FINAL TERM PROJECT

서왕규

2014004066

CONTENTS

01. Procedure

- **Procedure**

1. Train/Test samples generation
2. Training : Find mean vectors by K-means clustering
3. Training : Choose maximum distance
4. Test

- **Assume**

Training data has **no class**(Unsupervised learning)

No information of **how many sample point** are in each class

The number of cluster is **5**

We can evaluate prediction of test data is corrected or not

- **Evaluation**

T = # of match that is predicted to the right class.

F = # of match that is predicted to the wrong class.

Accuracy = T / T + F

02. Generate train/test set

Train/Test samples generation

- Train set
  - Consisted of 5 classes
  - Randomly generate **300 sample points** per a class

- Test set
  - Consisted of 6 classes which 5 classes are same as train set and 1 class is not same.
  - Randomly generate **100 sample points** per a class.

Class	X	Y	Z
0	N(1,4)	N(1,1)	N(1,2.25)
1	N(5,6.25)	N(7,9)	N(-3,4)
2	N(-4,9)	N(-10,2.25)	N(5,1)
3	N(-3,7.29)	N(6,2.25)	N(-1,9)
4	N(9,1)	N(0,2.25)	N(0,4)

Fig 1. Table of Train set distribution

Class	X	Y	Z
5	N(5,2.25)	N(-4,6.25)	N(4,12.25)

Fig 2. Table of extra class for test

Training samples

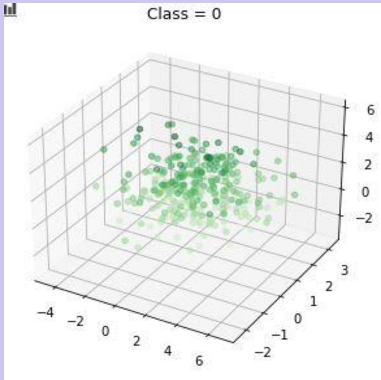


Fig 1. sample of class 1

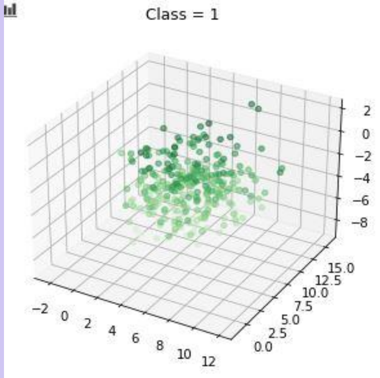


Fig 2. sample of class 2

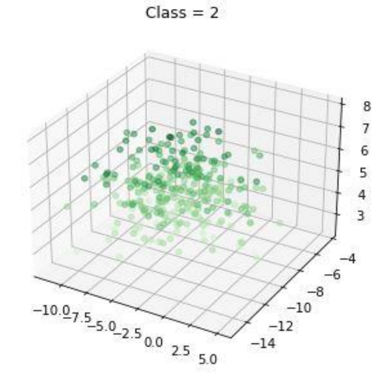


Fig 3. sample of class 3

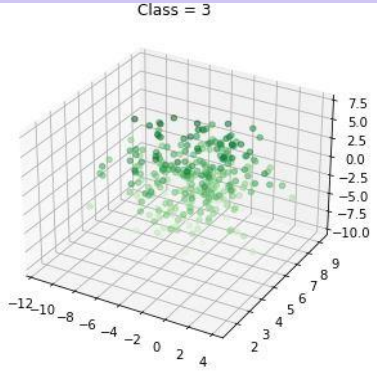


Fig 4. sample of class 4

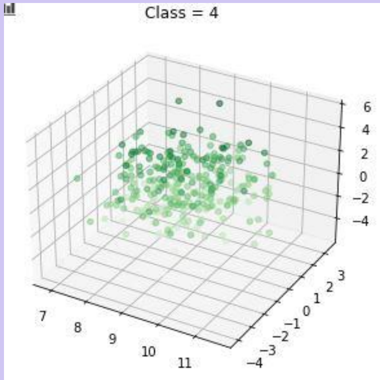


Fig 5. sample of class 5

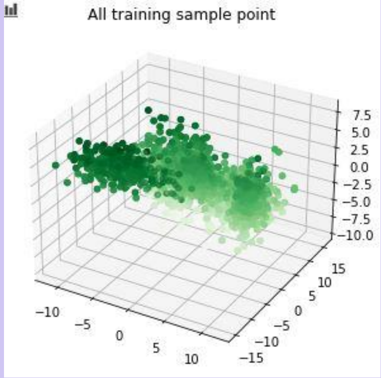


Fig 6. All training sample

Test samples

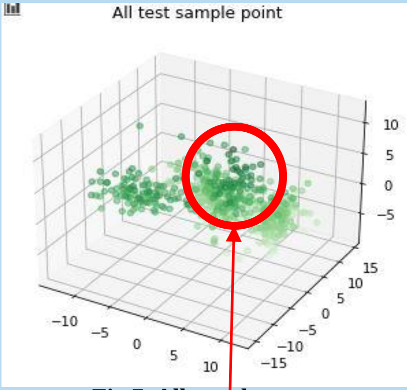


Fig 7. All test data set

Extra class

Training : Find mean vectors by K-means clustering

- K-means clustering(K=5)

I. Choose seed point randomly

II. Divide all point to a cluster which has **nearest** centroid from the point

III. Recalculate **centroid** of clusters

IV. Repeat 2,3 until **converge**

→ Find **mean vector** of 5 clusters

Mean vector	X	Y	Z
Class 0	1.0602	1.3509	0.9158
Class 1	5.3074	7.5098	-3.1818
Class 2	-3.8317	-9.8629	5.0319
Class 3	-3.5391	6.4290	-1.1532
Class 4	8.8335	0.1976	-0.3490

Fig 1. mean vector of cluster by K-means clustering

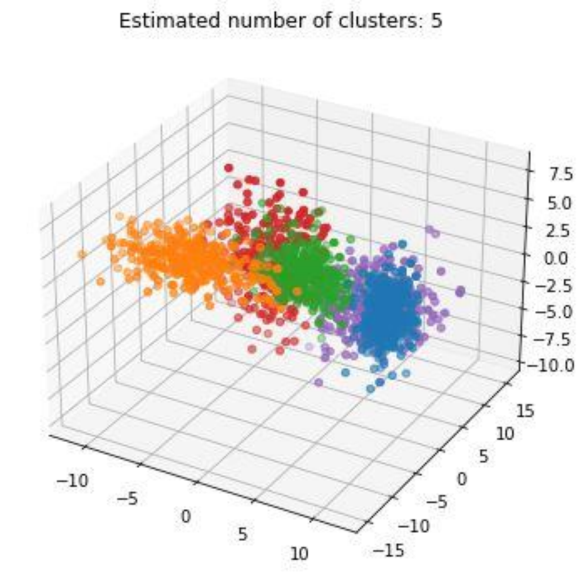


Fig 2. Result of K-means clustering

Training : Choose maximum distance

① Maximum likelihood for finding **covariance** of a cluster

✓ K-means clustering으로부터 나온 **centroid**와 그 **centroid를 포함하는 클러스터에 속하는 모든 점**들에 대한 가능성(likelihood)를 최대화하는 $\boldsymbol{\Sigma_z}$ 계산

$$Z_i = [x_i, y_i, z_i]_{i=1}^{N} \in Cluster(k)$$

$$P(Z|\Sigma_z) = \prod_{i=1}^{N} P(Z_i|\Sigma_z) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}^{3}|\Sigma_z|^{\frac{1}{2}}} exp(-\frac{1}{2}(Z_i - u_z)^T {\Sigma_z}^{-1}(Z_i - u_z))$$ **: likelihood**

$$\Sigma_z = argmax_{\Sigma_z} P(Z|\Sigma_z) = argmax_{\Sigma_z} \ln P(Z|\Sigma_z)$$

$\frac{\partial}{\partial \Sigma_z} \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}^{3}|\Sigma_z|^{\frac{1}{2}}} exp(-\frac{1}{2}(Z_i - u_z)^T {\Sigma_z}^{-1}(Z_i - u_z)) = 0$ **일 때의 Σ_z가 K-means 클러스터에 의해 나온 centroid와 해당 클러스터에 속하는 샘플포인트들의 likelihood function을 maximize한다.**

$\Sigma_z = \frac{1}{N}\sum_{i=1}^{N}(Z_i - u_z)^T {\Sigma_z}^{-1}(Z_i - u_z)$ **: empirical covariance**

✓ Gaussian distribution의 실제 확률 분포를 모를 경우, **Covariance matrix**를 샘플 데이터를 통해 구한 **empirical covariance Σ_z** 로 설정하면 maximum likelihood를 얻음

② From covariance matrix, find **maximum weighted distance**

- 각 클러스터는 maximum distance를 갖음
- 클러스터의 centriod와 test sample point의 Distance의 계산은 euclidean distance가 아닌, 클러스터 covariance matrix의 σ_x, σ_y, σ_z값에 따라 x, y, z에 weight를 부여하여 계산

$\boldsymbol{Distance(k(C_{kx}, C_{ky}, C_{kz}), P_i(x_i, y_i, z_i))}$ **: 클러스터 k와 i번째 sample의 거리**

$$\boldsymbol{Distance(k, P_i)} = \frac{|C_{kx}-x_i|^2}{\sigma_x{}^2} + \frac{|C_{ky}-y_i|^2}{\sigma_y{}^2} + \frac{|C_{kz}-z_i|^2}{\sigma_z{}^2}$$

- $\boldsymbol{Distance(k, P_i)}$ < c (c:constant)

타원체의 **내부의 점**일 경우에만 cluster에 속한 것으로 판정

	$\sigma_x{}^2$	$\sigma_y{}^2$	$\sigma_z{}^2$
Class0	3.6803	2.2208	2.9527
Class1	4.8532	6.1886	4.2016
Class2	8.6057	3.8973	1.0286
Class3	5.4869	2.2065	8.3600
Class4	1.7645	2.5653	4.1939

Fig 1. Empirical variance

C(constant)	# of Correct
9	545
10	**549**
11	545
12	543

Fig 2. Choose maximum distance

04. Result

Compare probability density function with Empirical variance and mean

- K-means clustering을 통해 클러스터의 **centroid(Empirical mean)**를 찾고, 샘플데이터에 label(class)를 할당(Unsupervised)
- Maximum likelihood 를 통해 각 클러스터의 **empirical covariance**를 계산하고, 클러스터의 x, y, z의 variance를 이용하여 euclidean distance 가 아닌 **weighted distance를 계산**

Class	X	Y	Z
0	N(1,4)	N(1,1)	N(1,2.25)
1	N(5,6.25)	N(7,9)	N(-3,4)
2	N(-4,9)	N(-10,2.25)	N(5,1)
3	N(-3,7.29)	N(6,2.25)	N(-1,9)
4	N(9,1)	N(0,2.25)	N(0,4)

Fig 1. Defined probability density function

Class	X	Y	Z
0	N(1.0602, 3.6803)	N(1.3509, 2.2208)	N(0.9158, 2.9527)
1	N(5.3074, 4.8532)	N(7.5098, 6.1886)	N(-3.1818 , 4.2016)
2	N(-3.8317 , 8.6057)	N(-9.8629 , 3.8973)	N(5.0319 , 1.0286)
3	N(-3.5391 , 5.4869)	N(6.4290 , 2.2065)	N(-1.1532 , 8.3600)
4	N(8.8335 , 1.7645)	N(0.1976 , 2.5653)	N(-0.3490 , 4.1939)

Fig 2. Empirical mean and variance from samples

❖ **Predict Condition**

➢ **Condition 1** : find the nearest cluster

Find nearest centroid of Cluster k

➢ **Condition 2** : check the distance

$$Distance(k, P_i) = \frac{|C_{kx}-x_i|^2}{\sigma_x^2} + \frac{|C_{ky}-y_i|^2}{\sigma_y^2} + \frac{|C_{kz}-z_i|^2}{\sigma_z^2} < 10$$

Data

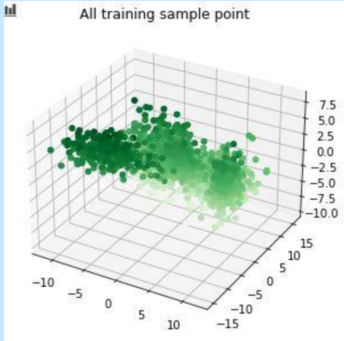


Fig 1. Unlabeled Train data

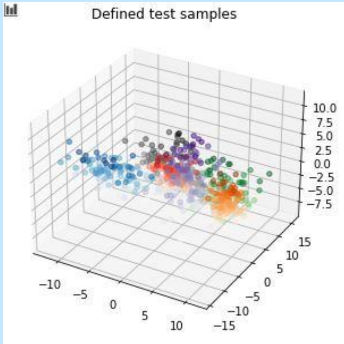


Fig 2. labeled Test data

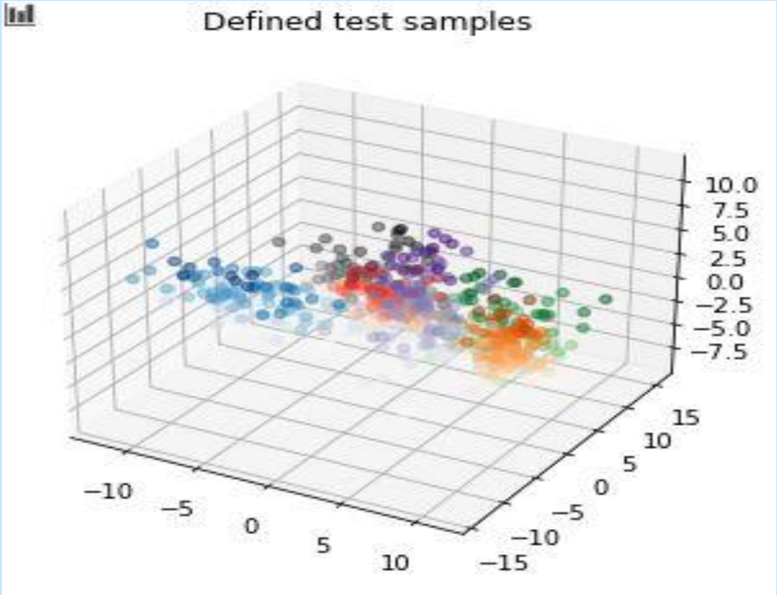


Fig 3. Defined Classes of test samples

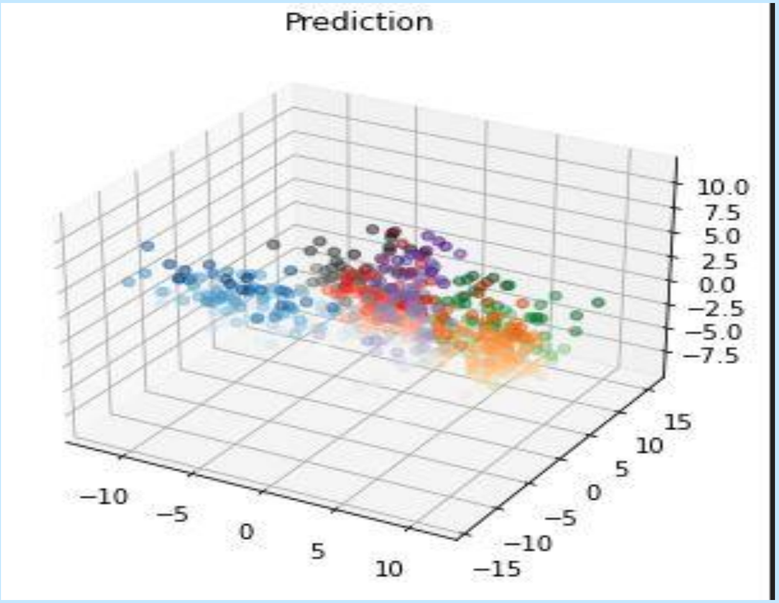


Fig 4. Result of prediction

04. Result

Class	# of correct	# of wrong	Accuracy
0	98	2	0.98
1	84	16	0.84
2	98	2	0.98
3	85	15	0.85
4	100	0	1
Extra	84	16	0.84
Total	549	51	0.915

Fig 1. Table of recognizing result

Cluster representation with Maximum distance

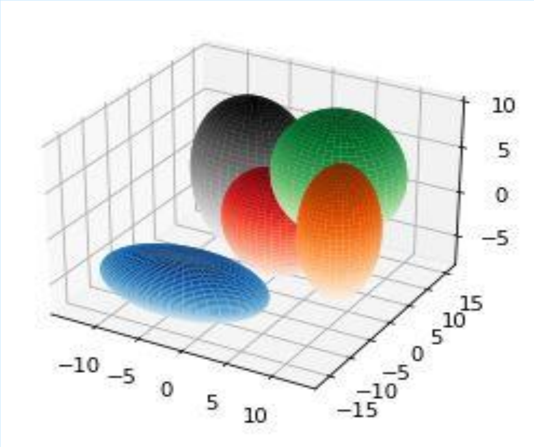


Fig 1. Vector space with maximum distance of each cluster

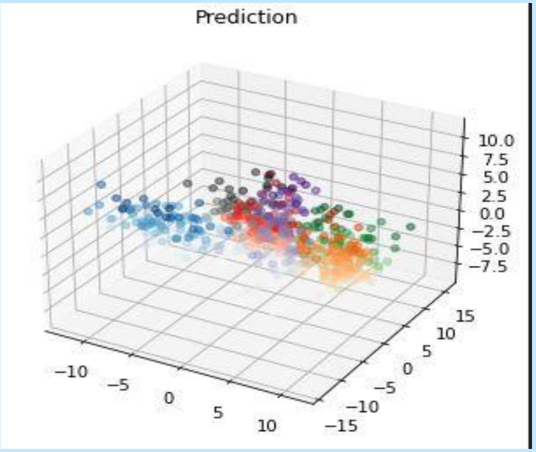
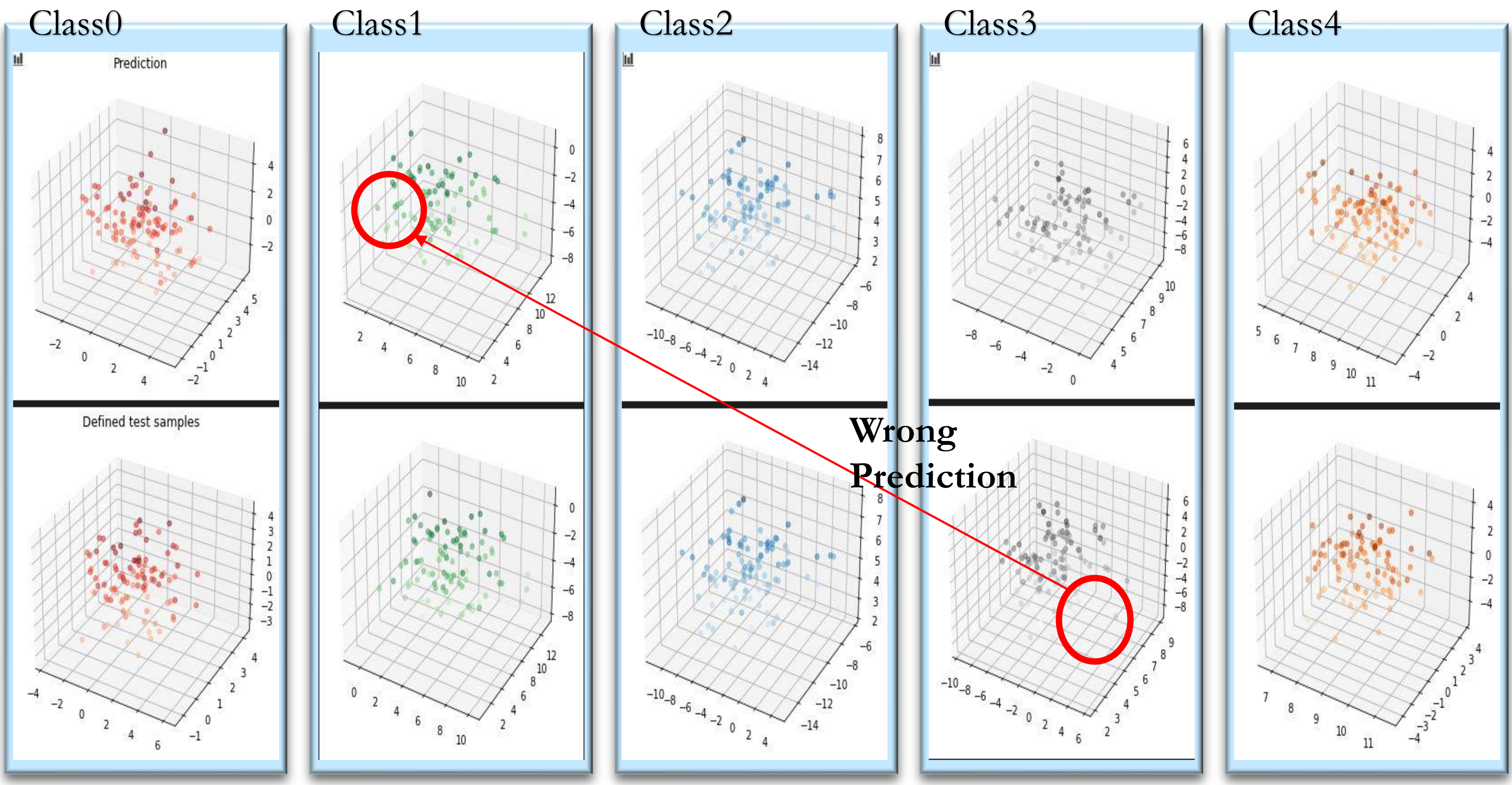


Fig 2. Result of prediction

- Training Sample point를 통해 구한 covariance matrix를 이용하여 축이 $\sqrt{10}\sigma_x, \sqrt{10}\sigma_y, \sqrt{10}\sigma_z$ 길이의 Recognize boundary 생성
- 해당 영역 밖의 점은 Figure2의 **보라색(Extra class)**로 판별



- Upper figure : Prediction
- Under figure : predefined classes

Conclusion

- 두 정규분포의 **Mean vector 차이는 작고 Variance가 큰 경우,** 잘못된 예측을 하기 쉽다.

Class	X	Y	Z
1	N(5,6.25)	N(7,9)	N(-3,4)
3	N(-3,7.29)	N(6,2.25)	N(-1,9)

Class	# of correct	# of wrong	Accuracy
1	84	16	0.84
3	85	15	0.85

- Labeling 되어 있지 않는 데이터셋으로 **K-means clustering**과 Maximum likelihood를 활용하여 **recognize**을 위한 클러스터 영역을 생성할 수 있다.