Lab Report For INT104 Coursework 2

Xu Chen
2257453
TA: Dr Shengchen Li

***Abstract*—Supervised learning is an important part of traditional machine learning. This report is based on a data set of INT104 student final exam results from the previous year, and compares decision trees, random forests, support vector machines, naive Bayes and ensemble learning for data classification. In addition, we map these models to two dimensions and use tsne dimensionality reduction technology to visualize the hyperplane of the model, so as to better understand how to select models and extract features under classification problems, and explain the causes of phenomena based on theory and experiment.**

I. Introduction

In this report, our goal is to use traditional supervised learning to complete four classification tasks based on index, gender, grade, total, MCQ, Q1, Q2, Q3, Q4 and Q5, with a sample size of 619, to distinguish which students belong to the programme .

In this article, all methods below will perform data normalization operations in advance

$$x' = \frac{x - \text{mean}}{\text{std}} \tag{1}$$

All the following machine learning methods will choose the cross-validation method. First, the training set and the test set are divided according to 4:1, and then the 5-fold cross-validation method is used for the training set, and the average is obtained to obtain the test set accuracy.

#	Feature	Average Mutual Information
1	Grade	0.198751
2	Total	0.187396
3	MCQ	0.104358
5	Q2	0.087815
7	Q4	0.086718
8	Q5	0.058307
6	Q3	0.048980
4	Q1	0.040192
0	Gender	0.032749

TABLE I
The Importance of Feature Based on Average Mutual Information

Table 1 shows me using the average of 10 groups of random numbers to calculate the mutual information of each feature and item. The higher the ranking, the higher the importance. This is an important basis for how I will extract different features to calculate the accuracy. I will directly select the top Four as initial features, and then incremented one by one.

Figure 1 mainly explains why we later use tsne dimensionality reduction technology as observation instead of other dimensionality reduction techniques. It can be seen from

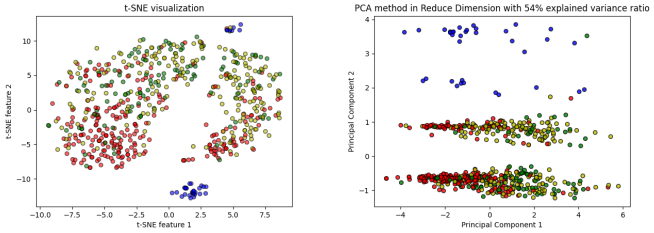

Fig. 1. Comparison of PCA and TSNE in two dimensions

PCA that the data points are very close together, which is not conducive to hyperplane drawing.(Related parameters in tsnecomponents=2, random state=60, perplexity=100, learning rate=50, iteration =1000, early exaggeration =70)

II. Decision Tree and Random Forests

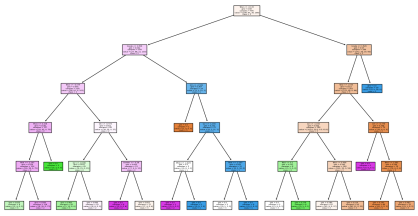

Fig. 2. Initial decision tree

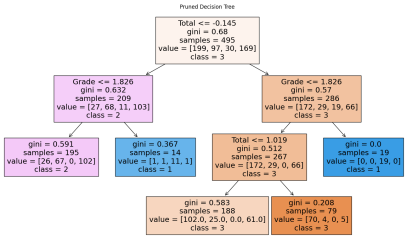

Fig. 3. Post-pruning decision tree

Figures 2 and 3 show the decision tree for all features (see Table 2 below for the reason: the test set has the highest accuracy without deleting any features), excluding the initial state under the index and the actual classification basis after

post-pruning processing.(Related parameters : max depth=5, min samples split=2)

TABLE II
CLASSIFICATION OF DECISION TREES UNDER DIFFERENT FEATURES

Feature	Cross-validation accuracy	Test accuracy
Grade, Total, MCQ, Q2	0.569	0.580
Grade, Total, MCQ, Q2 ,Q4	0.573	0.588
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.575	0.596
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.567	0.604
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.553	0.604
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.561	0.604
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender (Post-puring)	0.553	0.653

TABLE III
CLASSIFICATION OF RANDOM FORESTS UNDER DIFFERENT FEATURES

Feature	Cross-validation accuracy	Test accuracy
Grade, Total, MCQ, Q2	0.581	0.629
Grade, Total, MCQ, Q2 ,Q4	0.587	0.645
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.591	0.620
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.577	0.661
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.591	0.661
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.591	0.653

From Table 2 and Table 3, it can be seen that decision tree and random forest have the highest test set accuracy without taking features.Figure 4 shows that both algorithms classify items 1 and 4 most accurately. Figure 5 shows that the advantage of random forest over decision tree is better generalization ability. This is because nonlinear hyperplane is much better at classifying data points in discrete space.

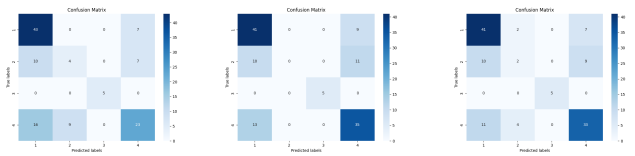

Fig. 4. Confusion matrix for decision trees and random forests

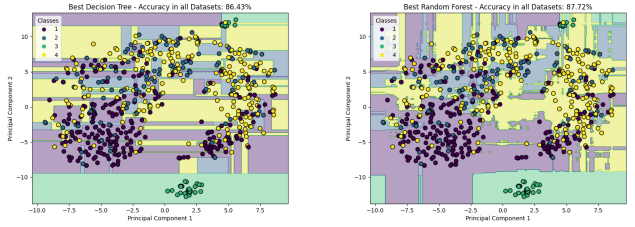

Fig. 5. Actual classification situation of decision tree and random forest dimensionality reduction to 2 dimensions

III. LINEAR SUPPORT VECTOR MACHINES AND NONLINEAR SUPPORT VECTOR MACHINES

It can be seen from Table 4 that the test set has the highest accuracy when nonlinear support vector machine does not extract features. Compared with the previous significant improvement from random forest to decision tree, the nonlinear linear improvement of support vector machine here is not very significant. Figure 5 shows that just a simple cut in two

TABLE IV
CLASSIFICATION OF LINEAR SUPPORT VECTOR MACHINES AND NONLINEAR SUPPORT VECTOR MACHINES UNDER DIFFERENT FEATURES

Feature	LSVM Cross-validation accuracy	LSVM Test accuracy	NLSVM Cross-validation accuracy	NLSVM Test accuracy
Grade, Total, MCQ, Q2	0.606	0.637	0.600	0.629
Grade, Total, MCQ, Q2 ,Q4	0.585	0.677	0.597	0.661
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.595	0.661	0.595	0.645
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.589	0.661	0.583	0.645
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.583	0.661	0.585	0.653
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.591	0.677	0.600	0.693

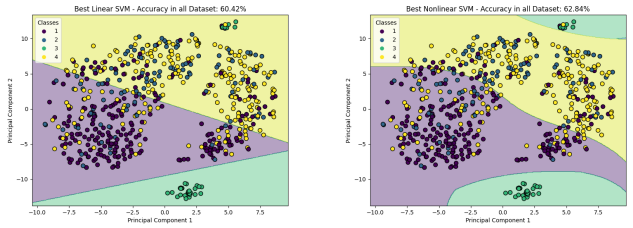

Fig. 6. Linear support vector machine and nonlinear support vector machine in 2-dimensional classification

dimensions is obviously unable to cope with the intertwined distribution of the data, and the higher linear accuracy is entirely caused by whether the upper left corner needs to be cut.The two confusion matrices are similar to the random forest and decision tree above.

IV. NAIVE BAYES

TABLE V
CLASSIFICATION OF NAIVE BAYES UNDER DIFFERENT FEATURES

Feature	Cross-validation accuracy	Test accuracy
Grade, Total, MCQ, Q2	0.464	0.637
Grade, Total, MCQ, Q2 ,Q4	0.456	0.677
Grade, Total, MCQ, Q2 ,Q4 ,Q5	0.460	0.693
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3	0.472	0.661
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1	0.478	0.653
Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender	0.474	0.645

It can be seen from Table 5 that Naive Bayes does not extract the least features or the most features with the highest accuracy. It is worth noting that his performance on the test set is particularly bad. The main reason is that Naive Bayes assumes that features are independently and identically distributed. For discrete feature classifiers, it calculates the conditional probability of each value of each feature corresponding to each class. In the prediction phase, given a new observation, Naive Bayes uses these observed feature values

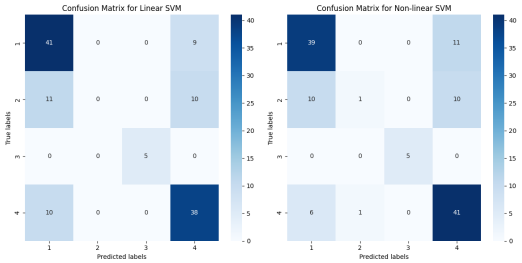

Fig. 7. Confusion matrices for linear support vector machines and nonlinear support vector machines

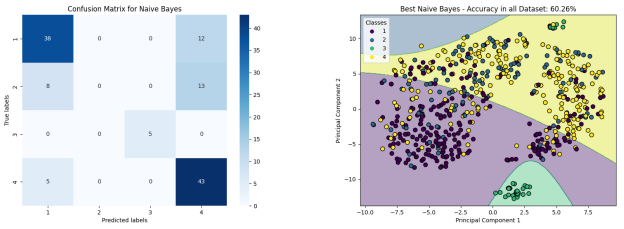

Fig. 8. Naive Bayes' confusion matrix and classification in 2 dimensions

combined with the learned probabilistic model to estimate the probability of belonging to each class. When it is assumed that the distribution of independent and identically distributed features in different categories is not very different, its impact on classification will be smaller. One evidence is in Figure 8. In the two-dimensional graph we can see that Bayes is completely incapable of classifying the blue data points.

V. Ensemble Learning

TABLE VI
CLASSIFICATION OF CURRENT STATE OF ART MODELS UNDER DIFFERENT FEATURES

Model with specific feature	Cross-validation accuracy	Test accuracy
Descision Tree (Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender)	0.553	0.653
Random Forests (Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1r)	0.591	0.661
NLSVM (Grade, Total, MCQ, Q2 ,Q4 ,Q5, Q3, Q1, Gender)	0.600	0.693
Naive Bayes(Grade, Total, MCQ, Q2 ,Q4 ,Q5)	0.460	0.693
Ensemble	0.597	0.653

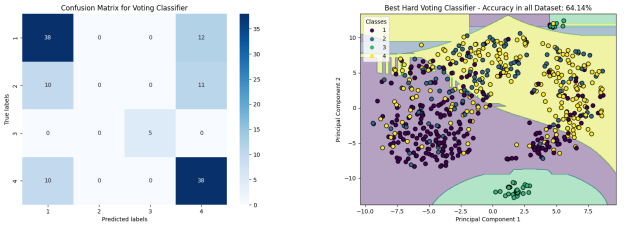

Fig. 9. Confusion matrix and 2-dimensional classification of ensemble learning

My ensemble learning uses a hard voting method to perform ensemble learning on the top performers. They consist of three classifiers: decision tree based on all features, nonlinear support vector based on all features and naive Bayes based on which performs best Characteristics. As can be seen from Table 6, this integration method does not have the highest accuracy in the test set. Of course, I use stacking and energy models to estimate that the highest accuracy can reach 90 percent. This does not mean that I cannot just focus on the advantages of this integrated learning. where. If we carefully observe Figure 9, we can find that the 2-dimensional graph shows a large number of perceptions of our human body. It can be said that reducing the variance and error between models is not the main means to improve the performance of the model. We can look back at my previous 2-dimensional random forest The figure achieves 87 percent accuracy, which shows that our dimensionality reduction is more important and more efficient than ensemble learning in solving discrete classification problems.

VI. Conclusion and Discussion

In this report ,we compares decision trees, random forests, support vector machines, naive Bayes and ensemble learning for data classification. In addition, we map these models to two dimensions and use tsne dimensionality reduction technology to visualize the hyperplane of the model, so as to better understand how to select models and extract features under classification problems, and explain the causes of phenomena based on theory and experiment.