the data in the Data File to facilitate our search for the wrong data. By looking at the broken line statistics of the number of reported results, we find a cliff drop at 2022/11/30 reported results, as known in Figure 1.

(Figure 1)

Considering the data error is relatively serious, which may cause great interference to our following analysis results, we replaced the data through the method of data fitting before and after. The data used fitting before and after and the replacement results are shown in Table 1.

Table 1: Reported Number Errors in the Data File

Date	Contest Number	Reported Number
...
2022/12/2	532	24646
2022/12/1	531	22628
2022/11/30	530	2569 (25295)
2022/11/29	529	23735
2022/11/28	528	26051
...

B. word

According to the actual rules of Wordle, the target word to be guessed must be composed of 5 English letters. However, in the process of data cleaning, we find through observation that there are letters missing, letters wrong or the number of letters exceeded the limit of 5 in some word data. For example, the word data with contest number#525, #545, and #314 have errors in the word column. By inquiring relevant information published on the official website of the New York Times, we have corrected the wrong word, as shown in the following table.

Table 2: Word Errors in the Data File

Date	Contest Number	Word (Errors)	Word (Corrections)
2022/12/16	545	rprobe	probe
...
2022/12/11	540	naïve	naive
...
2022/11/26	525	clen	clean
...
2022/4/29	314	tash	trash
...

4 Model I: Reported Results Number Prediction Model

4.1 Overview

After data cleaning, we re-generate the line figure for the number of reported results.

Figure

According to the observation and analysis of the new line figure, the data fluctuate greatly in the early period, and gradually tend to be stable and show a slow decline in the later period. Generally speaking, the data of the number of reported results is still in a state of fluctuation over time, which belong to the non-stationary time series. Based on the above analysis, we decide to build the ARIMA model to solve the problem.

4.2 Autoregressive Integrated Moving Average (ARIMA) model

The basic idea of ARIMA model is to treat the data sequence formed by the prediction object over time as a random sequence, which is approximated by a certain mathematical model. Once it is identified, the model can predict future values from past and present values of the time series. The model can be divided into three parts. The bottom layer is the AR model, which uses the same variable such as the previous periods of x , namely x_1 to x_{t-1} , to predict the performance of x_t in the current period and assumes that they are linear relationships. The formula is:

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t$$

Where c is a constant term; ε_t is assumed to be the random error of the mean equal to 0 and the standard deviation equal to σ (zero mean white noise sequence); Sigma is assumed to be invariant for any t .

On this basis, we add MA model to optimize AR model. MA model focuses on the accumulation of error terms in the autoregressive model, which can effectively eliminate random fluctuations in the prediction. Its formula is defined as

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^q \Theta_i \varepsilon_{t-i}$$

However, due to the limitation that AR model requires time series to have stationarity, we still need to continue to optimize it

For a non-stationary time series, after the elimination of its local level or trend, it will show a certain homogeneity, in other words, some parts of the series are very similar to others. This kind of non-stationary time series can be transformed into stationary time series after difference processing, which is called homogeneous non-stationary time series, in which the degree of difference is the order of homogeneity.

If I take ∇ as the difference operator, then we have

$$\nabla^2 y_t = \nabla(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}$$

For the delay operator B , there is

$$y_{t-p} = B^p y_t, \forall p \geq 1$$

Therefore, it can be concluded that

$$\nabla^k = (1 - B)^k$$

Given a homogeneous non-stationary time series y_t of order d , then given that $\nabla^d y_t$ is a stationary time series, it can be called the ARMA (p, q) model, that is

$$\lambda(B)(\nabla^d y_t) = \theta(B)\varepsilon_t$$

In it

$$\lambda(B) = 1 - \sum_{i=1}^p \lambda_i B^i \quad \theta(B) = 1 - \sum_{i=1}^q \theta_i B^i$$

Are respectively autoregressive coefficient polynomial and moving average coefficient polynomial. ε_t is zero mean white noise sequence. The proposed model can be called the autoregressive summation moving average model, denoted as **ARIMA** (p, d, q) .

4.3 Model Building

We group the 359 data sets we had (the data as of November 30 had been cleaned up) and use the data before November 30 as the training set and the rest as the test set. IBM SPSS Statistic 20 are used to establish the ARIMA model, and the following results are obtained:

Table 3: Specific Data of the ARIMA Model

Fitting Statistics	Stationary R^2	R^2	$RMSE$	$MAPE$	MAE	normalized BIC
Mean Value	0.277	0.981	12846.088	5.929	6775.711	19.029

We take 328 pieces of data before November 30 as the training set to test our model, and the test results are shown in the Figure 3.

Figure 3

The images of the real data are generated and compared with the images of the training set, and zoom in at the end of the image where the testing set is made.

Figure 4

The results are shown in the figure above. We can obviously see that the results of the predicted value of the training set coincide with the real data to a high degree, which indicates that our model has high reliability and accuracy.

4.4 Results

Apply the model to the problem we need to solve, and after importing all the data sets, forecast the data after December 31, 2022, until March 1, 2023. Based on our model projections, we conclude that the range of reports for March 1, 2023 is: ()

Figure 5

5 Model II: Relativity Between Word Attributes and Hard Mode Percentage

5.1 Overview

In order to study the relationship between word attributes and the proportion of Hard Mode in report scores, after searching and learning relevant materials, we define word attributes and divided them into 3 criteria, respectively.

- 1) Letter frequency, that is, the proportion of the number of times that each letter in the word appears in the sum of all letters in the word data to the total number of letters. The calculation formula is

$$F_D(\text{letter}) = \frac{\text{the number of times that the letter appears in the word data}}{\text{the total number of letters}}$$

(For example, if the letter "e" appears 15 times in a total of 25 words which all have 5 letters, its letter frequency is $F_D(e) = \frac{15}{25 \times 5} = 0.12$).

- 2) Frequency of combined sounds, similar to letter frequency, that is, the proportion of the number of combined sounds in all word data contained in this word to the total number of words, the calculation formula is

$$F_D(\text{Double Letter}) = \frac{\text{the number of times that the double letters appears in the word data}}{\text{the total number of words}}$$

- 3) Word frequency. The data mainly comes from walfram, a professional word frequency statistics website, which records the usage frequency of each word worldwide. The frequency of the word can be expressed as $F(\text{word})$.

Since there is a large gap between some words, it is not good for comparison, we use sigmoid function for data processing. The functional relation is

$$F_s(x) = \frac{1}{1 + e^{-x}}$$

Its domain of definition is limited to $[-2.5, 2.5]$. The processed values are sorted, and the value of the 180th word in the word frequency sorting of 359 words is processed as 0.5. The image of the specific result is shown in Table 4.

Table 4: Word Frequency and Data Processing

Contest Number	Word	$F(\text{word})$	$F_s(\text{word})$
286	their	0.001799	0.924142
314	other	0.001396	0.923157
...
204	goose	2.56×10^{-6}	0.503492
240	badge	2.44×10^{-6}	0.5
144	cling	2.39×10^{-6}	0.496508
...
84	howdy	3.74E-08	0.076843
107	parer	1.84E-08	0.075858

5.2 Single Letter Frequency

We first count the letter frequency of each letter in the data set and select the nine letters with a higher letter frequency. Then, we divide words in the data set into three categories according to the percentage of Hard Mode, namely high percent, medium percent, and low percent respectively, to study the relationship of individual letter frequency with the percentage of Hard Mode.

As can be seen from Figure 6, the frequency of words fluctuates slightly in the three categories of Hard Mode percentage, but there is no significant overall change and tends to be average. As a result, it can be concluded that the single letter frequency has no significant effect on the percentage of Hard Mode.

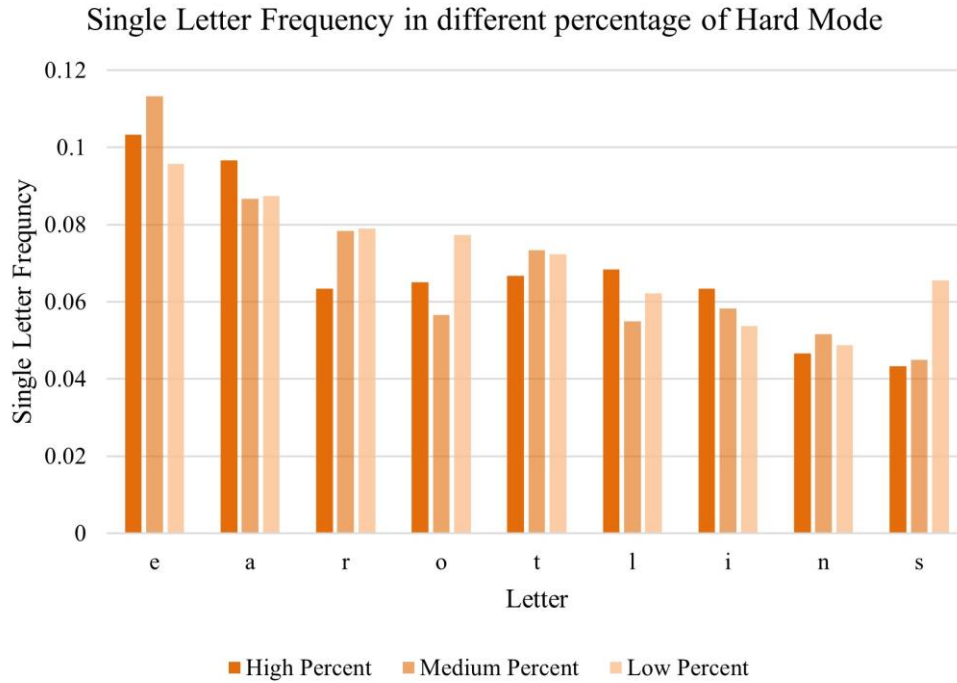


Figure 6: Single Frequency in Different Percentage of Hard Mode

5.3 Double Letter Frequency

After considering the single letter frequency, we also consider the double letter frequency. Similar to the idea when considering the single letter frequency, we first count the frequency of all double letter combinations in the data set and select 6 two letter combinations with higher frequency. Then, we divide words in the data set into three categories according to the percentage of Hard Mode, namely high percent, medium percent, and low percent respectively, to study the relationship of double letter frequency with the percentage of Hard Mode. As can be seen from Figure 7, most of the double letter frequency do not change much, and a few have huge ups and downs, but there is no regularity. As a result, it can be concluded that the double letter frequency has no significant effect on the percentage of Hard Mode.

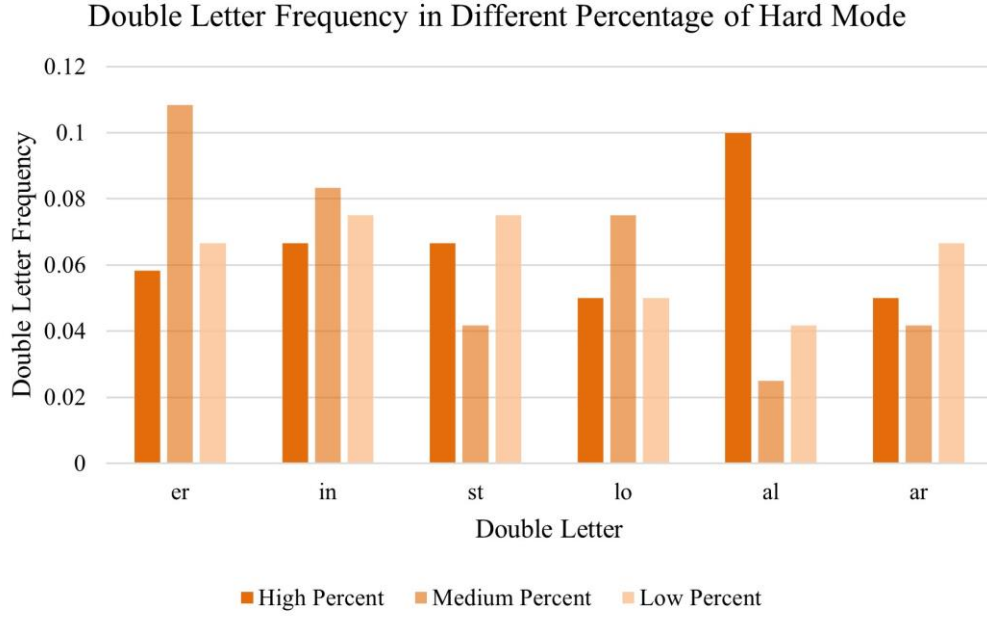


Figure 7: Double Letter Frequency in Different Percentage of Hard Mode

5.4 Word Frequency and Letter Frequency

Furthermore, we draw a 3-dimensional figure of the Hard mode percentage, word frequency and word frequency. Among them, the percentage of hard mode is defined as

$$P_{Hard Mode} = \frac{\text{Number in Hard Mode}}{\text{Number of reported results}}$$

The word frequency $F_S(word)$ is transformed by the sigmoid function.

$$F_S(x) = \frac{1}{1 + e^{-x}}$$

The letter frequency of the word is the sum of letter frequency of its letters.

$$F_D(word) = \sum_{i=1}^5 F_D(letter[i])$$

The X axis is $F_S(word)$, the Y axis is $F_v(word)$, and the Z axis is $P_{Hard Mode}$.

As can be seen from Figure 8, the data points are scattered and have no obvious regularity. As a result, it can be concluded that the word frequency has no significant effect on the percentage of Hard Mode.

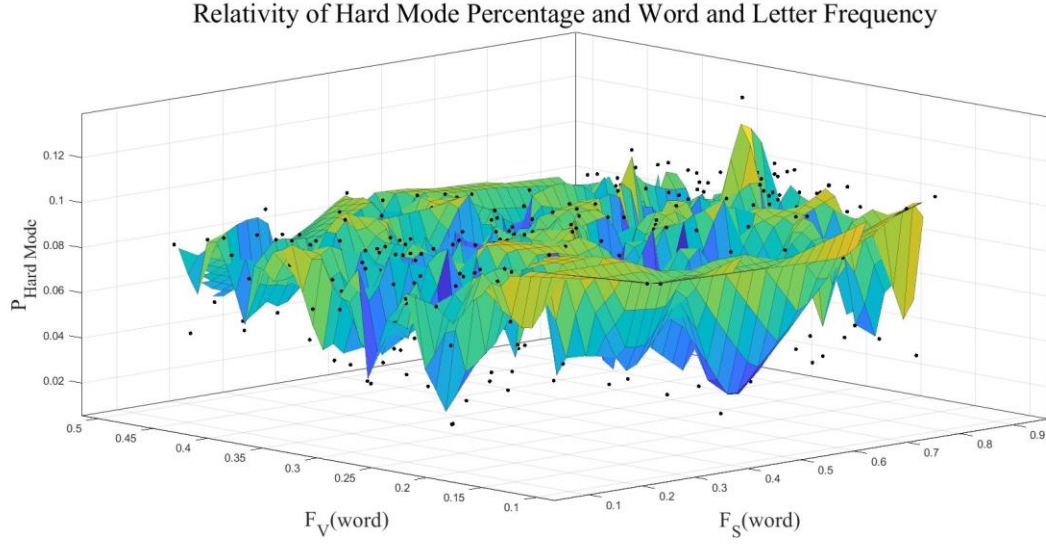


Figure 8: Relativity of Hard Mode Percentage and Word and Letter Frequency

5.5 Results

In this section, we discuss the relativity of Hard Mode percentage and the word attributes from 3 aspects: single letter frequency, double letter frequency and the word frequency. All of them shows that the attributes of the word **do not** affect the percentage of scores reported that were played in Hard Mode.

6 Model III: Percentage Distribution Prediction Model

6.1 Overview

For the distribution of the report of the attempt times of a word, we divide the factors which are affecting it into three types: letter frequency, word frequency and letter repetition times, a total of seven parameters (the letter frequency of each letter should be taken into account), and the target output results are also seven (1,2,3,4,5,6,X). On this basis, we set up BP neural network, optimized by error feedback, and predict the distribution of future report results.

6.2 Back Propagation Neural Network Model

BP neural network is a multi-layer feedforward neural network trained according to the error backward propagation algorithm. The propagation process can be divided into forward propagation process and back propagation process. In the forward propagation process, BP neural network processes data through hidden layer neurons to output corresponding results. In the back propagation process, BP neural network constantly adjusts the parameter values of neurons in each layer by comparing the error between the real result and the predicted result. Thus, the error can be reduced to achieve the ideal effect of the model. The principle of adjustment is that the error is decreasing, so the weight adjustment should be proportional to the error gradient.

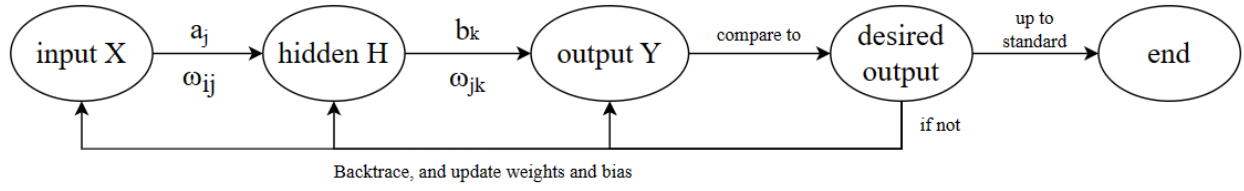


Figure 9: Schematic Diagram of the BP Neural Network Principle

Assume that the number of nodes in the input layer is n , the number of nodes in the hidden layer is l , and the number of nodes in the output layer is m . The weight from input layer to hidden layer is ω_{ij} , the weight from hidden layer to output layer is ω_{jk} , the bias from input layer to hidden layer is a_j , the bias from hidden layer to output layer is b_k , the learning rate is η and the excitation function is $g(x)$. $g(x)$ takes the Sigmoid function, of the form as

$$g(x) = \frac{1}{1 + e^{-x}}$$

As shown in the BP network above, the output of the hidden layer is

$$H_j = g\left(\sum_{i=1}^n \omega_{ij}x_i + a_j\right)$$

The output of the output layer is

$$O_k = \sum_{j=1}^l H_j \omega_{jk} + b_k$$

In view of the error, we still need to set out the equation to take the error

$$E = \frac{1}{2} \sum_{k=1}^m (Y_k - O_k)^2$$

Where Y_k is the expected output. If we call $Y_k - O_k = e_k$, the error can be denoted as

$$E = \frac{1}{2} \sum_{k=1}^m e_k^2$$

In the above formula, $i = 1 \cdots n, j = 1 \cdots l, k = 1 \cdots m$

The updated formula of weight is

$$\begin{cases} \omega_{ij} = \omega_{ij} + \eta H_j (1 - H_j) x_i \sum_{k=1}^m \omega_{jk} e_k & \text{input - hidden} \\ \omega_{jk} = \omega_{jk} + \eta H_j e_k & \text{hidden - output} \end{cases}$$

The updating formula of bias is:

$$\begin{cases} a_j = a_j + \eta H_j(1 - H_j) \sum_{k=1}^m \omega_{jk} e_k & \text{input} - \text{hidden} \\ b_k = b_k + \eta e_k & \text{hidden} - \text{output} \end{cases}$$

Finally, determine whether the iteration is over.

6.3 Model Building

BP neural network system is applied to our study, and its structure is as follows:

- Input layer -- Each input layer of prediction data contains 7 neurons, each representing a parameter describing the attributes of the data (frequency of the first, second, third, fourth, and fifth letters, number of letter repeats, or word frequency).
- Hidden Layer -- This layer has one neuron for each case in the training dataset. The neuron stores the values of the predictor variables for the case along with the target value. In this case, there are 10 neurons in this layer.
- Output layer -- Through the calculation and fitting of the data, the calculated (1,2,3,4,5,6, X) is derived as the prediction result.

A new consideration in this model is the number of letters repeated as a parameter. Through our own experience and searching for relevant strategy data, we've noticed that, unless we can't recall another word or are absolutely certain, when considering possible words, players tend to choose words with fewer repetitions (or even zero) in order to eliminate more misinformation. As a result, it often takes more attempts to find the answer when the letter is repeated.

We select 70% of the data as the training set, 15% as overfitting validation and 15% as the test set, and then obtain the prediction $h(x_i)$ and real values y_i .

To evaluate the accuracy of the model, we can use RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2}$$

In the process of establishing the model, we adjust two parameters: the number of overfitting and the number of hidden layer neurons. We build a 3-dimension figure that reflects the RMSE of our model when the number of overfitting and hidden layer neurons change.

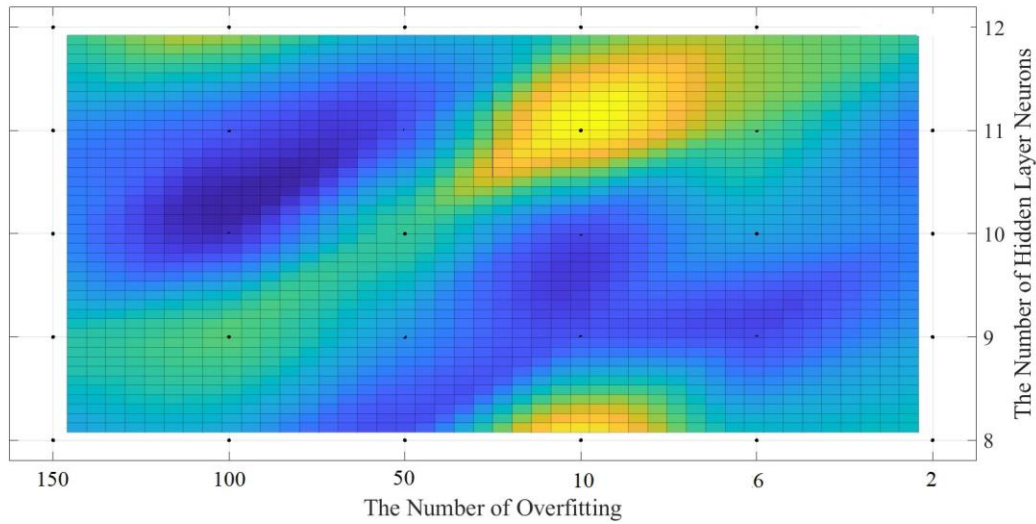
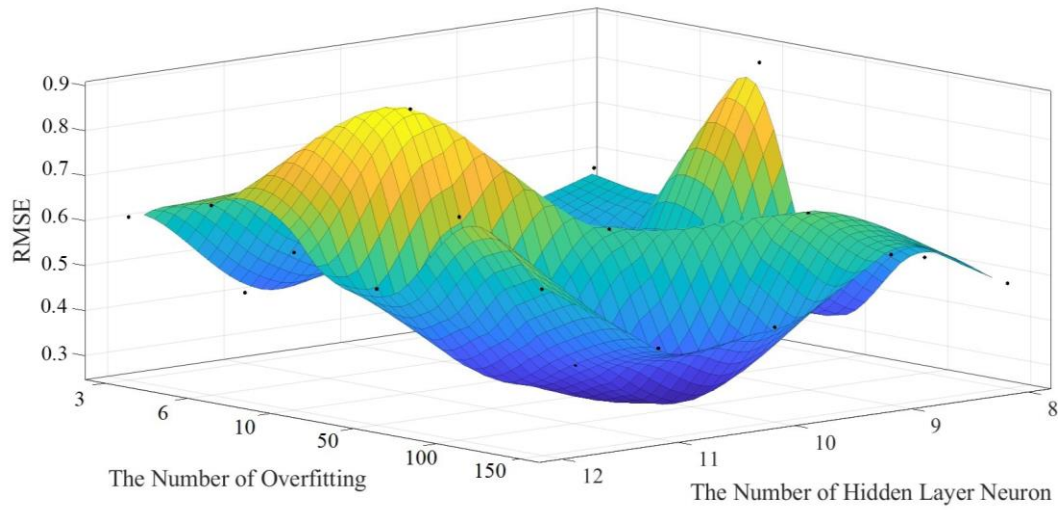


Figure11: The Variation Trend of RMSE with the Number of Overfitting and Hidden Layer Neurons

It can be found that RMSE of the model is lower when the number of overfitting is set to 100 and the number of hidden layer neurons is set to 10, which means the prediction results are better.

6.4 Results

The word that need to be predicted is “EERIE”, and the data of the word can be seen in Table 5.

Table 5: Parameters of the word “EERIE”

Parameter	$F_D(l_1)$	$F_D(l_2)$	$F_D(l_3)$	$F_D(l_4)$	$F_D(l_5)$	$F_s(\text{word})$	Repeat Letter
Value	0.13	0.13	0.06	0.07	0.13	0.283	3

After training the BP Neural Network Model, we get the output below, which is the prediction of the distribution of the reported results.

Table 6: Prediction of “EERIE”

Tries	1 try	2 tries	3 tries	4 tries	5tries	6tries	7 or more tries (X)
Percentage	0	4	20	30	23	17	6

7 Model IV: Words Difficulty Classification Model

7.1 K-means Model

K-means clustering algorithm is a clustering analysis algorithm with iterative solution. It divides the sample set into k subsets, constituting k classes, and divides n samples into k classes. The distance between the center of each sample and its class is minimum, and each sample belongs to only one class (which is the characteristic of hard clustering algorithm).

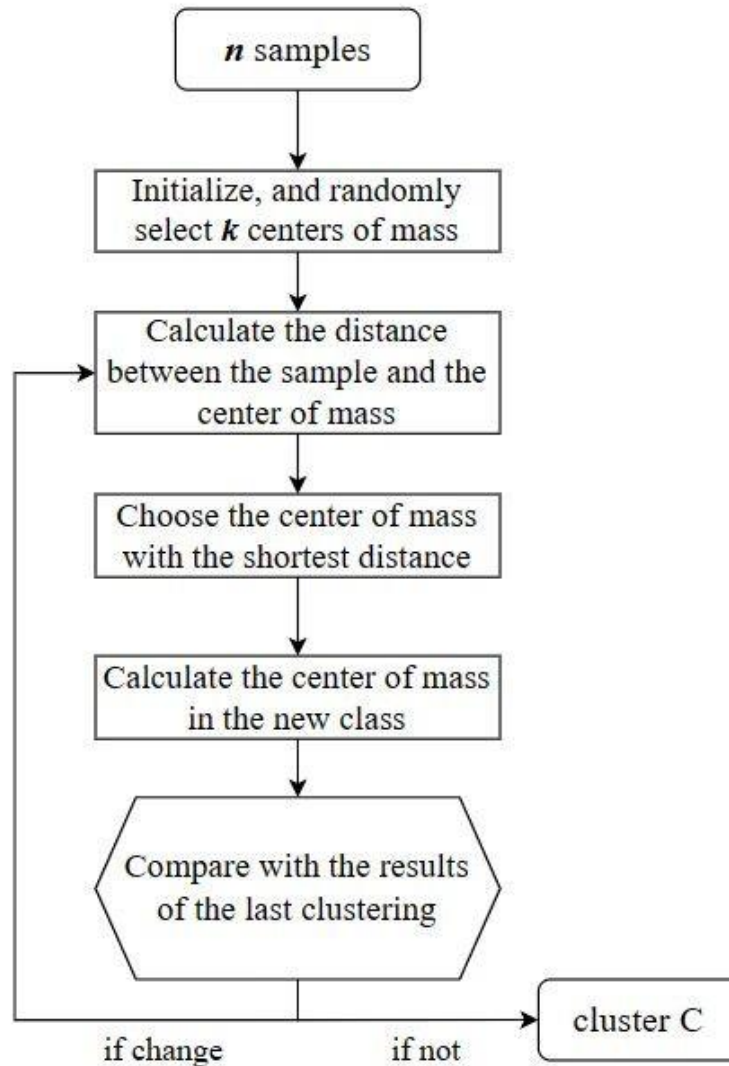
Its specific construction process is as follows:

Step 1: Initialization. The sample of k is randomly selected as the center of the initial clustering.

Step 2: Cluster the samples. For the clustering center selected during initialization, the distance between all samples and each center is calculated, the default Euclidean distance, and each sample is aggregated into the class of its nearest center to form the clustering result.

Step 3: Calculate the class center after clustering, and calculate the centroid of each class, namely the mean value of samples in each class, as the new class center.

Step 4: Then perform steps (2) and (3) again until the clustering results no longer change.



7.2 Model Building

The data we use for the K-means Model has a total of 7 parameters, as the percentage distribution. We use SPSS to build the K-means Model and get the classification of each word. After obtaining the classification of word difficulty, to facilitate visualization, we simplify the 7 parameters to 2 parameters by linear combination. Next, it is represented in the rectangular coordinate system, and the data points of the same classification are given the same color to distinguish them from the other classification.

Figure

The figure shows the classification of word by difficulty directly.

8 Features of the Data Set

- Since January 10, the number of reported results has shown an overall trend of crazy growth, tripling in just 23 days.
- While the number of reported results remained stable at around 20000 for several days before and after, the number of reported results dropped sharply to 15,554 on 2022/12/25. Combined with real life, we speculated that it might be because of the influence of Christmas.

- The proportion of hard mode report results in the total number of report results continues to increase, and finally, as the total number of report tends to stabilize, it is roughly stable around 10%. Combined with our statistical analysis of the total reported result data, we conclude that the likely reason for this change is that the number of "elite" players who choose hard mode is relatively stable compared to the number of regular players who churn out. And as some interested players become more familiar with the game, they are willing to play it in hard mode.

9 Strengths and Possible Improvements

References

Dear Puzzle Editor of the New York Times:

Thanks for taking the time out of your busy schedule to read my letter. Hope our advice can help.
MCM Team #2313119