Bowen Dong

Zhaoxing Wu

STAT 499

Winter 2024

Classifying high-dimensional data In Higher Accuracy

In winter quarter DRP learning, I tried to explore the classification problems posed by data with large features (P) and small samples (N) and its possible solutions. My mentor and I ended up focusing on Penalized Discriminant Analysis (PDA). PDA is a variant of Fisher's Linear Discriminant Analysis (LDA) that improve the classification of high-dimensional data by incorporating a penalty term (λ) to prevent overfitting. When I joined this project, I did not have any research or experience in computer algorithms. The PDA and LDA were very unfamiliar to me. But through this experience, I learned about these two different algorithms and how to classify the higher dimensional data.

In the early half of the quarter, I mainly understood the binary classification, the curse of dimensionality and projection pursuit. I read the first book called *R in Action* (Kabacoff, 2022) which emphasized that the results of binary classification are predicted by a set of variables or features, such as the ability to repay a loan, medical condition, or spam. Mostly through machine learning and then verifying its accuracy on another dataset. Then the article includes a couple of important parts which are classifying with decision trees, and random forest classifier, support vector machines to assess the accuracy and understanding of complex models. Then, I read the second book called *Applied multivariate statistical* (Johnson, 1992), which emphasized curse of dimensionality and projection pursuit. Curse of Dimensionality is a phenomenon where traditional classification methods falter as the dimensionality of the dataset increases. This challenge is important to our project, leading our exploration to use advanced techniques to maintain, or even enhance, classification accuracy in such high-dimensional data. To solve the curse of dimensionality, the book gives a way called projection

pursuit which is a statistical technique aiming to discover meaningful low-dimensional representations of high-dimensional data by seeking projections that optimize specific criteria, such as separability or preservation.

In the latter half of the quarter, The main book I read is *A Projection Pursuit Index for Large p Small n Data* (Lee, Eun-Kyung, and Dianne Cook. 2010). The author's motivation for writing this article is to improve the shortcomings of Fisher LDA, Fisher LDA can better distinguish between large-sample and less-features data, but in less-sample and more- features data, Fisher LDA is prone to overfitting preventing us from getting a more accurate categorization. The article proposes a method called Penalized Discriminant Analysis (PDA) is a technique used to improve the classification of high-dimensional data by incorporating a penalty term to prevent overfitting. Penalty term called λ works by "shrinking" the importance of less essential predictors, This means it effectively reduces the influence of predictors that don't help much with classification. λ corrects for between class and within class to ensure that we can find the features that are critical to the problem, and thus ignore the features that Not-so-important. And PDA Projection Pursuit index measures the separability of the classes after projection: a value close to 1 indicates good class separability, while a value close to 0 indicates poor separability.

Finally I applied the R package inside the article called ClassPP on my simulated data, there are 1000 features in my data but only 10 examples, it are divided into two classes. After testing in the Rstudio, I found that PDA method can plot more detailed 1D or 2D space compared to Fisher LDA. this is mainly dependent on the regularization of λ.

In concusion, this quarter's DRP was a great learning opportunity. I am very happy that I had the opportunity to participate in it, and I am very thankful to my mentor, Zhaoxing Wu, who was very patient and willing to put in the extra work hours to meet with me, and who was able to correct me and give me the right advice when I didn't know what I was doing wrong. My mentor was very helpful in completing this quarterly DRP. However, I don't think I studied deeply enough and didn't apply these method in real data. I will apply these two algorithms to real data and understand them thoroughly in my subsequent studies.

Sources:

Lee, Eun-Kyung, and Dianne Cook. “A Projection Pursuit Index for Large p Small n Data.” *Statistics and Computing*, vol. 20, no. 3, 2010, pp. 381–92, https://doi.org/10.1007/s11222-009-9131-1.

Johnson. (1992). *Applied multivariate statistical* ..

Kabacoff, R. (2022). *R in action*. Manning Publications.