Lu Guannan

20454477

2017年11月11日

Project1 report

Compare and select NN as classifier

This problem is a two-classification problem. There are in total 57 continuous features separated by commas in the file and labels in another file. I test three classification by using cross-validation and select neural networks. The result are as follows:

	5-Fold CV	5-Fold CV after normalization
SVM	0.815	0.928
NN	0.815	0.942
LogisticRegression	0.923	0.923

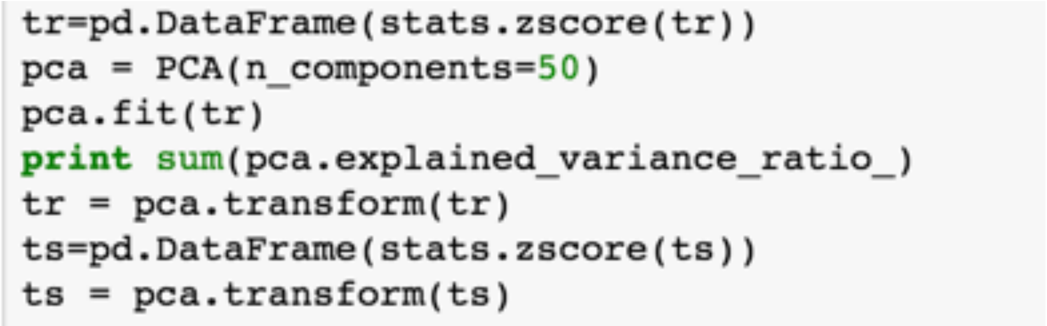
there are two steps for normalization

1. Use z_score to normalize the scale of data

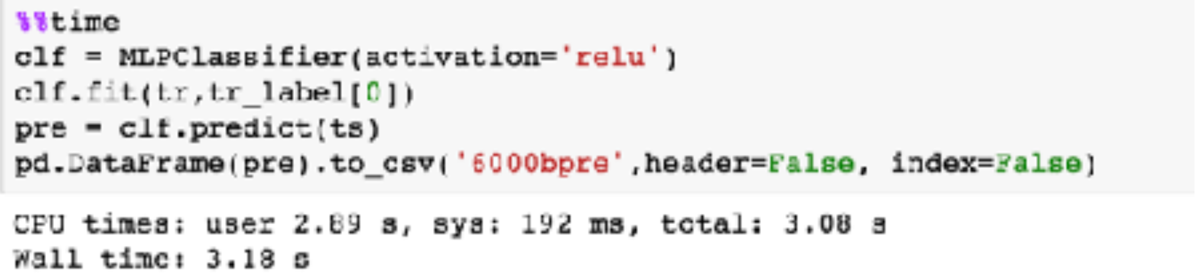
$$z = \frac{(x - \mu)}{\sigma}$$

2. Use PCA to reduce the dimension from 57 to 50

Finally. We choose NN as my classifier. And I use all the data to train the model.



```
tr=pd.DataFrame(stats.zscore(tr))
pca = PCA(n_components=50)
pca.fit(tr)
print sum(pca.explained_variance_ratio_)
tr = pca.transform(tr)
ts=pd.DataFrame(stats.zscore(ts))
ts = pca.transform(ts)
```



```
%%time
clf = MLPClassifier(activation='relu')
clf.fit(tr,tr_label[0])
pre = clf.predict(ts)
pd.DataFrame(pre).to_csv('6000bpre',header=False, index=False)
```

```
CPU times: user 2.89 s, sys: 192 ms, total: 3.08 s
Wall time: 3.18 s
```