Ensemble Model

In the last two sections, we have implemented unsupervised learning and supervised learning on Fashion-MNIST classification respectively. Now, we combine unsupervised learning and supervised learning together to build a joint model with a better accuracy of classification.

PCA-SVM

We first combine principal component analysis (PCA) and support vector machine (SVM) to create a classification pipeline. In the first stage, we apply PCA to reduce the dimensions of feature space as well as make features orthogonal to each other. In the second stage, we select components with great deviation among principal components, and use them as input to fit a SVM classifier.

We use the elbow method with the cutoff 1.75 and take all the components which have standard deviation above that value, which gives 28 components. We use those 28 components to fit a multiclass kernel SVM with RBF kernel, which is demonstrated in the classification part. Before performing PCA, we preprocess the data by subtracting the mean and scale it to variance 1.

The confusion matrix and mis-classification rate for each class are shown in the following two tables. The overall mis-classification rate of PCA-SVM classifier is 0.112 and the testing accuracy is 0.888, which is better than random forest but worse than the origin SVM.

Ytest	Ytest_pred_psvm									
	0	1	2	3	4	5	6	7	8	9
0	855	3	16	24	2	1	93	0	6	0
1	6	978	2	14	0	0	0	0	0	0
2	16	0	808	14	78	0	78	0	6	0
3	23	13	13	900	36	1	12	0	2	0
4	1	2	73	22	848	0	53	0	1	0
5	0	0	0	0	0	946	0	32	2	20
6	135	1	71	30	66	0	689	0	8	0
7	0	0	0	0	0	28	0	938	0	34
8	4	0	8	2	1	2	8	3	970	2
9	1	0	0	0	0	15	0	38	1	945

(a) Confusion Matrix

class	mis−classification
0	0.179
1	0.019
2	0.185
3	0.105
4	0.177
5	0.047
6	0.262
7	0.072
8	0.026
9	0.056
overall	0.112

(b) Mis-classification Rate

Figure 1: PCA-SVM

Though our PCA-SVM classifier does not improve the classification performance compared with the origin SVM, the classifier still explores some features of the data that the origin model may not have explored. Our next step is to combine PCA-SVM and SVM, in additional to random forest and k-NN model to build an ensemble classifer.

Ensemble Model

The classification models are limited to certain type of data structure. However, in real-world problems, it is difficult to find a model that well satisfied the data structure of the problem. Thus, ensemble multiple classification models and combine the information together may result in a more robust model which has a better performance than the single model.

We combine four different models, k-NN, random forest, kernel SVM and PCA-SVM, to build an ensemble classification model. The model tuning and performance of the single model is demonstrated in the classification section. The following figure shows the architecture of the ensemble model.

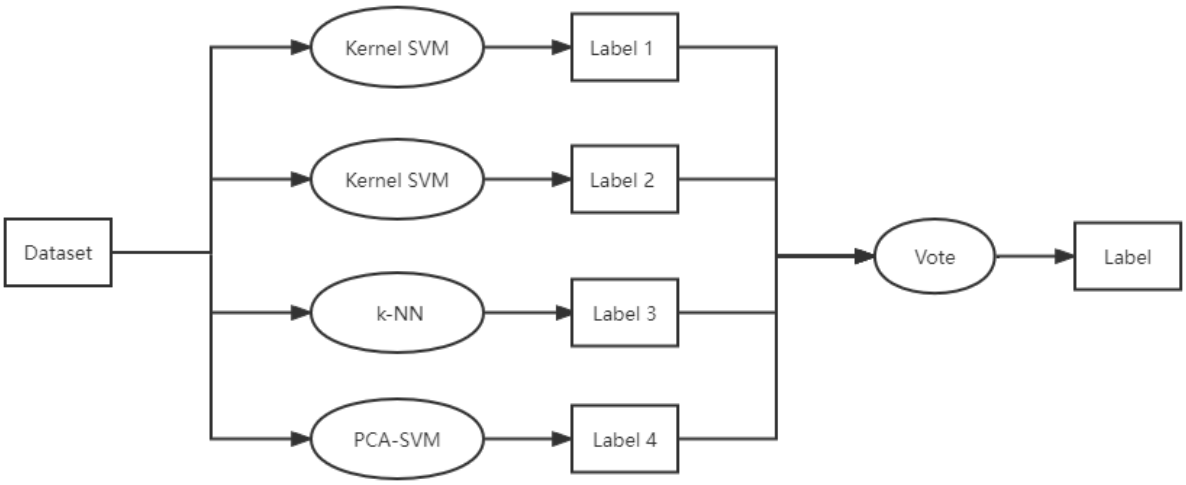

Figure 2: Ensemble Model Architecture

The classification pipeline has two stages. In the first stage, we let k-NN, random forest, kernel SVM and PCA-SVM classify input data respectively. Each classifier will give a predicted label for the input. In the second stage, we let those classifiers vote for the class of input, and the prediction result is the class that most classifiers vote for.

Now, we fit the ensemble model with Fashion-MNIST data. The confusion matrix and mis-classification rate for each class are shown in the following two tables. The overall mis-classification rate of the ensemble model is 0.089 and the testing accuracy is 0.911, which outperforms all of the single model.

	Ytest_pred_en									
Ytest	0	1	2	3	4	5	6	7	8	9
0	888	0	12	17	1	0	76	0	6	0
1	2	988	0	7	0	0	3	0	0	0
2	16	1	839	16	73	0	53	0	2	0
3	26	9	8	917	25	0	14	0	1	0
4	3	1	61	25	871	0	37	0	2	0
5	0	0	0	0	0	953	1	29	3	14
6	126	0	54	24	52	0	740	0	4	0
7	0	0	0	0	0	8	0	963	0	29
8	3	0	4	2	2	2	5	1	981	0
9	0	0	0	0	0	3	0	26	3	968

(a) Confusion Matrix

class	mis−classification
0	0.165
1	0.011
2	0.142
3	0.09
4	0.149
5	0.013
6	0.203
7	0.055
8	0.021
9	0.043
overall	0.089

(b) Mis-classification Rate

Figure 3: Ensemble Model

Conclusion

Finally, we take a comparison between the ensemble model and all other single models. The testing accuracy of different algorithms are shown in the table. We find that the ensemble model indeed utilizes the advantages from different single models and have the best performance among all algorithms.

Table 1: Model Summary

	kNN	Random.Forest	Linear.SVM	Kernel.SVM	PCA.SVM	Ensemble.Model
Accuracy	0.8559	0.8846	0.8122	0.9083	0.8877	0.9108

The computational cost for the ensemble model is relatively high since we need to fit each single model respectively. However, if we have already trained some classifiers and want to further improve the model's performance, the computational cost can be reduced.

Additionally, the ensemble model is very flexible since it is convenient to add new models or remove inappropriate ones to improve the performance of the ensemble model.