论文写作

Q3

Develop and summarize a model to classify solution words by difficulty. Identify the attributes of a given word that are associated with each classification. Using your model, how difficult is the word EERIE? Discuss the accuracy of your classification model.

开发并总结了一个模型，按难度对解词进行分类。识别与每个分类相关联的给定单词的属性。用你的模型，EERIE这个词有多难?讨论你的分类模型的准确性。

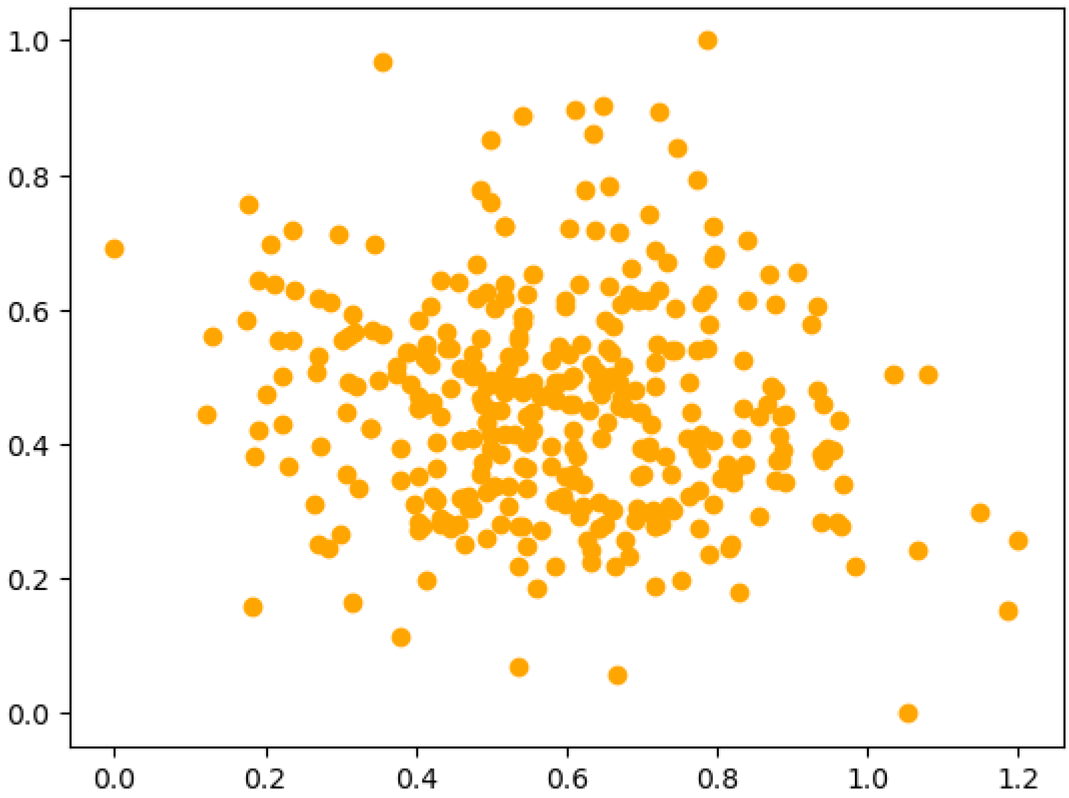
问题分析

There are four tasks we need to accomplish:

1. Classify the samples by difficulty
2. Identify the attributes of a given word associated with each classification
3. How hard is the word EERIE?
4. Discuss the accuracy of your classification model

This is a scatter plot of mean and standard deviation

主要指标的散点图



第三位要求我们分类，但是数据集中并没有事先给定的难度等级。

所以第一步，我们使用聚类分析，对均值与标准差按距离进行无监督的Kmeans聚类，并拟合出概率分布直方图去直观的进行难度分类

第二步，我们使用随机森林分类模型，对单词属性的五个指标（列出），对第一步聚类所形成的target进行映射，拟合出一个分类器，方便后续对单词的预测。

第三步，我们将EERIE的单词属性做评估，再放入训练好的随机森林模型中做预测，给EERIE这个单词做难度定性。

Accordingly, the model construction and analysis of the third question can be divided into four steps：

1. Using percentage and included attributes to do cluster analysis

   In this section，we use K-means clustering analysis to cluster the mean value and standard deviation according to the distance, and fit the probability distribution histogram for difficulty classification

2. Using word attributes to map classification by classification model

   We use the random forest classification model to map the five indicators of word attributes (EN, TR, GS, YS, CS) and the target formed by the clustering in the first step, and fit a classifier to facilitate the subsequent word prediction.

3. Put EERIE into a classifier for sorting

   We evaluated the word attributes of EERIE, then put them into the trained random forest model for prediction, and determined the difficulty of the word EERIE.

4. Get a representation of the model's performance on the test set

模型原理

K-means clustering is a common unsupervised learning algorithm used to divide a data set into multiple different groups (clusters). The similarity of data points within each group is high, while the similarity between different groups is low. The optimization goal of K-means algorithm is to minimize the Sum of distance (SSE, Sum of Squared Errors) between each data point and the clustering center of the cluster it belongs to. Therefore, the main steps of k-means clustering include selecting the appropriate K value, initializing the clustering center, allocating data points to the cluster, calculating the center point of the cluster, and iterating until convergence is reached.

The principle of random forest classification is the same as that of random forest regression used in the second question, so I will not repeat it here.

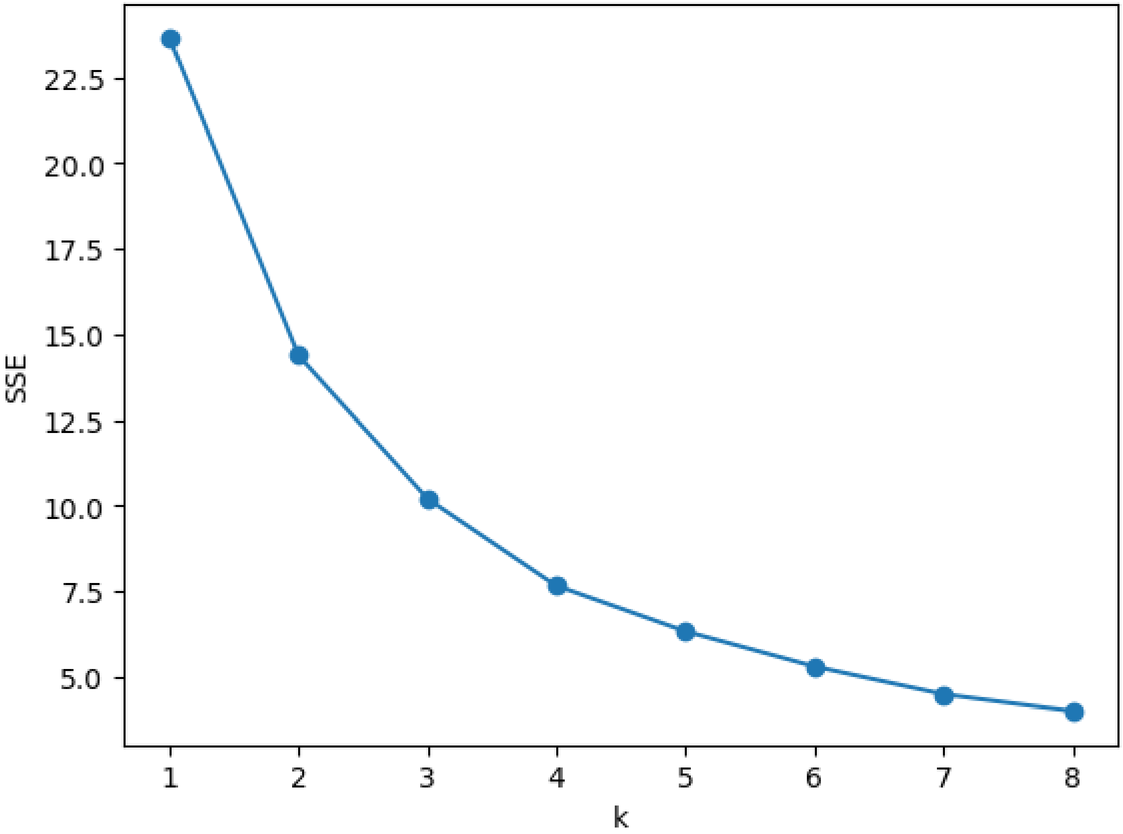
模型假设

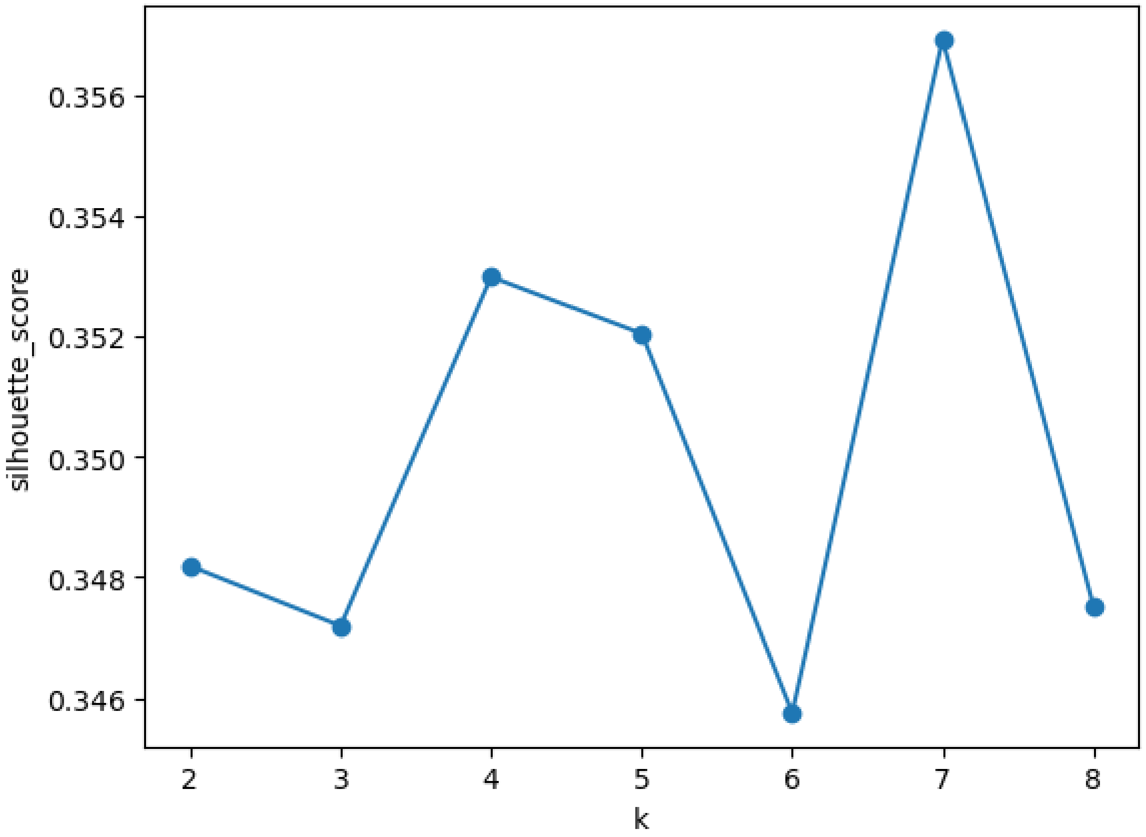
1. Suppose that the difficulty of the word is described only by the percentage distribution of the number of reports, independent of other factors such as the percentage of difficult patterns chosen.
2. Suppose that the percentage distribution of the number of reports is close to a normal distribution, which can be well described by means and standard deviations alone.
3. Suppose that the mean of the distribution of reporting times can better reflect the difficulty of words than the standard deviation. In K-means analysis, the weight ratio of mean to standard deviation is 6:5.
4. Suppose that noise caused by chance will not obscure the characteristics of standard deviation and mean

主要模型

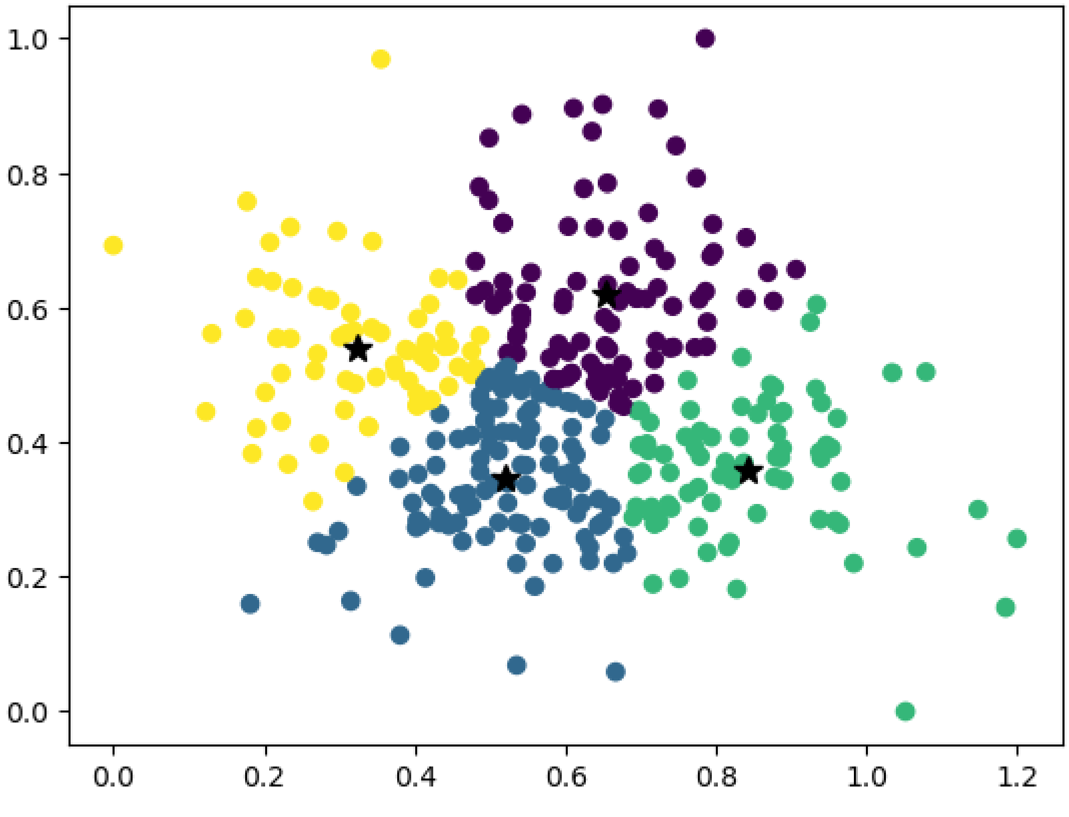
The first step is to construct the clustering model with K-means clustering

1. The data is normalized and the mean is multiplied by 1.2, making the weight ratio of mean to standard deviation 6:5.
2. Using elbow method and Silhouette Coefficient, the optimal clustering number was determined to be 4（左右各一张图）



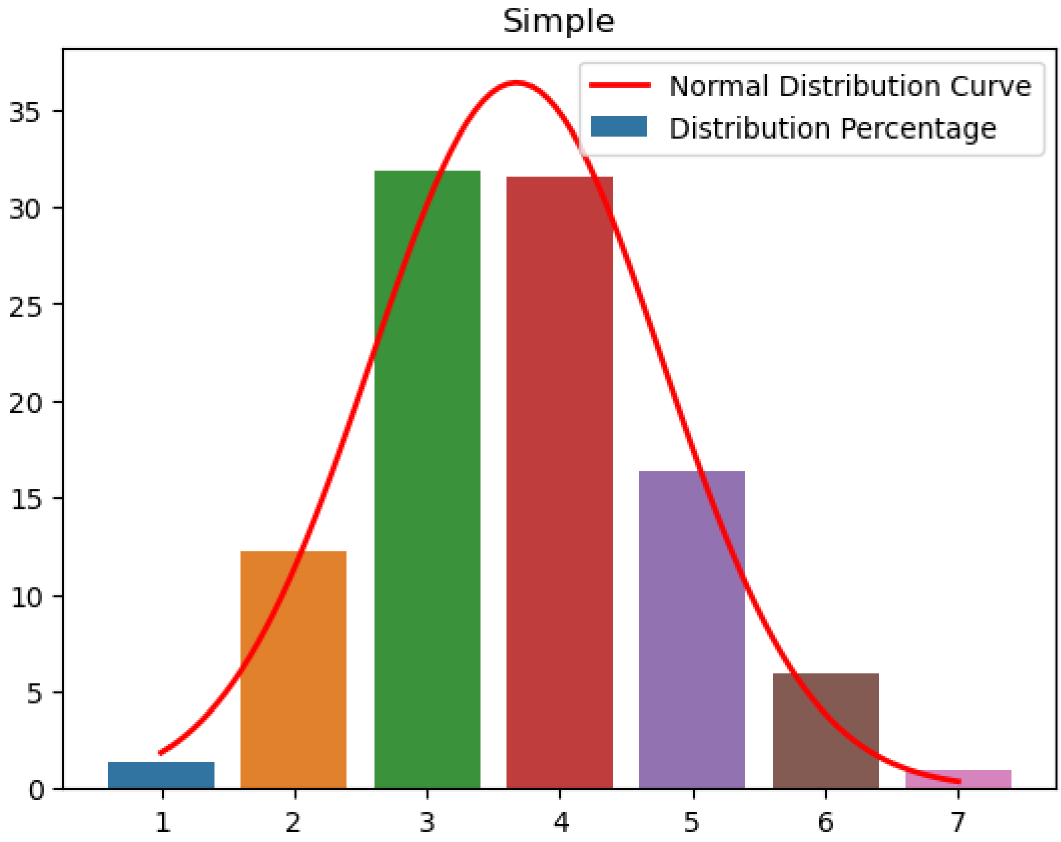


3. Here is the clustering result (the asterisk is the clustering center)

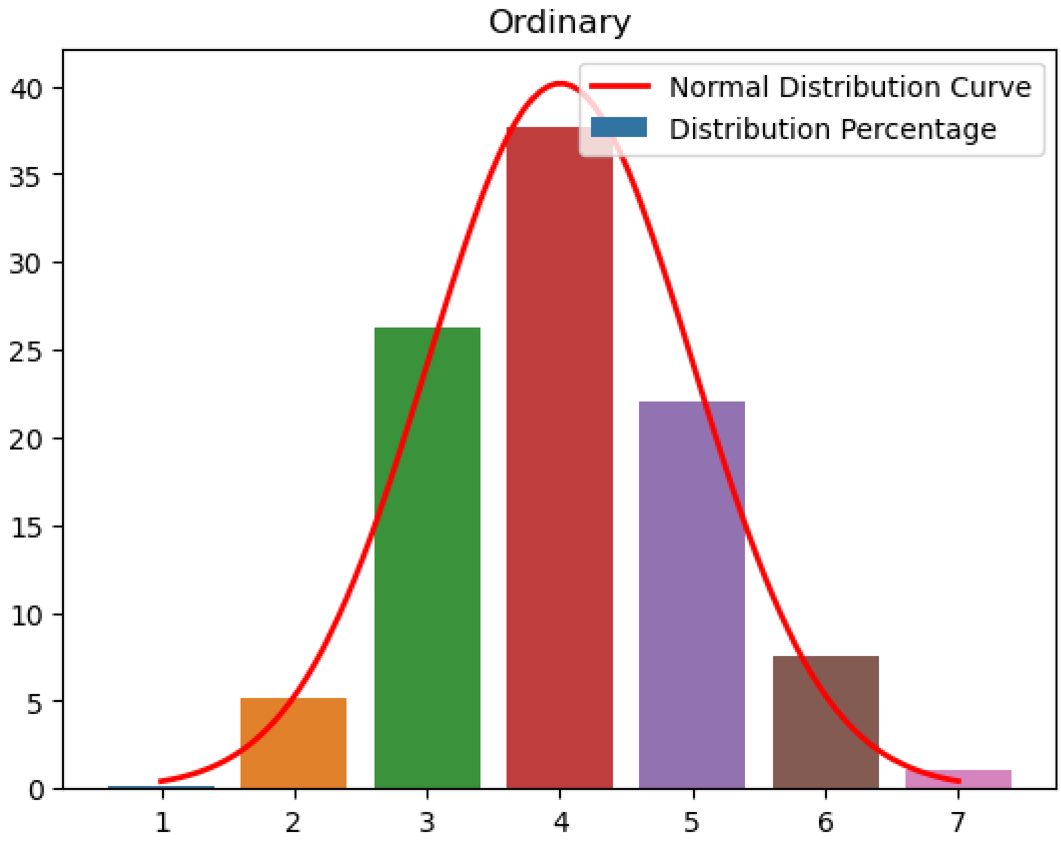


4. The distribution histogram and normal distribution diagram of different difficulty are summarized

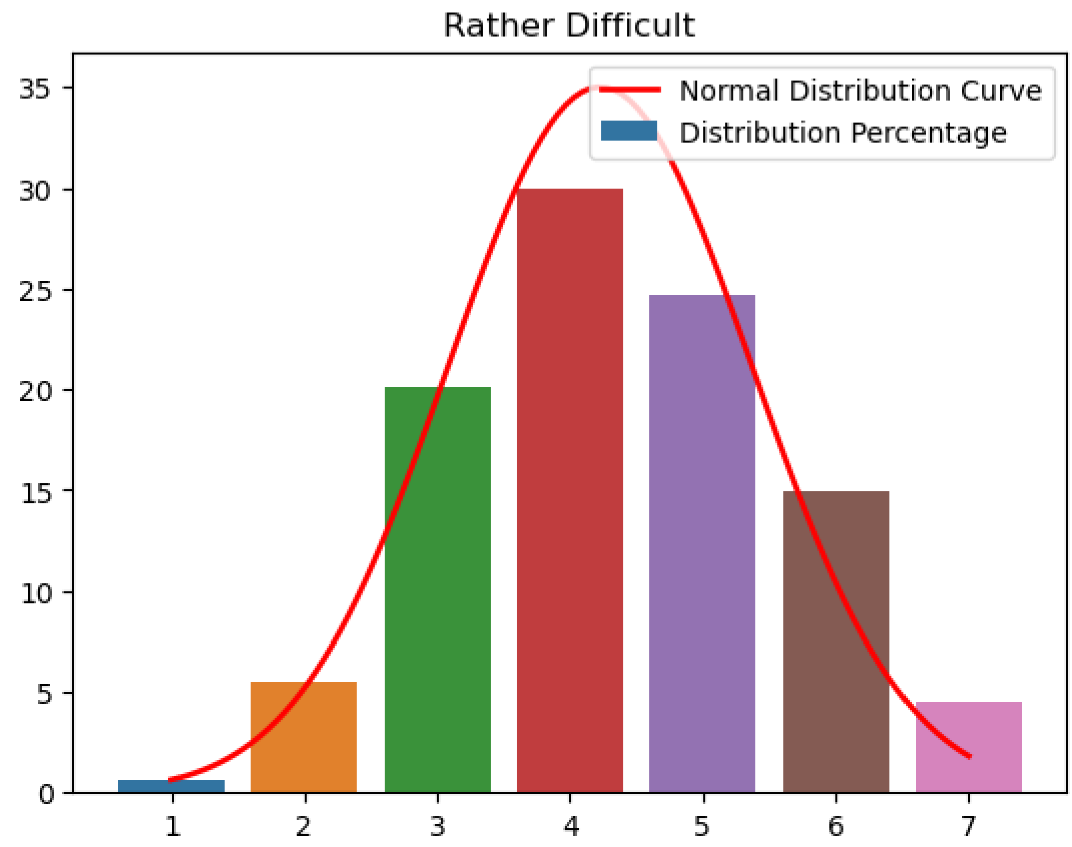
- Simple



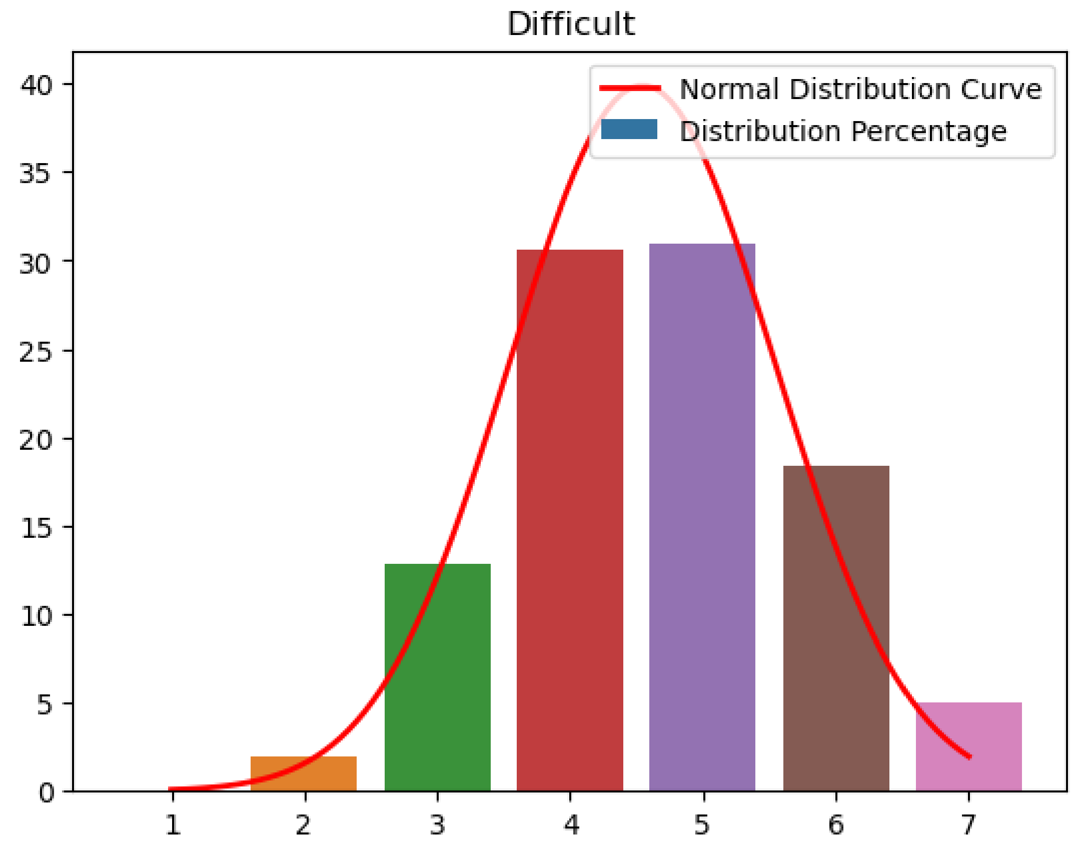
- Ordinary



- Rather Difficult



- Difficult

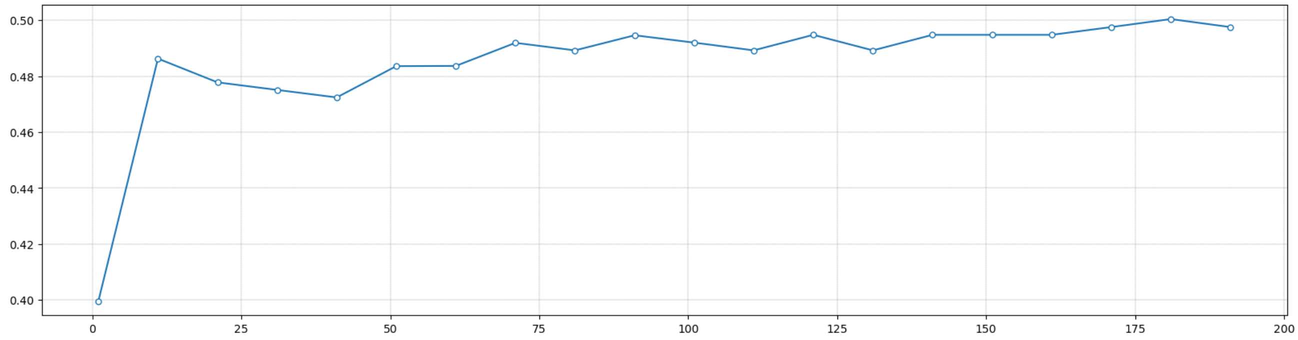


Second, the random forest classification model is no longer used to classify word features

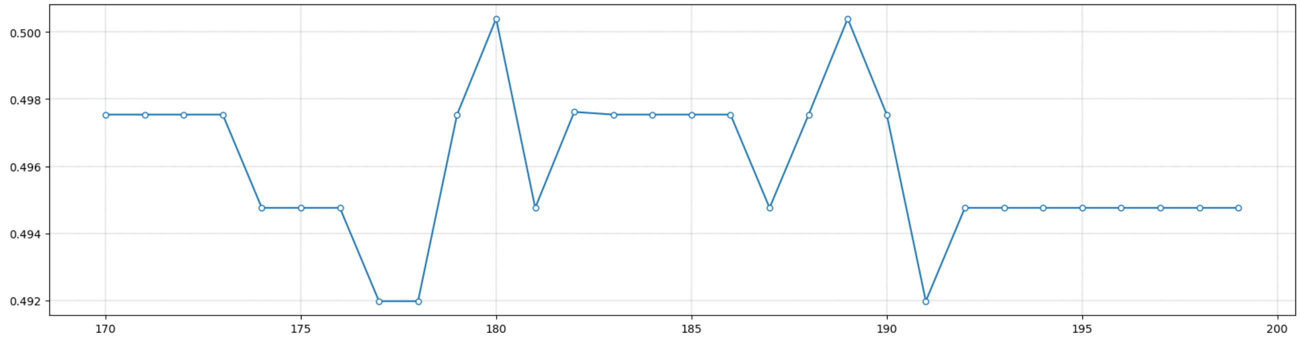
1. Adjustment parameter

   The optimal parameters of random forest are determined by grid search and learning curve

large-scale learning curve



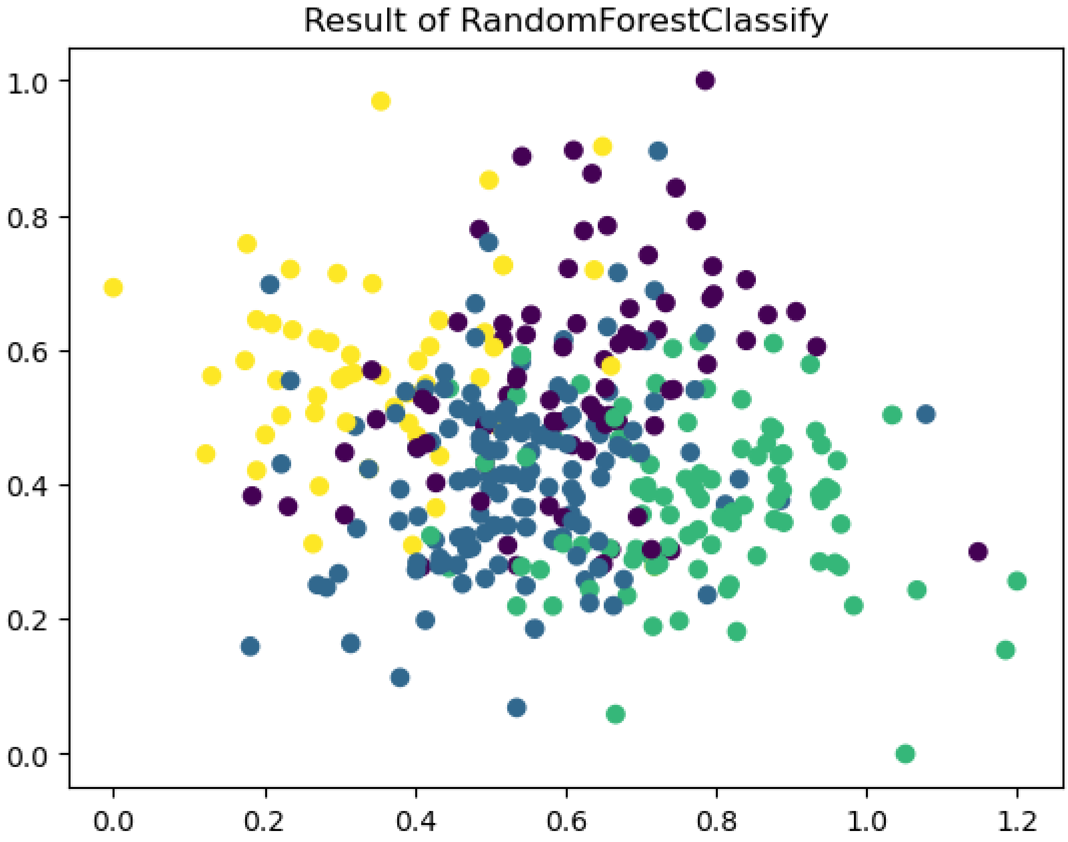
Small region learning curve



parameter	Value
`n_estimators`	189
`max_depth`	10
`min_samples_leaf`	6
`min_samples_split`	None

2. Training random forest classification models

   Classified image



Accuracy for the training set： 0.736

Accuracy for the testing set： 0.523

3. The importance of the feature

Feature	Degree of importance (%)
entropy	37.54
text_rate	15.58
green_score	21.12
yellow_score	14.30
conbine_score	11.44

4. Predict word difficulty （EERIE）

Predicted classification：（Difficult）

Classification probability

classification	probability

classification	probability
Simple	9.54
Ordinary	14.09
Rather difficult	13.63
difficult	62.73