MINI-PROJECT 1  
ECE/CS 7720 MACHINE LEARNING AND PATTERN RECOGNITION

Junyoung Shin

1. {'mean': array([2.00167345, 2.99978405, 1.01581082, 5.02328933, 8.00876704]),

'covariance':

array([[ 9.8012580e-01, 4.4428000e-03, -9.0496000e-03, 4.2132000e-03, -1.7546900e-02],

[ 4.4428000e-03, 5.0562140e-01, -8.2040000e-04, 3.0652000e-03, 1.3997700e-02],

[-9.0496000e-03, -8.2040000e-04, 2.5488618e+00, 1.0818200e-02, 1.6520500e-02], [ 4.2132000e-03, 3.0652000e-03, 1.0818200e-02, 7.1529040e-01, -2.6669000e-03],

[-1.7546900e-02, 1.3997700e-02, 1.6520500e-02, -2.6669000e-03, 3.4940553e+00]])

I was able to obtain a rather similar value between manually calculating covariance and using the cov function. However, if we're not careful when calculating the covariance, our estimate might be too low. This is because when we divide by the sample size n-1 instead of n we're dividing by a smaller number which can exaggerate differences. Furthermore, . The difference between the two results become more pronounced for smaller samples. For larger samples, like 10000, the difference between dividing by n and n-1 is minimal, so the covariance estimates don't vary much. However, for smaller samples, like 50, this difference can lead to significant disparities in the covariance estimates.

1. (eigenvalues=array([3.49453594, 2.5486865 , 0.98006515, 0.50546825, 0.71519887]), eigenvectors=array([[-7.03352407e-03, -5.55756252e-03, -9.99782971e-01,

-9.39645460e-03, 1.62898985e-02],

[ 4.66613456e-03, -5.24939904e-04, -9.66078024e-03, 9.99839268e-01, -1.43547420e-02], [ 1.75192386e-02, 9.99812917e-01, -5.58801340e-03, 4.75022489e-04, 5.97963628e-03], [-8.96717817e-04, 5.91141091e-03, -1.61795474e-02, -1.45024472e-02, -9.99746044e-01], [ 9.99810496e-01, -1.75506259e-02, -6.90483833e-03, -4.75370223e-03, -8.19847578e-04]])

1. Choose 2 features pairwise : In order to plot the data in a 2D graph, I chose two features pairwise. Each color of the arrow indicates class1-4. (["darkred", "darkblue", "green", "purple"])

A graph with a colorful arrow

Description automatically generatedA graph with a colorful arrow

Description automatically generated with medium confidenceA graph with a colorful arrow

Description automatically generatedA graph with a colorful arrow

Description automatically generatedA graph with a red arrow pointing to the right

Description automatically generatedA graph with a colorful arrow

Description automatically generated with medium confidenceA graph with arrows pointing to different colors

Description automatically generatedA graph with a colorful arrow

Description automatically generated with medium confidenceA graph with a colorful triangle

Description automatically generated with medium confidenceA graph with a colorful line

Description automatically generated with medium confidenceA graph with a purple arrow

Description automatically generatedA graph with a colorful arrow

Description automatically generated with medium confidenceA graph with a number of colored triangles

Description automatically generated with medium confidenceA graph with arrows pointing to different directions

Description automatically generatedA graph with arrows pointing to different colors

Description automatically generatedA graph with arrows pointing to different directions

Description automatically generatedA graph with arrows pointing to the same direction

Description automatically generatedA graph with arrows pointing to the same direction

Description automatically generatedA graph with colorful arrows

Description automatically generatedA graph with a purple arrow

Description automatically generated

Fig. 1 : Visualizing features pairwise along with PCA

Since each dataset has a large portion of overlap between others, the scatter plot shows that they are difficult to separate linearly.

A yellow and blue dot

Description automatically generatedA yellow circle with blue dots

Description automatically generatedA yellow and blue dot

Description automatically generatedA yellow and blue dot diagram

Description automatically generated with medium confidenceA yellow and blue dot

Description automatically generatedA yellow and blue dot

Description automatically generatedA yellow and blue dot

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow and blue dot

Description automatically generatedA yellow and blue dots

Description automatically generatedA yellow circle with blue dots

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow and blue dot

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow and blue dot

Description automatically generatedA yellow circle with blue and green dots

Description automatically generatedA yellow circle with blue and green dots

Description automatically generated

Fig 2. Visualizing raw data pairwise along 5 different features

d-1) Euclidian Results

[[2.64611741 2.20771217 2.34467901 2.32012411] --- class 2

[3.55237915 3.86145832 3.75682108 3.72628583] --- class 1

[2.22643973 2.07990511 2.08309569 2.14207158] --- class 2

[2.46389511 2.21738798 2.28695745 2.2811197 ]] --- class 2

d-2) Mahalanobis Results

[[2.01484404 1.61857867 1.6849854 1.70594432] --- class 2

[2.65473063 2.45215169 2.46515197 2.38294785] --- class 4

[2.46437299 2.21699309 2.1997465 2.17972384] --- class 4

[2.62535835 2.35088984 2.58182838 2.48620851]] --- class 2

d-3) classifying each datasample

Explained in (d-1) and (d-2). While Euclidean distance treats features as independent and equally weighted, Mahalanobis distance adjusts for the covariance structure of the data by incorporating the covariance matrix. This adjustment allows Mahalanobis distance to provide a more accurate measure of distance, particularly when dealing with correlated features or differences in variances. This means that although a certain data might seem to be closer to class 1 in Euclidian, when converted to Mahalonobis distance, the factors defining the distance between the class and the data itself has shifted, therefore showing a difference in the results.