Assignment 3: Unsupervised Learning and Dimensionality Reduction

Minming Zhao (gtID: 902685774)

In this report we will solve two same binary classification problems as previous assignments. The purpose of this report is to document the clustering procedure including k-means and Expectation Maximization, and perform feature dimension reduction algorithms including PCA, ICA, Random Projections and LDA. We use scikit learning package in python to implement all these algorithms.

Problem #1: Adult census income prediction

Money making is always an interesting question, isn't it? This is to predict whether one could make income more than $50k/y or not based on census data. With this study prediction, it help us understand how to make more money by locating some key features who did make over $50k/y.

There are 25800 observation in training set, in testing set. There are 64 features including dummies variables after changing categorical strings to dummies variables.

Problem #2: Student alcohol addiction prediction

To study student possibility to be alcohol addictive with relation to their school, gender, age, address, parents' job, their study time etc. Knowing better about student alcohol addiction could help families and teachers pay more attentions to those students who are more likely to be alcohol addictive and help students to study hard and let them enjoy other beautiful aspects of the life other than alcohol addiction.

There are 316 observation in training set, 79 in testing set. There are 38 features including dummies variables after changing categorical strings to dummies variables.

1. Clustering algorithms (k means, EM)

1.1 Adult data Clustering

We use KMeans in scikit-learn package to do K-means clustering algorithm. Since we have binomial data classification, we set n-cluster = 2 naturally and leave random_state at 0. Thus the Kmeans algorithm gives us accuracy 0.6155.

And we use GassianMixture in scikit learn package for the Expectation Maximum algorithm. We leave n_components = 2 and random_state at 0. It returns accuracy of 0.371.

We also have some clustering performance evaluations as below for both clustering algorithms.

Shorthand	full name
homo	homogeneity score: each cluster contains only members of a single class
compl	completeness score: all members of a given class are assigned to the same cluster.
v-meas	V measure: the harmonic mean of homo and compl.
ARI	adjusted Rand index: measures the similarity of the ground truth and prediction
AMI	adjusted mutual information: measures the agreement of the ground truth and prediction
silhouette	silhouette coefficient: evaluation of model clustering distances itself without ground truth

name	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
k-means++	2	1.06s	0	0	0	-0.006	0	0.511	0.616
k-means++	5	2.66s	0.001	0	0.001	-0.003	0	0.438	0.332
k-means++	10	4.68s	0.002	0	0.001	-0.002	0	0.478	0.096
k-means++	20	8.54s	0.011	0.002	0.004	0	0.002	0.298	0.076
k-means++	50	17.70s	0.016	0.002	0.004	0	0.002	0.258	0.032

As we could notice, in a supervise learning the accuracy drops as n cluster increases as we are dealing with a binomial problem. The homo and compl score all very low but they increases as number of clusters increases, which indicates that the number of clustering we have right now is not big enough and some labels are still mixed inside the clusters we have. We should increase the cluster numbers to have better score. However the Silhouette score drops especially after n_clusters at 20 and above, which indicate the clustering distance drops and elements in each clusters become more and more similar. Since we have 25800 oberservations in this dataset, it makes senses as well to have more than 50 clusters for an unsupervised learning because it does not know ahead we are dealing with binominal labels. However we could notice that the accuracy or the matchness of clustered labels with actual ground truth lables hold relatively reasonable / acceptable level, at number of clusters at 2, it achieves 61.6% accuracy. We define the accuracy as the matchness of predicted labels versus ground truth labels.

name	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
GaussianMixture	2	0.88s	0.021	0.028	0.024	-0.061	0.021	0.001	0.371
GaussianMixture	5	3.50s	0.067	0.032	0.044	0.049	0.032	-0.464	0.267
GaussianMixture	10	13.92s	0.078	0.036	0.05	0.078	0.036	-0.306	0.147
GaussianMixture	20	30.70s	0.073	0.023	0.035	0.044	0.022	-0.558	0.015
GaussianMixture	50	81.18s	0.126	0.028	0.046	0.027	0.028	-0.634	0.035

Similarly we could see results for GaussianMixture. Relatively speaking GaussianMixture has worse accuracy but better homo score and compl score. And it computes slower than kmeans. Based on Silhouette score we can see it clusters poorly and hold distances small in this dataset clustering.

1.2 Student Alcohol Clustering

Similar clustering analysis could be applied to student alcohol dataset. The observation is similar that k means seems to have better clustering accuracy compared to ground truth in this case than GaussianMixture although GassianMixture performs a little faster than kmeans.

name	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
k-means++	2	0.04s	0.121	0.082	0.098	0.095	0.08	0.127	0.658
k-means++	5	0.04s	0.076	0.023	0.035	0.009	0.019	0.086	0.237
k-means++	10	0.04s	0.121	0.025	0.041	0.007	0.018	0.083	0.101

k-means++	20	0.05s	0.147	0.023	0.04	0.006	0.012	0.035	0.057
k-means++	50	0.12s	0.301	0.037	0.066	0.003	0.014	0.122	0.009

name	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
GaussianMixture	2	0.01s	0.121	0.082	0.098	0.095	0.08	0.113	0.342
GaussianMixture	5	0.02s	0.036	0.011	0.017	0.012	0.007	0.068	0.073
GaussianMixture	10	0.03s	0.138	0.029	0.048	0.02	0.022	0.071	0.095
GaussianMixture	20	0.02s	0.177	0.028	0.049	0.008	0.017	0.11	0.016
GaussianMixture	50	0.05s	0.322	0.04	0.071	0.005	0.017	0.002	0.003

2. Apply the dimensionality reduction algorithms to the two datasets

The key code to apply dimension reduction algorithms in scikit-learn is below. Both datasets share the same dimensionarlity reduction algorithm codes. For simplicity in comparison, we set components holds at 5.

Principal Component Analysis:

```
PCA_data = PCA(n_components = 5,whiten=False)
PCA_data.fit(data)
PCA_data_trans = PCA_data.transform(data)
```

Independent Component Analysis:

```
ICA_data = FastICA(n_components = 5)
ICA_data.fit(data)
ICA_data_trans = ICA_data.transform(data)
```

Random Projection:

```
transformer = GaussianRandomProjection(n_components=5,eps=0.1)
RP_data_trans = transformer.fit_transform(data)
```

Linear Discriminant Analysis:

```
transformer = LinearDiscriminantAnalysis(solver="svd",n_components = 5)
LDA_data_trans = transformer.fit_transform(data)
LDA_data.fit(X=data,y=labels)
LDA_data_trans = LDA_data.transform(data)
```

We select component number at 5 and hold it constant so as to compare different algorithms. After dimension reduction, we will same observation rows but with 5 features which is generated by the dimension reduction features.

3. Reproduce your clustering experiments on the data after dimensionality reduction

3.1 Adult data

Then we reproduce/rerun clustering kmeans and GassianMixture for Expectation Maximum on the data transformed by dimensionality reduction. And we could see the results as following. We could notice that

the accuracy of k means still hold the same after dimension reduction while Gaussian Mixture performs much better accuracy from initially 0.37 to 0.774.

Name on PCA reduction	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
k-means++	2	0.23s	0	0	0	-0.006	0	0.642	0.616
Gaussian Mixture	2	0.14s	0.073	0.105	0.087	0.192	0.073	-0.061	0.774

ICA improves accuracy in both kmeans and GaussianMixture from 0.616 and 0.37 to 0.753 and 0.755 levels.

Name on ICA reduction	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
k-means++	2	0.23s	0.016	0.047	0.023	0.063	0.016	0.574	0.753
Gaussian Mixture	2	0.31s	0.0129	0.223	0.0245	0.0209	0.0129	0.777	0.755

RP still holds same accuracy for k means while worsen in GaussianMixture as the accuracy drops from 0.37 to 0.231.

Name on RP reduction	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
k-means++	2	0.24s	0	0	0	-0.007	0	0.53	0.619
Gaussian Mixture	2	0.13s	0.073	0.105	0.087	0.192	0.073	0.008	0.231

Hence, it seems the ICA transformation provides best feature dimension reduction while still holds better prediction accuracy.

3.2 Student Alcohol

Similar operation could be applied to student alcohol dataset. To contrast, we observed ICA and RP both performs better than that before feature transformation. However PCA does not change the accuracy, which is not too bad since it at least did the feature reduction and provide much fast processing time with same accuracy.

Name	n_clusters	time	Homogeneity score	Completeness score	v_measure score	adjusted_rand score	adjusted_mutual_info_score	Silhouette score	Accuracy
PCA k-means++	2	0.03s	0.121	0.082	0.098	0.095	0.08	0.257	0.342
PCA Gaussian Mixture	2	0.01s	0.121	0.082	0.098	0.095	0.08	0.248	0.342
ICA k-means++	2	0.04s	0.002	0.002	0.002	-0.016	-0.001	0.234	0.592
ICA Gaussian Mixture	2	0.04s	0.006	0.004	0.005	0.034	0.002	0.248	0.674
RP k-means++	2	0.04s	0	0	0	-0.004	-0.002	0.237	0.582

RP Gaussian Mixture	2	0.02s	0.03	0.023	0.026	0.078	0.02	0.226	0.684

4. Rerun your neural network learner on the newly projected data

4.1 Adult data

Without transformation we already have neural network model from previous assignments. For this dataset, we had accuracy 0.244 in test dataset.

As the table below shows, we were able to increase the accuracy when we apply neural network learner onto the transformed data after feature reduction.

ICA and RP seems to perform better than PCA here probably due to the natural of independent observation of this dataset.

n=2	without transformation	PCA	ICA	RP
Training accuracy	0.250039	0.482209	0.772132	0.755969
Test accuracy	0.244341	0.490233	0.755659	0.75938

4.2 Student Alcohol

The initial neural network has some overfitting, similarly ICA and RP feature transformation improves the testing accuracy. However PCA reduces testing accuracy due to it missed some information, in order to keep better accuracy, I believe more components need to be kept for PCA.

n=2	without transformation	PCA	ICA	RP
Training accuracy	0.917722	0.838608	0.822785	0.822785
Test accuracy	0.721519	0.632911	0.772152	0.772152

5. Treating the clusters as new features and rerun neural network learner

5.1 Adult data

When we add the clusters back to data to treat them as new features, we could rerun the neural networks and compare the accuracy.

n=2	without transformation	w/ PCA	w/ ICA	w/ RP
Training accuracy	0.250039	0.250039	0.750233	0.250039
Test accuracy	0.244341	0.244341	0.755814	0.244341

As we can see here, after adding new features, PCA and RP does not change the overall neural network accuracy while with ICA still holds a good accuracy. The processing speed become faster with features from clustering.

5.2 Student Alcohol data

Now with new features from clustering, all algorithms holds good training accuracy, especially PCA and RP, and RP seems to performs best because it is able to improve training accuracy while still could hold

good testing accuracy to avoid overfitting. The processing speed become faster with features from clustering.

n=2	without transformation	w/ PCA	w/ ICA	w/ RP
Training accuracy	0.917722	0.98417	0.89240	0.93670
Test accuracy	0.721519	0.62025	0.75949	0.77215

6. Conclusion

In conclusion, k means and Expectation Maximization clustering have been applied to two dataset (Adult and Student Alcohol dataset). Without any feature dimension reduction, kmeans performs better in both datasets than Expectation Maximization. After feature dimension reduction, PCA and ICA performs well than RP in Adult dataset while ICA and RP performs better in Student Alcohol dataset. The processing time improves significantly with feature reduction transformed dataset. Finally when adding new clusters as new features back to dataset, neural network accuracy was observed to increase in both training set and testing set for both dataset, also the processing time also improved compared to original without feature transformed features.

7. Reference

http://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation
http://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.GaussianMixture
http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans
http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html
http://scikit-learn.org/stable/modules/random_projection.html#random-projection