Clustering Kaggle tabular playground – July 2022

Tao Shan

Describe the data

The data is Kaggle manufacturing control data that can be clustered into different control states. Your task is to cluster the data into these control states. You are not given any training data, and you are not told how many possible control states there are. This is a completely unsupervised problem, one you might encounter in a real-world setting. Totally it has 32 columns and 98k rows.

Objective

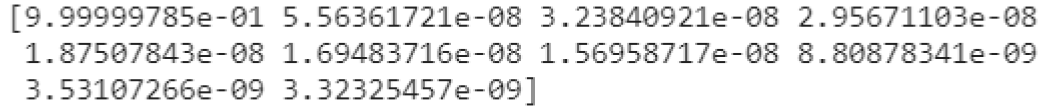
This model aims to find the correct groups in the kaggle competition, based on 32 columns of numerical continuous data. We don't know how many groups it should be, so the optimal groups we also need to find it.

Steps:

Check multicollinearity

PCA:

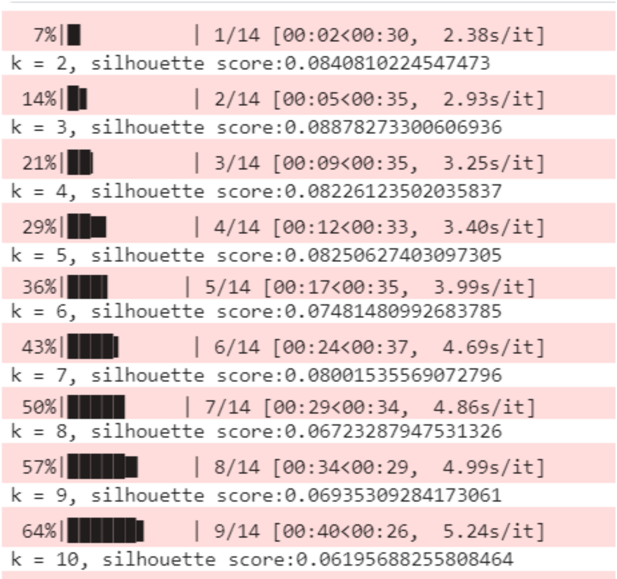
Explained variance ratio:



```
[9.99999785e-01 5.56361721e-08 3.23840921e-08 2.95671103e-08
 1.87507843e-08 1.69483716e-08 1.56958717e-08 8.80878341e-09
 3.53107266e-09 3.32325457e-09]
```

This is the solution for top 10 attributes generated by PCA. So the 1st attributes is the most important.

Check silhouette score:



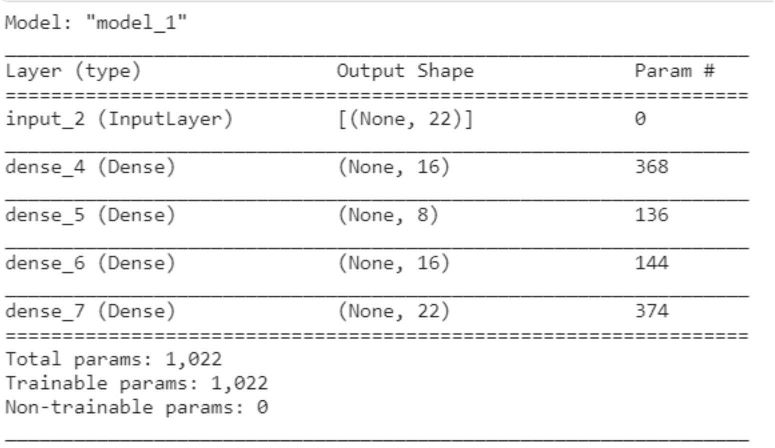
```
  7%|█          | 1/14 [00:02<00:30,  2.38s/it]
k = 2, silhouette score:0.0840810224547473
 14%|█▌         | 2/14 [00:05<00:35,  2.93s/it]
k = 3, silhouette score:0.08878273300606936
 21%|██▏        | 3/14 [00:09<00:35,  3.25s/it]
k = 4, silhouette score:0.08226123502035837
 29%|██▊        | 4/14 [00:12<00:33,  3.40s/it]
k = 5, silhouette score:0.08250627403097305
 36%|███▌      | 5/14 [00:17<00:35,  3.99s/it]
k = 6, silhouette score:0.07481480992683785
 43%|████▎      | 6/14 [00:24<00:37,  4.69s/it]
k = 7, silhouette score:0.08001535569072796
 50%|█████     | 7/14 [00:29<00:34,  4.86s/it]
k = 8, silhouette score:0.06723287947531326
 57%|█████▋     | 8/14 [00:34<00:29,  4.99s/it]
k = 9, silhouette score:0.06935309284173061
 64%|██████▍    | 9/14 [00:40<00:26,  5.24s/it]
k = 10, silhouette score:0.06195688255808464
```

As an example for Kmeans, I checked silhouette score from k = 2 to 14. I choose k = 7 since the silhouette score is not low, and keep a sufficient number of k.

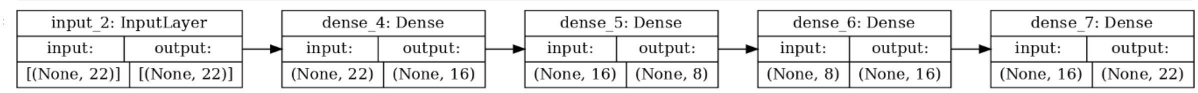
Similiarly, for Variational Bayesian estimation of a Gaussian mixture, I choose k = 5, silhouette score = 0.070

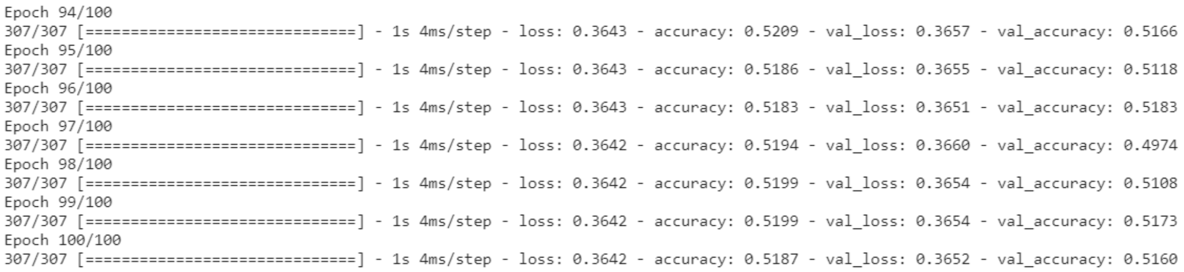
For DBSCAN, I choose eps=4, min_samples=3 based on different trials of silhouette score.

Auto encoder (optional)

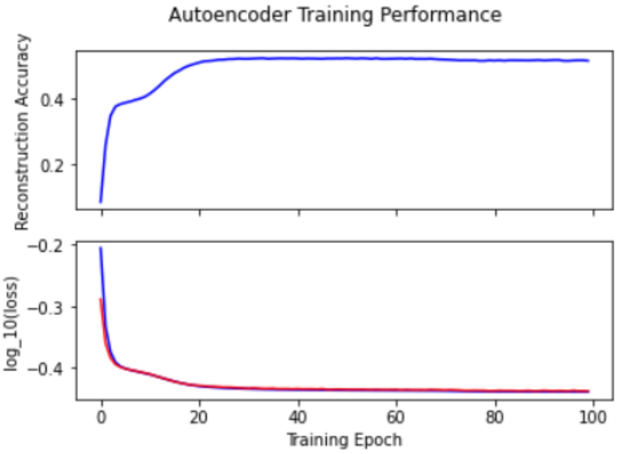


```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 22)]              0
_________________________________________________________________
dense_4 (Dense)              (None, 16)                368
_________________________________________________________________
dense_5 (Dense)              (None, 8)                 136
_________________________________________________________________
dense_6 (Dense)              (None, 16)                144
_________________________________________________________________
dense_7 (Dense)              (None, 22)                374
=================================================================
Total params: 1,022
Trainable params: 1,022
Non-trainable params: 0
_________________________________________________________________
```





```
Epoch 94/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3643 - accuracy: 0.5209 - val_loss: 0.3657 - val_accuracy: 0.5166
Epoch 95/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3643 - accuracy: 0.5186 - val_loss: 0.3655 - val_accuracy: 0.5118
Epoch 96/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3643 - accuracy: 0.5183 - val_loss: 0.3651 - val_accuracy: 0.5183
Epoch 97/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3642 - accuracy: 0.5194 - val_loss: 0.3660 - val_accuracy: 0.4974
Epoch 98/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3642 - accuracy: 0.5199 - val_loss: 0.3654 - val_accuracy: 0.5108
Epoch 99/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3642 - accuracy: 0.5199 - val_loss: 0.3654 - val_accuracy: 0.5173
Epoch 100/100
307/307 [==============================] - 1s 4ms/step - loss: 0.3642 - accuracy: 0.5187 - val_loss: 0.3652 - val_accuracy: 0.5160
```



The auto encoder's performance is not high enough to use this. The main reason is number of attributes not high enough. So this method we can choose to apply if there is a high number of attributes

Models

Model 1: Kmeans with k = 7: silhouette score = 0.080, Kaggle competition leader board: 0.238
Model 2: Variational Bayesian estimation of a Gaussian mixture, I choose k = 5, silhouette score = 0.070, Kaggle competition leader board: 0.249

Model 3: Variational Bayesian estimation of a Gaussian mixture, I choose k = 7, silhouette score = 0.064, Kaggle competition leader board: 0.2573

Findings

Silhouette score calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample, it shows the separability for each clustering algorithm, but it's unnecessary shows the same trend as the true grouping. The best method on the leader board is Variational Bayesian estimation of a Gaussian mixture with k = 7

Revisit

I can try more options such as DBSCAN and try more variables. Also need to find more meanings within the original data.

Appendix

Data: https://www.kaggle.com/competitions/tabular-playground-series-jul-2022/data

Code: https://www.kaggle.com/taos2000/clustering-pca-kmeans-bayes-auto-encoder