Linear Discriminant Analysis*

Aatmun Baxi, Andy Chen, Johnny Mo, Abhi Vemulapati

I. Objective

Supervised classification of datasets has proven to be one of the most versatile types of problems encountered in the study of machine learning. While the most prominent solutions to supervised classification, such as logistic regression, can often tackle a diverse array of problems, certain methods can more accurately classify datasets that satisfy certain assumptions. The method of linear discriminant analysis (LDA) is one such example [5]. LDA provides a fix to the most significant shortcoming of logistic regression, which is its failure to find stable solutions to binary classification problems when classes are well-separated [2]. As a binary classifier, LDA assumes that input features are normally distributed in order to find a closed-form solution to the optimal separating hyperplane between the two classes. In the multiclassification problem, LDA uses diagonalization techniques in order to detect the features of an input dataset that minimize covariance among features within a class and maximize covariance among features between classes. The latter technique can also be used as a method of dimensionality reduction.

Understanding LDA requires understanding of multivariate Gaussian distributions, Bayesian probability, and numerical linear algebra. Our research will make use of the standard machine learning reference [2], some statistical references [5], [7], and a reference for relevant topics in advanced numerical linear algebra [9].

II. Steps

We aim to complete the following tasks in this project:

A. Mathematical Understanding

Our study of LDA will begin with an analysis of the underlying mathematical foundations of both the binary and multiple class techniques. This will include how LDA accomplishes dimension reduction while retaining discriminative information in the data, which will require understanding of the composition of the within-class scatter matrices and between-class scatter matrices labelled S_w and S_b respectively. We may also discuss Fisher's discriminant [4], which computes decision boundaries in a similar way but relaxes some assumptions about the within-class covariances. We plan to reference our statistical learning references ([5] and [7]) during our examination of the math involved in LDA while also seeking out other various online resources that may explain the functionality of the LDA technique at a more computational level.

B. Applications

To demonstrate the abilities of LDA as both a dimension reduction technique and a classification model, we will present results for LDA applied on at least two data sets: these will include a high-attribute binary classification problem and a low-attribute multiclass classification problem. Tentatively, the data we have chosen for the former is a set of chest X-ray images classified by pneumonia positive patients and pneumonia negative patients from Kermany et al [3]. For the latter, we are considering applying LDA to one of several options, including Fashion MNIST [6], classification of star types [1], and classification of flower species using images [8].

III. Deliverables

We will present a report covering, in detail, the mathematical background of LDA, the LDA results on our example datasets, and an annotated Python implementation of the LDA model on our aforementioned applications. Analysis of our results will include considerations of the advantages and disadvantages of the LDA algorithm in our chosen applications and comparisons with other classification methods. The results, full analysis of the results, and a high level overview of the LDA method will be communicated in a slide presentation. The totality of the project, including the proposal, report, Python notebook, and slides will be made available via a Github repository under a permissive license.

References

[1] D. Baidya. Star Dataset to Predict Star Types. https://www.kaggle.com/deepu1109/star-dataset, 2019. Accessed: 10/29/20.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] D. S. Kermany et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131, Feb 2018.

[4] R.A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[5] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1(10). Springer series in statistics New York, 2013.

[6] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, 2017.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[8] A. Mamaev. Flowers Recognition. https://www.kaggle.com/alxmamaev/flowers-recognition. Accessed: 2020-10-29.

[9] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, volume 50. Siam, 1997.

*This work was not supported by any organization