Project 2 - Fisher's Linear Discriminant Analysis

Linear Algebra and Learning from Data

1. Fisher's LDA via a toy problem

Task 1: Show that an optimal solution is given by

$$v = (\Sigma_A + \Sigma_B)^{-1}(m_A - m_B).$$

Answer: According to the Fisher's linear discriminant problem, we are looking for the best separation vector v for the classes A and B. However, the separation ratio can be written as a generalized Rayleigh quotient with:

$$S = (m_A - m_B)(m_A - m_B)^T \ and \ M = (s_A + s_B)$$

By writing the problem this way, the vector v is also an eigenvector associated with the largest eigenvalue of $M^{-1}S$ so we have:

$$M^{-1}Sv = \lambda v$$

In our case:

$$M^{-1}(m_A - m_B)(m_A - m_B)^T v = \lambda v$$

As the product $(m_A - m_B)^T v$ is a scalar, v has to be proportional to $M^{-1}(m_A - m_B)$.

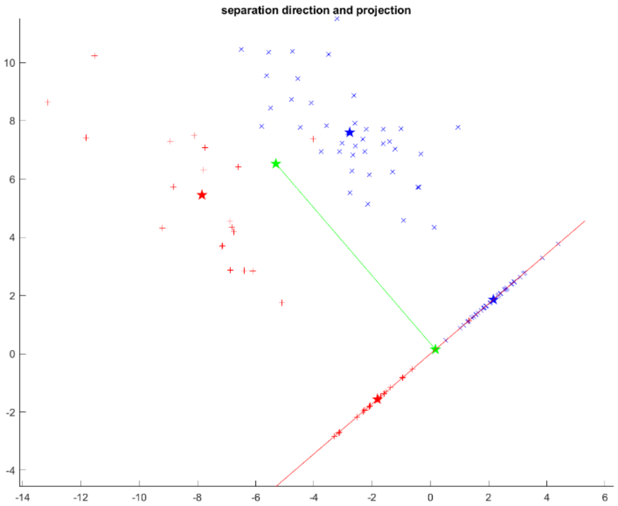


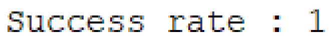
Figure 1. Projection with the separation vector v direction (green)

Classification: Using the computed separation vector from above, we can classify a sample vector x based on its projection. Specifically, if then we can guess x belongs to class A, otherwise, to class B (*can you explain why?*).

Answer: By being projected on v, data only have a parameter remaining: the position on the line directed by v. As this projection is theorically the best projection to separate the data, we can suppose that there exists a c such that the proposition is verified. However, the data can be to sparse or mixed for having a great separation by projection (red cross classified as blue).

Task 2: Use the separation vector to build a classifer. Test and report the success rate of your classifier on the testing dataset. Note that the success rate = 1 - missrate, where missrate = (missA + missB)/(sizeofTestA + sizeofTestB)

Answer: The Fisher's LDA method seems to work well on the toy dataset as we obtain a success rate of 1 on the test sets.

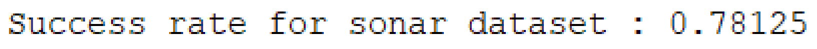


```
Success rate : 1
```

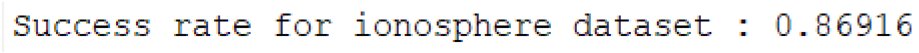
2. UCI Benchmark Problems

Task 3: Use the first 70% of data points as training data, and the rest as testing data. Test and report the success rate for classification by LDA, using the default threshold.

Answer: The model seems to be slightly more adapted for the ionosphere dataset. However, we can consider both success rates as good results.



```
Success rate for sonar dataset : 0.78125
```



```
Success rate for ionosphere dataset : 0.86916
```

Task 4: Try different threshold c, with value between $v^T m_A$ and $v^T m_B$. Observe how sensitive is the classification rate to the choice of c.

Answer: To analyze the best choice of c we have plotted the variation of the success rate according to the value of c between $v^T m_A$ and $v^T m_B$ for each dataset. For both, the best cs are near the center of the segment as chosen by intuition in the first part. However, the fluctuation of the success is very sensitive and two near cs can produce distant success rates.

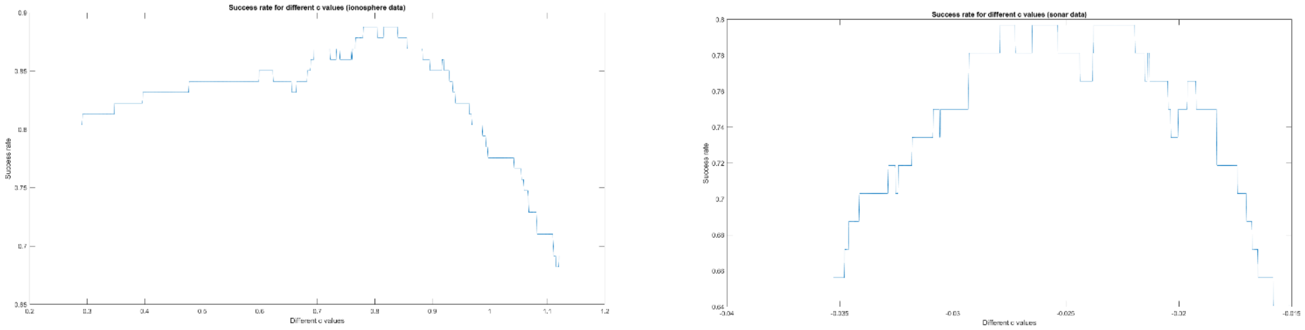


Figure 2. Variation of success rates according to the value of c

Task 5: For the ionosphere, the matrix $S_A + S_B$ become singular. Can you figure out why? How to address this issue? In general, when we encounter near singular matrices, what can we do?

Answer: For ionosphere data, $S_A + S_B$ become singular due to the high number of zero coefficient. According to matlab documentation, the use of least square method is recommended for sparse matrix like $S_A + S_B$ in our case so that's what have been used in task 3. With the same method as in sonar study we obtain a warning as we try to invert a singular matrix. To address this issue, we could also have tried to keep the same method and use the pseudoinverse with the *pinv* function. Principal components analysis could also help to reduce the matrix dimensions and at the same time the number of 0 in $S_A + S_B$.