**ECE 8527: Introduction to  
Machine Learning and Pattern Recognition**

# HW No. 4: CLASS-DEPENDENT VS. CLASS INDEPENDENT GAUSSIAN MODELS

For this assignment, you will generate two multivariate 2D Gaussian distributions. The first class will have a mean of and a covariance matrix of . The second class will have a mean of and a covariance matrix of . Generate 10,000 training data points (2D vectors) and 10,000 evaluation data points for each class (the total amount of training data is 20,000 points and the total amount of evaluation data is 20,000 points).

The tasks to be accomplished in this homework assignment are:

1. Plot a scatter plot of the data showing the first class in blue and the second class in red.
2. Pool both classes in the training set into one set of training data. Estimate the mean and covariance. Generate a transformation matrix using Principal Components Analysis (PCA). Transform the data using this matrix and plot a scatter plot. What do you observe?
3. Classify the data using a simple classifier based on the halfway point between the two means. What is the error rate? Draw the decision surface on the scatter plot.
4. Next, do a traditional class-dependent analysis and classify the data using a maximum likelihood classifier, as you did in HW #3. Assume the priors are equal and make sure you plug in each class’ covariance matrix. What is the error rate?
5. For step 4, draw a scatter plot of the transformed data and show the decision surface. Compare and contrast this to the findings of steps 1-3.

What did you learn from this analysis? If you rotated the scatter plots by a 45° angle would the results change? Can you prove this?

# Exercise 1 & 2

With provided mean and covariance matrix for desired data set, numpy subfunction of multi-variant Gaussian is used to generated desired data sets. The creation of the data set is controlled under the same class as the PCA form of the data set. Two scatter plots are generated with the original data set and the PCA form of the data set. The following two diagram shows the two generated data sets.

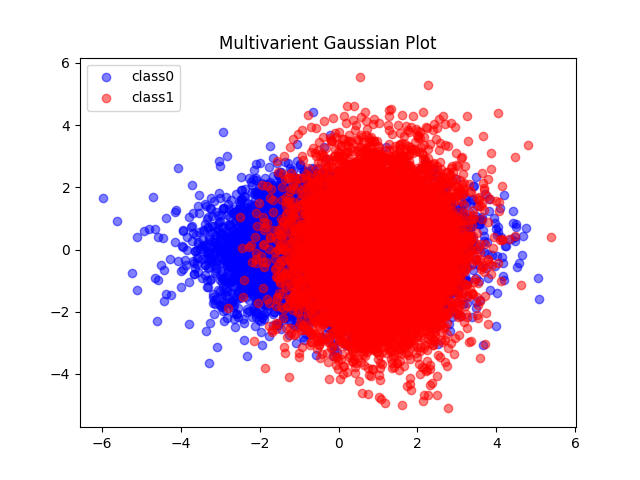


Figure Multi-Variant Gaussian Plot

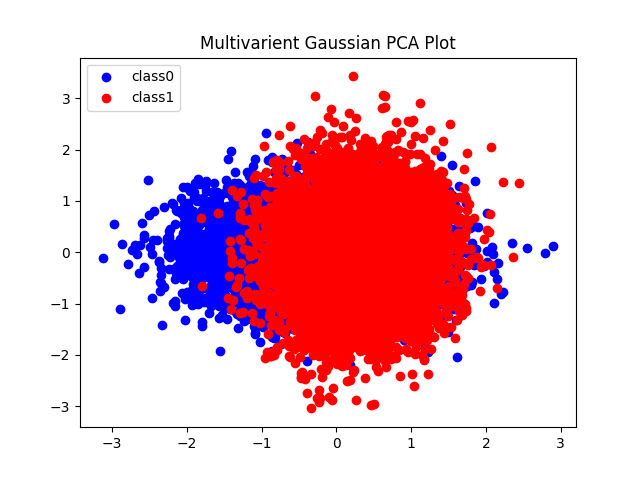


Figure Multi-Variant Gaussian PCA Plot

By comparing these two scatter plots, it shows that the original data set has a larger range of the scatter region while the PCA version of the data has a smaller scatter region. The PCA version of the scatter plot also shows that it is a little titled comparing to the original plot.

# Exercise 3

In this exercise, the plot is generated based on the theoretical error rate calculation that takes in the “distance” between the data point with each class’ mean value. This error rate for the evaluation data set shows as 33.33%. The decision surface for this classification is also drawn via computing an array of data points and selecting points where it shows as 0 by using the contour function. The following diagram shows the result of the contour plot for the evaluation data set.



Figure PCA Classified Data

Since the method of calculating the decision surface is based on the distance in between the plot and the mean of each class, the decision surface is shown as a straight line in between these two mean values.

# Exercise 4

In this exercise, the data is pooled into the maximum likelihood classifier and then classified from the train data set and ran with the evaluation data set. The classifier gives out a result of 31.44% with equal prior. The classifier was constructed and used in the previous homework but modified so that all construction is done in the class instead of sub-functions outside of the class with individual output. The calculation for the error rate is taken via log probability and then computed for the result of the class.

# Exercise 5

The decision surface of the maximum likelihood classifier is also plotted together with the scatter plot of the evaluation data set. The decision surface is calculated by finding the contour level of 0 when the difference in between two classes are zero. The following diagram shows the decision surface plotted together with the scatter plot of the evaluation data set.



Figure 4 Maximum Likelihood Classified Data

By comparing this decision surface towards the PCA data decision surface, the main difference in between these two is that the PCA data’s decision surface is a straight line while the Maximum Likelihood Classified data shows as a curve. One of the significant outcomes of this is that the ML Classifier shows a better error rate when comparing to the PCA one.

If the data is rotated by 45 degrees, the result should stay the same. A rotation of 45 degree is a linear transformation, thus it applies to “everything” in the data set. Hypothetically, when the data is rotated by 45 degrees, the axis can also be considered as being applied a rotation of 45 degrees. The relative data result would show the same relatively to the new axis, thus a data point is still being classified into its “old” class. As for how to show this in calculation, one method is to apply the rotational matrix to the Gaussian data and run the new data through the classification process.