Bonus report written by m5261108 Kazuki Fujita

1. Contents

In previous question: classification of Iris data using Kohonen net, we could not classify them with 100% accuracy because the data values are not in the separation hyperplane. In bonus contents, we categorized them in a different way. We used Gaussian Mixture Model as another method.

2. Gaussian Mixture Model

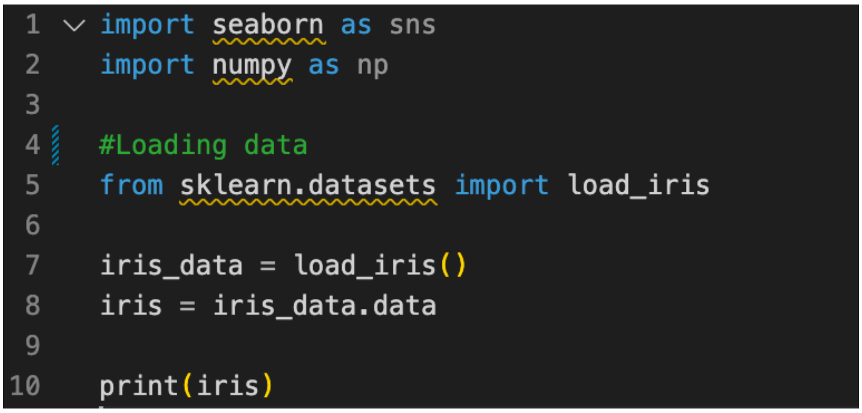
Gaussian mixture model (GMM) is one of the algorithms used for clustering. It represents a given data set as a superposition of multiple normal distributions. It yields the probability of belonging to each cluster for each sample.

3. Implementation

Our program with GMM was created in python (version: 3.7.9). The program created this time can be referred to at the following URL: https://github.com/K-Fujita-onkyo/neural_network/blob/main/Project4/iris.py

3.1. Loading Data

We used load_iris() in sklearn's datasets module to retrieve Iris data. In this case, we only need the Iris data, so we assigned only the data part of Iris_data to iris. This iris data is the same as the iris data in Project 4.

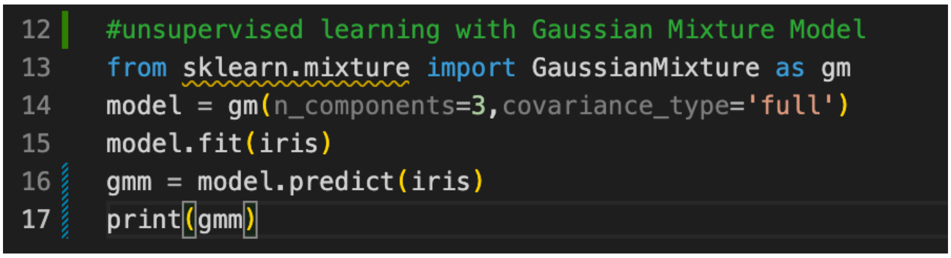


```
import seaborn as sns
import numpy as np

#Loading data
from sklearn.datasets import load_iris

iris_data = load_iris()
iris = iris_data.data

print(iris)
```

Picture1: Loading data in iris.py

3.2. Unsupervised learning with GMM

In this case, we used GaussianMixture in sklearn's mixture module for the calculation. There are two arguments, n_components is the number of components and covariance_type is the type of covariance. We set n_components as 3 and covariance_type as full. The training method is fit(T_data). T_data is the data to be trained. Data classification is done by predict().

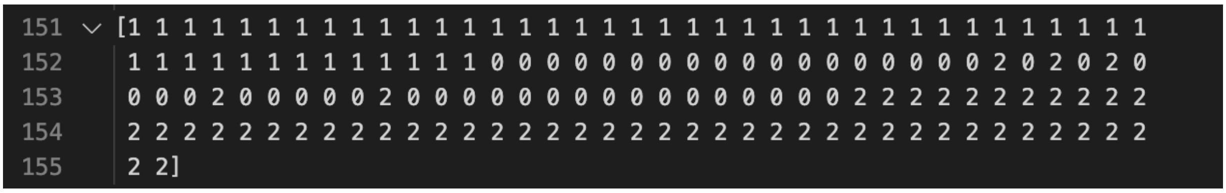


```
#unsupervised learning with Gaussian Mixture Model
from sklearn.mixture import GaussianMixture as gm
model = gm(n_components=3,covariance_type='full')
model.fit(iris)
gmm = model.predict(iris)
print(gmm)
```

Picture2: Loading data in iris.py

4. Evaluation method and Result

We verified that the Iris data were accurately classified. The results were as follows. There were five Iris-versicolor mistaken for Iris-virginica. It was found that the classification was considerably more accurate than a winner-take-all learning with Kohonen net.



```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 0 2 0
 0 0 0 2 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
```

Picture3: The result with GMM

In Table1, We compared Kononen and GMM results. Both algorithm accurately classified setosa. Kohonen classified versicolor more accurately than GMM. GMM accurately classified virginica, while Kohonen was only about 70% accurate. Therefore, we have found that GMM can classify Iris more accurately than Kohonen.

Algorithm	Species	Number of data	Corrects	Incorrects		
				Iris-setosa	Iris-versicolor	Iris-virginica
Winner-take-all learning with Kohonen net	Iris-setosa	50	50	-	0	0
	Iris-versicolor	50	48	0	-	2
	Iris-virginica	50	34	0	16	-
Gaussian Mixture Model	Iris-setosa	50	50	-	0	0
	Iris-versicolor	50	45	0	-	5
	Iris-virginica	50	50	0	0	-

Table1: Comparison of Kohonen and GMM

5. Discussion

We discussed the followings:

- Why did GMM classify more accurately than Kohonen?

5.1. Why did GMM classify more accurately than Kohonen?

This is because GMM classifies using probability distributions. Kohonen is that it is a *hard clustering method*. A limitation to Kohonen approach is that there is no uncertainty measure or probability that tells us how much a data point is associated with a specific cluster. However, GMM is a soft clustering. GMM gives the probability of belonging to each cluster for each sample.

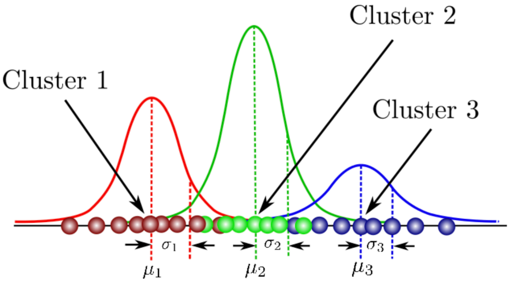


Figure 1: Gaussian function with K=3

A Gaussian Mixture is a function that is comprised of several Gaussians, each identified by k ∈ {1,..., K}. K is the number of clusters of our dataset. In Figure 1, we can see that there are three Gaussian functions, hence $K = 3$. Each Gaussian explains the data contained in each of the three clusters available.

The Gaussian density function is given by:

$$N(x\,|\,\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\,|\,\Sigma\,|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu))$$

Where x represents our data points, D is the number of dimensions of each data point. μ and Σ are the mean and covariance, respectively. This function corresponds to the maximum likelihood estimate method. This method is used to find the most applicable clusters. Therefore, Kohonen cannot probabilistically find clusters, whereas GMM can probabilistically find clusters. This difference led to this accuracy result.

6. Conclusion

We used Gaussian Mixture Model instead of Kohonen to classify Iris. The method results showed that GMM classifies more accurately than Kohonen. In discussion, we found that even when there are outliers, they are classified with considerable accuracy by probabilistic determination.

7. Reference

[1] Bishop, Christopher M. *Pattern Recognition and Machine Learning* (2006) Springer-Verlag Berlin, Heidelberg.
[2] Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective* (2012) MIT Press, Cambridge, Mass.