Machine Learning: Computer Exercise 4

机 53　陈声健　2015010509

1. Task Description

For this exercise, new data is provided. For the first task, we are required to use one kind of available algorithms to do feature selection and use one of the previous classification methods to classify E6 vs. E7. And then try the two approaches of cross-validation and observe their effect on the assessment of the performance and the time consumption. In the optional task, we should try R-SVM to do wrapped feature selection and classification of E6 and E7 cells. The optional task is not done here.

2. Experiment Design

In the first task, I used MLP as the classifier. There are many feature selection algorithms provided in sklearn package and I tried out most of them. For most of the feature selection algorithms, the procedure is quiet similar: specified the required parameters, fit and transform the X dataset, fit the classifier with the newly selected features on training set and evaluate the model on the test set.

In CV1, I first used all samples to do feature selection and build a new dataset with the selected features. And then I divided the dataset into 5 folds and do cross validation, leave 1 fold out each time, train the MLP on the other folds and test on this fold. Results of accuracy on the 5 tests are averaged and returned, and the feature selection time and cross validation time are recorded.

In CV2, I first divided the dataset into 5 folds. For each time, I left out 1 fold and do feature selection on the other folds and fit the model on them. And then test the model with the left out fold. The accuracy of each test was also averaged. The running time was recorded as mentioned above.

3. Method

I used *sklearn* and *tensorflow* packages on python3.6 for this exercise. For the feature selection algorithms, I tried *VarianceThreshold* with threshold = 0.8, *SelectKBest* with k = 500 and Chi-squared stats, *SelectFromModel* with SVC model and *ExtraTreesClassifier* with n_estimators = 50. For the experiments, the MLP was trained for 200 epochs.

For cross validation with feature selection, I used the *SelectKBest* algorithms with k = [20, 100, 200, 500, 1000], the dataset was divided into 5 folds, and the MLP was trained 100 epochs.

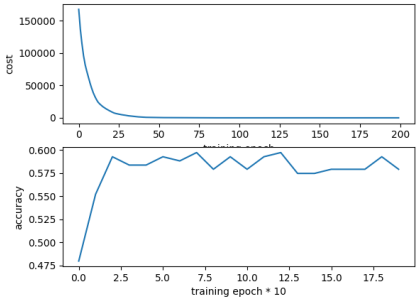
4. Result and Observation

As indicated in the file name of the dataset “no info”, the result of the classification is around 50% or a little bit higher (Fig.1). After feature selection, the size of dataset is reduced thus training can be finished in shorter time.

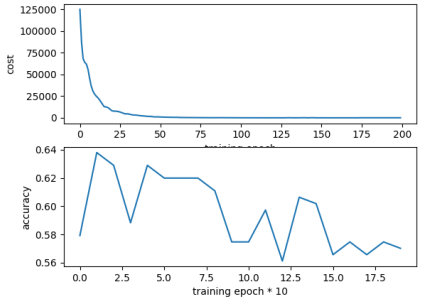
The result of the two schemes of CV is illustrated in table.1. As mentioned in the lecture, CV1 tend to be over optimistic. An example of “artificial fake class data” was also shown in the slide. However, in my experiment, I could not observe that much difference. The CV1 seemed to be a little bit more optimistic than CV2, but not that much. Intuitively, the do

feature selection and fit the model with the training data and then apply the model to new unseen data. Thus, CV2 seems to be more reasonable to evaluate the model. While CV1, which also take the test set into consideration in feature selection, will fit the data better and be more optimistic to the performance of the model.

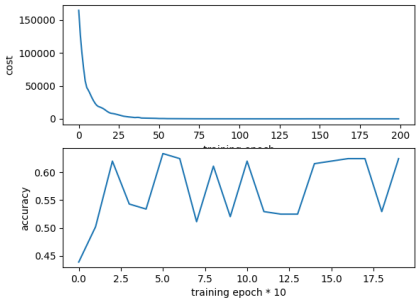
CV2 takes less time to do feature selection because the size set of the dataset became smaller (k-1 folds). However, it is observed that CV2 take more time in cross validation, which is mainly the process of training the MLP.



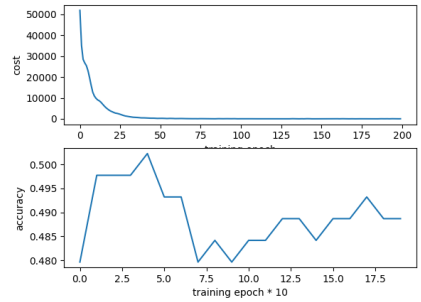
(a)



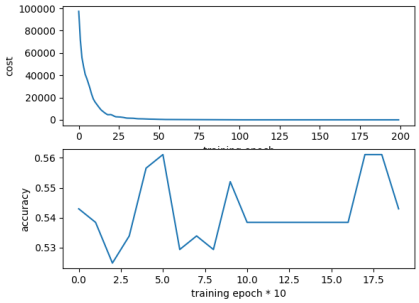
(b)



(c)



(d)



(e)

Fig.1 result of feature selection, (a) no feature selection; (b) VarianceThreshold; (c) SelectKBest; (d) SelectFromModel; (e) ExtraTreesClassifier

N_features	CV1			CV2		
	accuracy	Selection time / s	CV time / s	accuracy	Selection time	CV time
20	0.53017	0.04970	15.17790	0.59703	0.00014	25.45211
100	0.55853	0.04748	18.18312	0.50846	0.00014	31.09977

200	0.54246	0.04859	23.65527	0.53011	0.00013	33.73683
500	0.58229	0.05968	31.37909	0.53122	0.00014	38.61487
1000	0.57890	0.06759	40.14112	0.54709	0.00014	46.01771

Table. 1 result of two schemes of cross validation

5. Conclusion

In conclusion, in the term of evaluating model performance, the result of CV1 can be biased and over optimistic, and CV2's result is more reasonable. CV1 takes more time to do feature selection while CV2 takes more time to do cross validation. The dataset seems to contains little information as indicated in its name. The accuracy of the classification on E6 vs. E7 with several different models is not much better than random guess.

6. Reference

[1] sklearn.model_selection.KFold. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[2] Feature selection. https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection