A classification practice employing machine learning mehods

Jingtong.Cai21 (2144998)

Lab-D1-Group-4

May 17, 2023

1 Introduction

This report mainly discussed about the result of the coursework2 which requires to build a classifier to classify patients into the corresponding class, with given dataset containing 5000 patients' data labeled with 0 and 1 and 2. In this lab label2 data is deleted while the other 2 is classified accordingly. Classifiers built in this task are SVM, decision tree and Logistic Regression for supervised way and k-means for unsupervised way, while SVM performs best and k-means works well when k equals 3. The task is completed in Java language.

2 Dimensionality reduction

PCA is a dimensional reduction method used in machine learning to reduce the dimensionality of datasets with large number of features[1].

Before input original data into pca, firstly preprocess data by putting all data of label 0 in array[class_0] and corresponding label 1 in array[class_1]. Secondly, use concatenate method to concatenate them together as a big matrix. Then after standardization, the data is fitted into pca.

Determine the number of features of dimension is concerned with the specific dataset and corresponding application scenario. Cumulative sum of explained variance ratio can be used to decide the number of principal components. It is generally accepted that cumulative sum of explained variance ratio should reach to a specific threshold value to remain most of the information while at the same time reduce dimension. The result of Cumulative sum of explained variance ratio is shown in Fig1.

In addition, application scenario should also be considered since the aim of dimensional reduction is sometimes concerned with data visualization. In this case, it is more convenient to reduce the dimension to 2 or 3 so that principal components can be plotted in 2D or 3D graphs more easily. This task chooses to reduce the 15 features to 2 mainly out of this point of consideration which is shown in Fig2.

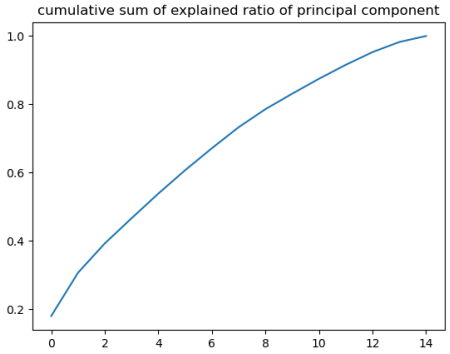

Figure 1: Cumulative sum of explained variance ratio

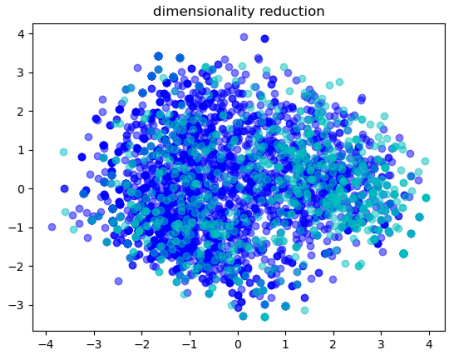

Figure 2: dimensional reduction to 2

3 Training Classifiers in a Supervised Way

Supervised learning is a machine learning approach that trains models in a labeled dataset[2]. Various types of classifiers are used in supervised learning and three classifiers are adopted in this task :SVM, decision tree, and logistic regression. Below are some general introductions of these algorithms:

SVM is a learning model that use classification algorithm to find best hyperplane that separate data into two groups[3]. Decision tree is a non-parametric algorithm that has a tree -based hierarchy and applies a divide and conquer strategy[4]. Logistic regression model predicts the probability that what class an instance belongs to by employing a sigmoid function and its input is the output of linear regression function.

To pre-process the input data, a cross-validation method provided by sklearn library is employed: from sklearn.model_selection import train_test_split, and the original dataset is split into trainset and test set randomly. It is a simple technique which is easy to implement. Before throw data into split set, a concatenate method provided by numpy library is used to concatenate labels 0 and 1 as y. Arrays [class_0] and [class_1] are concatenated as x and later fit into pca method as train_x. Train_x and y is the input data of split method.This adjustment will slightly improve the following accuracy.

With the aim of comparing three classifiers comprehensively, evaluation criterion f1_score is adopted. It is a combination of accuracy score and recall score. Table below shows the details of evaluation results of each classifier, from which a conclusion can be made that SVM performs best in precision of fitting prices, while logistic regression follows closely behind.

model	Precision	Recall	F1
SVM	0.743	0.679	0.696
Logistic Regression	0.741	0.679	0.695
Decision tree	0.675	0.571	0.604

4 Unsupervised Classification

Unsupervised leaning is a branch of machine learning different from supervised learning which attempts to search for similarity in an unlabeled dataset. The most prominent task includes clustering and association where k-means is a typical method of clustering, which means to partition the unlabeled data into a certain number of non-overlapping groups using some criteria such as the nearest distance between center point[5].

The Unsupervised classifier chosen in this lab is k-means mentioned above. K-means algorithm groups data into clusters so that the data points belong to the certain cluster have smaller distance between this centroid that between other. It has been actively applied in a wide range of fields and is simple to implement since the package is already provided in some java libraries.

To implement the k-means algorithm, first import k-means package from sklearn.cluster. Then instantiate kmeans by build a constructor. Use the constructor to call fit method in sklearn to input trained data as learning model. The input data here is the result after dimensionality reduction of pca. The number of group patients are divided into (usually k to represent) is the n_cluster variable passed into Kmeans constructor. Since there is no dedicated k that the patients should be grouped into, several different numbers of k are tried inside a for loop. In this task, numbers in 3 to 5 are tested. The classification results are plotted via scatter in Fig5.

$$s = \frac{disMean_{out} - disMean_{in}}{max(disMean_{out}, disMean_{in})}$$

Figure 3: silhouette coefficient formula

```
When cluster= 3
The silhouette_score= 0.4338015721859068
5306/5330

When cluster= 4
The silhouette_score= 0.39697134294635394
5267/5330

When cluster= 5
The silhouette_score= 0.3805121309069131
5277/5330
```

Figure 4: silhouette coefficient result

To make fully comparison of each condition, an evaluation criterion is import from sklearn.metrics to calculate the performance of a clustering: Silhouette Coefficient. 3 shows the definition of Silhouette Coefficient, the main calculation is the distance between points in this cluster and in other clusters. Additionally, silhouette_samples is applied to calculate the number of points whose Silhouette Coefficient is positive. Fig4 is an instance of execution result which is also displayed in the table below. From the comparison of visualize results and silhouette score, number 3 is chosen as the number of groups patients should be divided into.

n-cluster	silhouette coefficient	number of right
3	0.434	5315/5330
4	0.397	5252/5330
5	0.381	5274/5330

5 Conclusion

This report mainly focuses on the explanation of the principal of lab design and displays the results. First

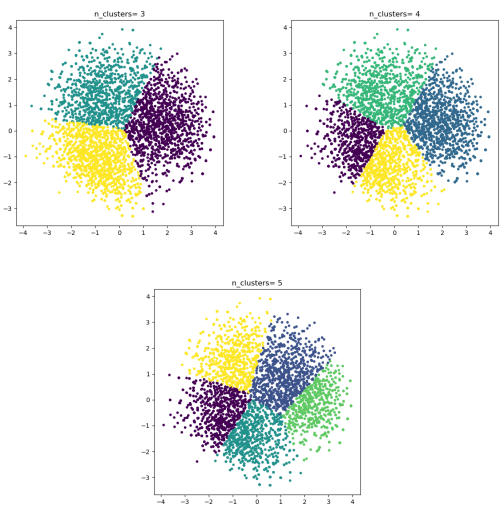

Figure 5: Three n-clusters

use pca to reduce the data dimension to 2, second supervised classification takes the results of pca as input, while cross-validation is employed to preprocess the data. Third part is concerned with unsupervised classification, while k-means is used and proved to perform well when number of clusters is 3.

There are few points to improve in this lab such as the cross-validation method is too straightforward and the dimensional reduction result is a bit distortion compared with the original data. However, the main requirement is basically met.

References

[1] J. Yang, D. Zhang, A. F. Frangi, and J. yu Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 131–137, 2004.

[2] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica (Slovenia)*, vol. 31, pp. 249–268, 2007.

[3] T. Joachims, "Making large scale svm learning practical," *Technical reports*, 1998.

[4] B. Charbuty and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, 2021.

[5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015.