Machine Learning Project Report

Li Sun
002433646

1. Classification and Prediction

1.1 Datasets 1-5

In this process, several classification algorithms are applied to category test data on the basis of the training dataset. The classification procedure is descripted as follow:

First, before performing classification algorithms, data preprocessing is executed. The missing values are filled with the means of their corresponding columns.

Then the Traindata samples are split into training dataset and validation dataset. Here *stratified cross validation* method is chosen so that each fold contains roughly the same proportions of the types of class labels.

In order to identify the best model for each dataset, 4 classifiers, SVM, Naive Bayes, KNN and Random Forest, are built. Each Model is generated by performing each classifier on the training dataset, then the model is applied on the validation data to generate prediction accuracy.

Finally, the comparison of prediction accuracy on these 4 classifiers are made. The classifier with the highest accuracy is selected and applied to the test dataset for classification.

As for the KNN algorithm, one simple approach for k selection is to set a reference range of K around the value of sqrt(n), which n is the observation size of the training dataset. The k with the highest accuracy is determined as the final k value for the KNN model.

Table 1 shows the prediction accuracy of 4 classifiers and the model selected for each dataset. The last column 'Choice' lists all the selected algorithms.

Table 1: Prediction accuracies of 4 classifiers for each dataset

	SVM	Naive Bayes	KNN	Random Forest	Choice
Dataset 1	0.9310	0.9655	0.9138 (k=13)	0.9310	**Naive Bayes**
Dataset 2	0.3055	0.7500	0.8056 (k=9)	0.8611	**Random Forest**
Dataset 3	0.3422	0.3048	0.3279 (k=79)	0.3322	**SVM**
Dataset 4	0.9015	0.4852	0.6526 (k=25)	0.9665	**Random Forest**
Dataset 5	0.5765	0.5203	0.5293 (k=11)	0.6396	**Random Forest**

1.2 Dataset 6

In this process, the real numeric targets in the test dataset are predicted.

Similarly as mentioned above, the data preprocessing is executed first. All the missing values are filled with the means of corresponding columns. Then the stratified cross validation method is applied to generate training and validation datasets.

Considering the target data type of numeric, KNN, Linear Regression Model (LM) with stepwise regression and Generalized Linear Regression (GLM) with Lasso, ridge or elastic net regularization models are built. The model with the lowest Mean Squared Error (MSE) is selected and applied to generate prediction accuracy for the test dataset. Table 2 lists the MSEs of the algorithms for dataset 6. The GLM with Ridge regularization which has the lowest MSE is selected as the final model.

Table 2: MSEs of the algorithms for each dataset

	MSE
KNN	3.768e+12
LM with Stepwise regression	2.506e+12
GLM with Lasso regularization	2.074e+12
GLM with Ridge regularization	**<u>1.988e+12</u>**
GLM with Elastic Net regularization	2.054e+12

2. Missing Value Estimation

There are nearly 4% missing values in the Dataset 1 and about 10% missing values in the Dataset 2. KNN algorithm is applied to do the imputation on these two datasets.

However there are nearly 83% missing values in the Dataset 3. Since the MICE algorithm can handle a large amount of missing data, it is utilized here to estimate the missing data in Dataset 3.

3. Multi-label Classification

For the Multi-label classification, the key is to transform data so that it is suitable to traditional classification algorithms. In this work, Binary Relevance (BR) approach is used. BR transforms the original multi-label dataset into several binary datasets, as many as different labels there are. The dataset here has 14 labels. As a result, 14 binary datasets are produced. For each binary dataset, a best model is chosen for prediction. Then all individual predictions are combined to generate the final output.

Here 5 algorithms are utilized. They are SVM, Naive Bayes, KNN (k=13), Random Forest and Logistic Regression. For each label all algorithms are used while the best one is picked out. Table 3 shows the accuracy using the BR transformation method with split ratio = 70:30 on stratified partitioning. The last column lists the best algorithm for each label.

Table 3: Prediction Accuracy using BR transformation method (ratio=70:30)

	SVM	NB	KNN(k=13)	RF	LR	Best
Label 1	0.7566	0.7566	0.7434	0.7566	0.6776	SVM
Label 2	0.6513	0.5855	0.6579	0.5855	0.6316	KNN
Label 3	0.7237	0.7434	0.7171	0.7434	0.6711	NB
Label 4	0.7434	0.7303	0.7105	0.7303	0.6579	SVM
Label 5	0.7829	0.7434	0.7763	0.7434	0.6711	SVM
Label 6	0.7829	0.6776	0.7697	0.6776	0.6974	SVM
Label 7	0.8224	0.7697	0.8224	0.7697	0.6645	SVM
Label 8	0.8158	0.7303	0.8092	0.7303	0.6513	SVM
Label 9	0.9342	0.8421	0.9342	0.8421	0.7829	SVM
Label 10	0.9276	0.8092	0.9276	0.8092	0.7632	SVM
Label 11	0.9211	0.7763	0.9211	0.7763	0.7500	SVM
Label 12	0.7303	0.6053	0.7237	0.6053	0.6645	SVM
Label 13	0.7303	0.6316	0.7171	0.6316	0.6447	SVM
Label 14	0.9803	0.9803	0.9803	0.9803	0.9803	SVM

In this project, different split ratios of stratified partitioning are applied to obtain the algorithms prediction accuracy on each label. The final algorithm is chosen by majority voting. Table 4 lists the best algorithms with different split ratios for each label. The last column 'Choice' gives the final selections, which are applied on test data to generate the final classification result.

Table 3: Best algorithms with different ratios

	60:40	65:35	70:30	75:25	80:20	Choice
Label 1	SVM	SVM	SVM	SVM	SVM	**SVM**
Label 2	KNN	KNN	KNN	KNN	NB	**KNN**
Label 3	SVM	SVM	NB	SVM	NB	**SVM**
Label 4	KNN	SVM	SVM	SVM	SVM	**SVM**
Label 5	SVM	SVM	SVM	KNN	KNN	**SVM**
Label 6	SVM	SVM	SVM	KNN	KNN	**SVM**
Label 7	KNN	SVM	SVM	KNN	SVM	**SVM**
Label 8	KNN	KNN	SVM	KNN	KNN	**KNN**
Label 9	SVM	SVM	SVM	SVM	SVM	**SVM**
Label 10	SVM	SVM	SVM	SVM	SVM	**SVM**
Label 11	SVM	SVM	SVM	SVM	SVM	**SVM**
Label 12	SVM	SVM	SVM	SVM	SVM	**SVM**
Label 13	SVM	SVM	SVM	SVM	SVM	**SVM**
Label 14	SVM	SVM	SVM	SVM	SVM	**SVM**