CDS503: Machine Learning

Assignment 1: Classification

1)

a. Report the class distribution. Is this a balanced or unbalanced data set?

Answer: It is a unbalanced data set.

Due to the class label's ratio of YES and NO is 136:155, not 1:1.

b. Please select and justify a suitable metric to evaluate the performance of your classification model.

Answer: I choose 'Validation Accuracy' to evaluate the performance of my classification model.

Confusion Matrix:

$$\begin{array}{ccc} \square & P & N \\ P' & TP & FP \\ N' & FN & TN \end{array}$$

P in the confusion matrix means Positive, and N means Negative. In the matrix, FP represents the number of samples that are actually negative but predicted to be positive, TN represents the number of samples that are actually negative but predicted to be negative, TP represents the number of samples that are actually positive but predicted to be positive, and FN represents the number of samples that were actually positive but predicted to be positive.

The accuracy rate is the proportion of correctly classified samples to the total number of samples

So :

Accuracy = $\frac{n_{correct}}{n_{total}} = \frac{TP+TN}{TP+TN+FP+FN}$

c. Given the size of the data set, which validation option (e.g., percentage split, k-fold cross validation) do you think is suitable to be used in your machine learning experiments. Specify the validation option you are selecting for your machine learning experiments. Briefly explain the reason for using the validation option.

Answer: I think "k-fold cross validation" is suitable to be used in my machine learning experiments.

Reason: 1.The size of the data set is too small (about 300), k-fold cross validation makes efficient use of limited data, and the evaluation results can be closer to the performance of the model on the test set than other validation options. Although using k-fold cross validation increases training time, the size of dataset is too small so the training time won't increase too much.

2.The size of the data set is too small (about 300), we don't need to use "percentage split" because it's suitable for large dataset(more than 2000). If we do this we will use less data to train the model and the accuracy will decline.

2)

Report the machine learning experiments you ran and identify the best performing model using Freq-PHOBinary and also the best performing model using Norm-PHO-Binary (both the best performing models should yield higher performance than your random baseline). Record all the results of your experiments in A1_Experiment_Sheet.xlsx and highlight the row indicating the best performing model.

Answer: I use svm, decision tree and knn model to run on the Freq-PHOBinary and Norm-PHO-Binary respectively.

Why I choose these three classification algorithms?

I choose svm because:1.The result of SVM is determined by a few support vectors, and the computational complexity depends on the number of support vectors, not the dimension of the sample space, which avoids the "curse of dimensionality". 2.In this experiment, the size of dataset is small and SVM is good at working with small size of dataset. 3.This experiment is a binary classification problem, SVM is hard to do multivariate classification but good at binary classification.

I choose decisiontree because:1.Decisiontree need less data to train the model compared to other algorithms. 2. Easy to understand and explain. tree structure visualization

I choose KNN because:1.Not sensitive to outliers.2.one of the disadvantages of KNN algorithm is that need large amount of calculation so the training time will be long, but in this experiment, the size of dataset is small so the training time won't be too long.

Here are the result of the experiment:

Feature Representation	**Machine Learning Algorithm**	**Parameters**	**Validation Option**	**Validation Accuracy**
Freq	DummyClassifie	strategy=	Validation Accuracy	0.5051724

	r	uniform		14	
Norm	DummyClassifier	strategy=uniform	Validation Accuracy	0.480689655	
Freq	**svm**	**kernel = 'linear'**	**Validation Accuracy**	**0.618505747**	
Norm	svm	C=8.0	Validation Accuracy	0.611724138	
Freq	knn	n_neighbors=61	Validation Accuracy	0.60183908	
Norm	knn	n_neighbors=17	Validation Accuracy	0.59137931	
Freq	decisiontree	criterion='entropy', max_depth=14,class_weight="balanced",splitter = 'best',	Validation Accuracy	0.535977011	The best model using Freq-PHOBinary is **svm**.
Norm	**decisiontree**	**criterion='entropy',max_depth=8,splitter = 'best',min_samples_split = 17**	**Validation Accuracy**	**0.692068966**	The best model using Norm-PHO-Binary is **decisiontree.**

Which feature representation produces a better model? Explain how you determine the best performing model based on the performance metric you have selected. Can you explain why one feature representation is better than the other?

Answer:

1.Norm-PHO-Binary representation produces a better model.

2.I use k-fold cross validation (k = 10) to calculate each model’s accuracy and compare to other models. If one model’s mean accuracy is highest, it is the best performing model.

Here are specific calculation steps:

First, divide all samples into 10 equal-sized sample subsets, then traverse the 10 subsets in turn, using the current subset as the validation set each time, and other samples as the training set, train and use Validation Accuracy to evaluate the model. Finally, the average value of 10 Validation Accuracy is used as the final evaluation indicator.

3. Because Norm-PHO-Binary is a standardization dataset compared to Freq-PHOBinary.

Data standardization can eliminate the influence of numerical imbalance on calculation results, in this example, assume that in Freq-PHOBinary one person’s feature such as “Emotion_Joy”’s number is much bigger than others, than the result will heavily biased on this person, but if we do data standardization such as Norm-PHO-Binary do, because the sum of all emotions is standardized to 15, everyone’s sum of number is 15, so the result won’t bias to the person whose number is much larger than others.

4)

Is your overall best performing model able to achieve very promising results (reach at least 0.8 of your selected performance metric)? Provide two suggestions on future strategies that can be used to improve the performance of your best performing model.

Answer: 1.No

2. (1)We have noticed that the data set was an unbalanced. So before training the model, we can preprocess the dataset to balance the dataset. Use unbalanced dataset to train will make the result bias, so we need to find best way to preprocess the dataset before training.

(2)We can use the neural network to find the best parameters of the model, in my experiment, The parameters are determined based on my experience and debugging based on the control variable method again and again, which is undoubtedly very inefficient and difficult to find the best parameters. Because the data set in this experiment is small, the model training time is short, so use the neural network can quickly find the best parameters, and it is very suitable for this experiment.