Worksheet 6: Linear Discriminant Analysis

Statistical Analysis and Document Mining

MSIAM 1st year / ENSIMAG 2nd year

In this tutorial, we analyse the behaviour of the linear discriminant analysis when the number of features d is large. As a rule, it is generally assumed that the number of training examples n is abundant. However, in many real-life applications, like medical data analysis or bioinformatics, this assumption does not hold. Therefore, we consider further the case when $n \not\gg d$.

1. Using properties of the matrix rank, show that the sample covariance matrix $\hat{\Sigma}$ is singular when $n < d$. Can we still apply the linear discriminant analysis (LDA) in this case?

2. Now, let's consider the case when n is slightly greater or equal d. Would be enough samples to estimate the covariance matrix?

3. Give the definition of the condition number of a normal matrix in the Euclidean space. What is the ill-conditioned matrix? What happened when we compute the inverse of an ill-conditioned matrix?

4. What is the spectral decomposition (vectorial form) of the covariance matrix's inverse? Using this decomposition, re-write the LDA discriminant function for class c.

5. It's known fact that when $n \approx d$, the sample covariance matrix is often ill-conditioned. Using the answer of the previous question, how this fact would impact on the LDA's decision rule?

6. In practice, one way to overcome the problems arose in questions 1-5 is to reduce dimension. How we can do it using principal component analysis?

7. Another solution would be the Friedman's regularization. The idea is to increase diagonal elements of the covariance matrix by a constant value: $\hat{\Sigma}^F := \hat{\Sigma} + \alpha \mathtt{I}$. How it helps to invert the matrix? (Use the property of eigenvalues of the matrix $\hat{\Sigma}^F$).

8. Next week, we have a labwork. For this, please install the following R packages: `caret`, `mlbench`.

References

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175.