Project Title: Directed Reading Project: Classify High-Dimensional Data

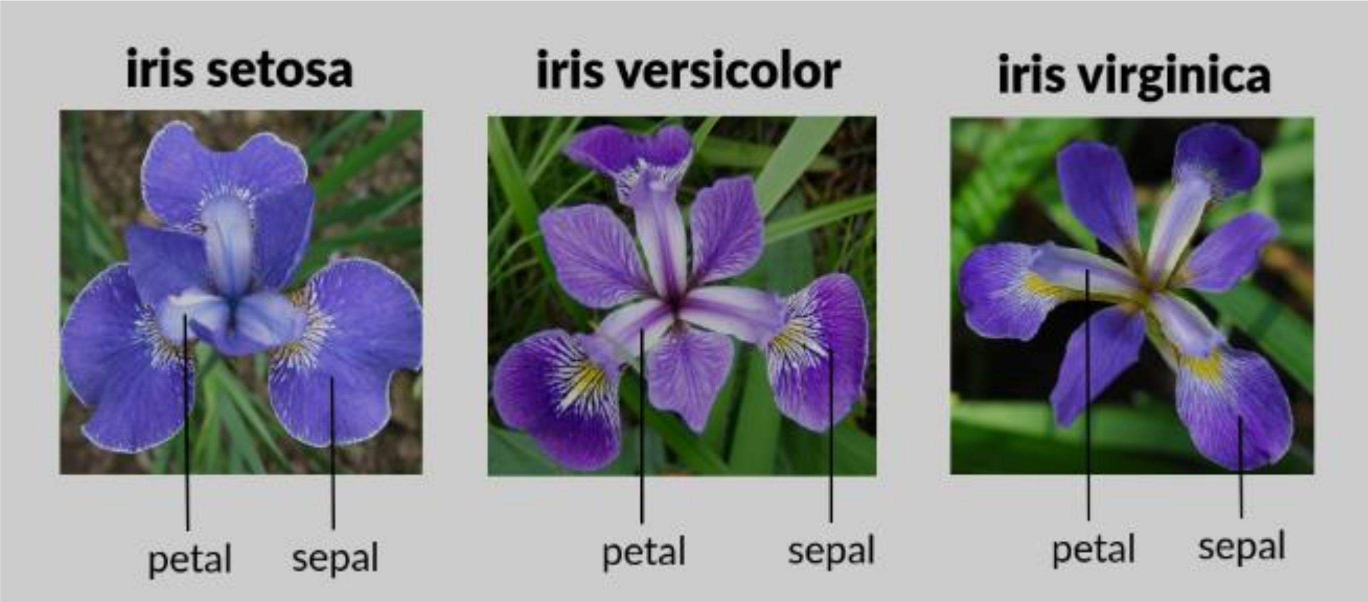
Graduate Student Mentor: Zhaoxing Wu

Undergraduate Student Mentee: Bowen Dong

DRP Winter 2024

Classification

Classification is the process of categorizing or organizing items into predefined categories based on their features or characteristics.

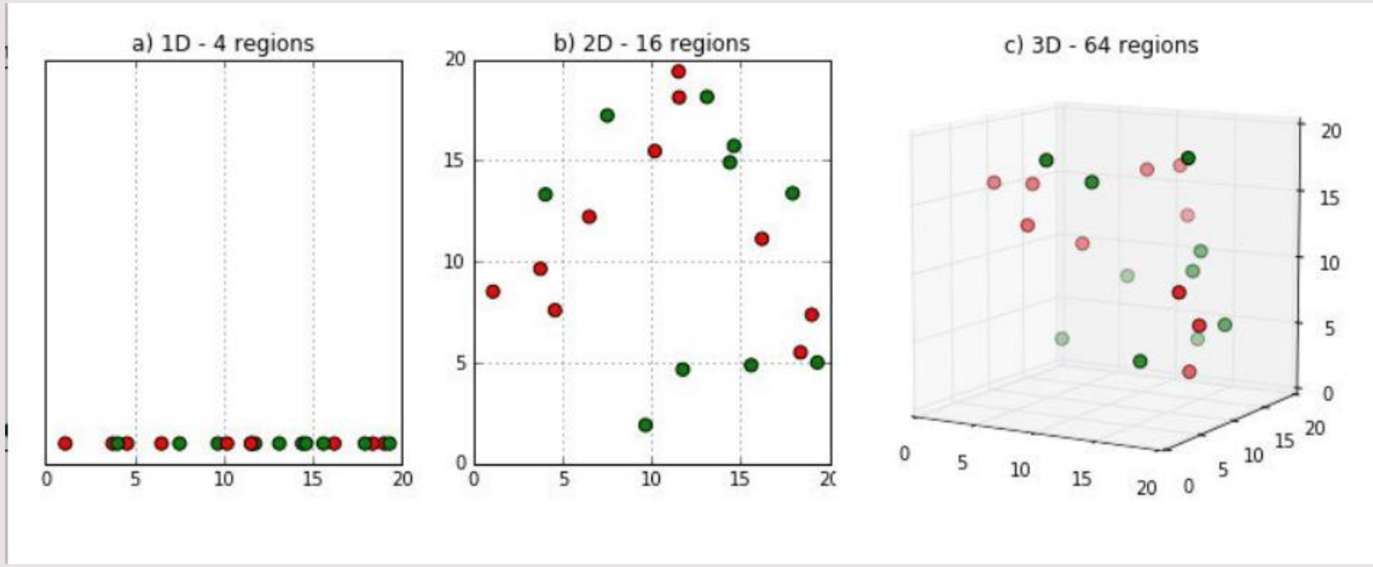


Srivani, Karpuram Dhanalakshmi. "Iris Flowers Classification Using Machine Learning." Analytics Vidhya, April 5, 2023

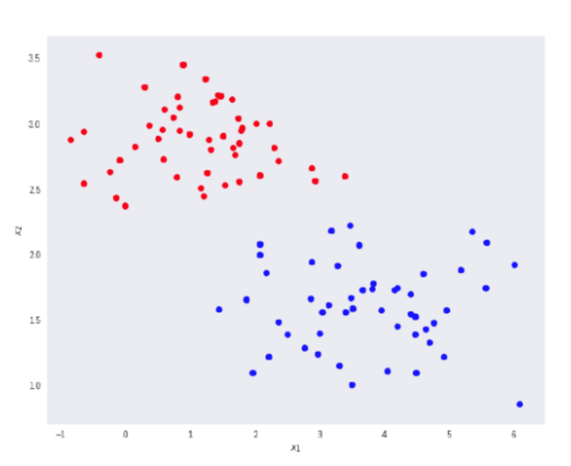
Curse of Dimensionality

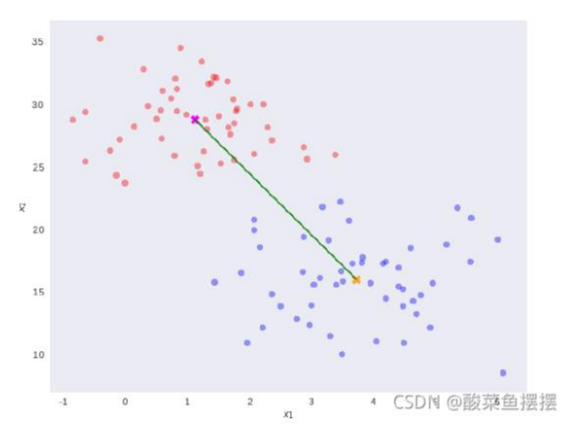
Curse of Dimensionality:

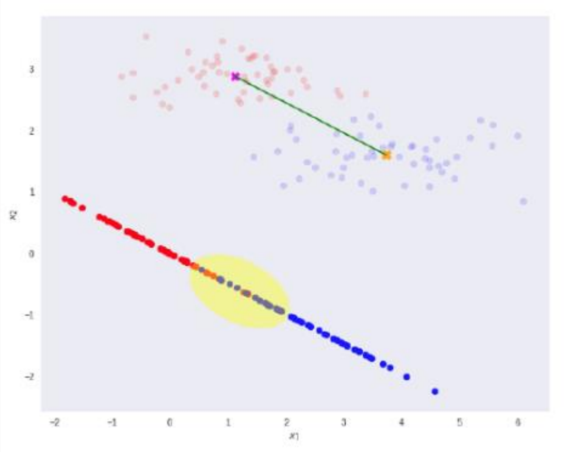
Classification accuracy decreases at higher dimensions.



Artificial Intelligence Stack Exchange, September 1, 1967.





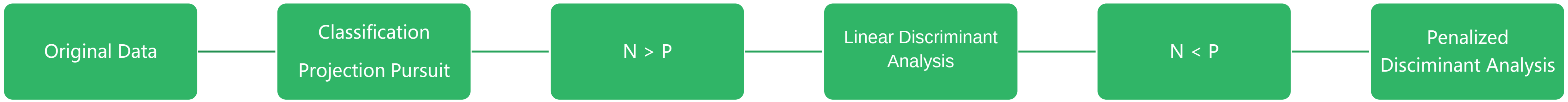


Projection Pursuit

freeCodeCamp.org. “An Illustrative Introduction to Fisher’s Linear Discriminant.” freeCodeCamp.org, February 8, 2021

Main Purpose

Improve the classification accuracy in situations where the number of variables is large compared to the number of observations.



Fisher's Linear Discriminant Analysis (LDA)

Averages all of the feature vectors for the samples within that class, providing a central point that summarizes the main characteristics of the class's data in the feature space.

Mean Vectors of each class ($\vec{m}_i$):

$$\vec{m}_i = \frac{1}{n_i} \sum_{\vec{x} \in D_i} \vec{x}$$

- $\overrightarrow{m_i}$: Average of class i features.
- n_i: Count of class i samples.
- D_i: Set of class i samples.
- $\vec{x}$: A dataset feature vector.

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Fisher's Linear Discriminant Analysis (LDA)

Mean Vectors of each class ($\vec{m}_i$):

$$\vec{m}_i = \frac{1}{n_i} \sum_{\vec{x} \in D_i} \vec{x}$$

Between-Class Scatter Matrix (S_B):

$$S_B = \sum_{i=1}^{c} N_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T$$

Within-Class Scatter Matrix (S_W):

$$S_W = \sum_{i=1}^{c} \sum_{\vec{x} \in D_i} (\vec{x} - \vec{m}_i)(\vec{x} - \vec{m}_i)^T$$

- S_B: Measures variance between classes.
- S_W: Measures variance within classes.
- N_i: Count of class i samples.
- c: Number of classes.
- T: Matrix or vector transpose.

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Fisher's Linear Discriminant Analysis (LDA)

Between-Class Scatter Matrix (S_B):

$$S_B = \sum_{i=1}^{c} N_i(\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T$$

Within-Class Scatter Matrix (S_W):

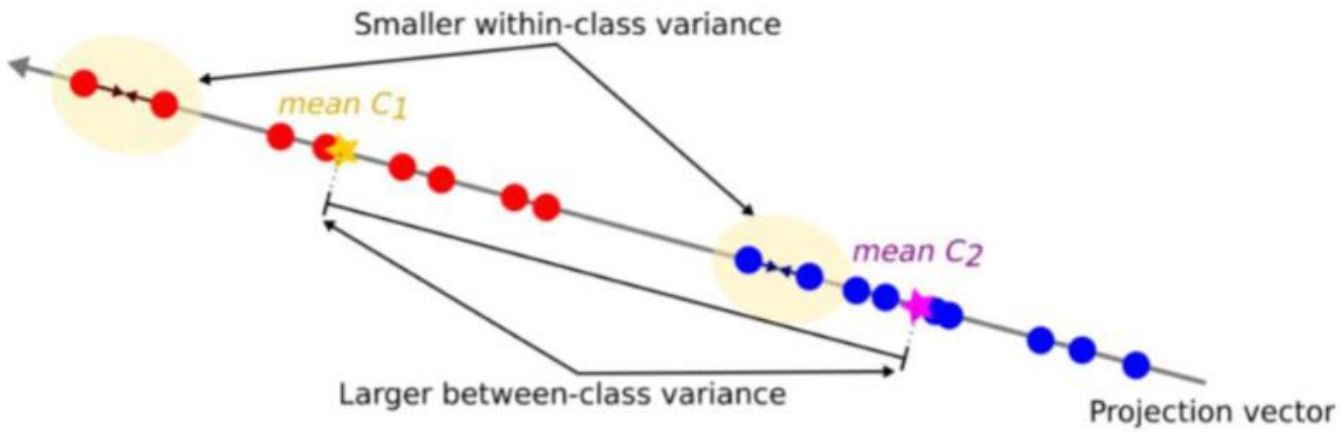
$$S_W = \sum_{i=1}^{c} \sum_{\vec{x} \in D_i} (\vec{x} - \vec{m}_i)(\vec{x} - \vec{m}_i)^T$$

Fisher's Criterion (J):

The criterion J that LDA aims to maximize is defined as:

$$J(\vec{w}) = \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}$$

where $\vec{w}$ is the linear discriminant vector (direction vector) that maximizes this ratio.



Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

LDA Projection Pursuit (PP) index

LDA projection Pursuit (PP) index is a measure used to evaluate the effectiveness of the Linear Discriminant Analysis (LDA) projection A onto a k-dimensional space

$$\begin{cases} 1 - \frac{\|A^T S_W A\|}{\|A^T (S_W + S_B) A\|}, & \text{for } \|A^T (S_W + S_B) A\| \neq 0, \\ 0, & \text{for } \|A^T (S_W + S_B) A\| = 0 \end{cases}$$

- A is an orthogonal projection matrix.
- S_W is the within-class scatter matrix.
- S_B is the between-class scatter matrix.
- $\| \cdot \|$ denotes a norm of the matrix.

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

PENALIZED DISCRIMINANT ANALYSIS

Formed by: Trevor Hastie, Andreas Buja, Robert Tibshirani

https://projecteuclid.org/journals/annals-of-statistics/volume-23/issue-1/Penalized-Discriminant-Analysis/10.1214/aos/1176324456.full

LDA limitation: lack of accuracy in high-dimensional data analysis

PDA Improvement: introducing a penalty term

Result: Improving the overall performance of the discriminant analysis.

PDA Projection Pursuit (PP) index

LDA Projection Pursuit (PP) index

$$\begin{cases} 1 - \frac{\|A^T S_W A\|}{\|A^T (S_W + S_B) A\|}, & \text{for } \|A^T (S_W + S_B) A\| \neq 0, \\ 0, & \text{for } \|A^T (S_W + S_B) A\| = 0 \end{cases}$$

PDA Projection Pursuit (PP) index

$$I_{PDA}(A, \lambda) = 1 - \frac{|A^T((1-\lambda)W_s + n\lambda I_p)A|}{|A^T((1-\lambda)(B_s + W_s) + n\lambda I_p)A|}$$

Lee, E.-K., & Cook, D. (2009). A projection pursuit index for large P Small N Data. Statistics and Computing, 20(3), 381–392.

Data Simulation

Sample: 100

Feature: 1000

Two Class: Class one and Class two

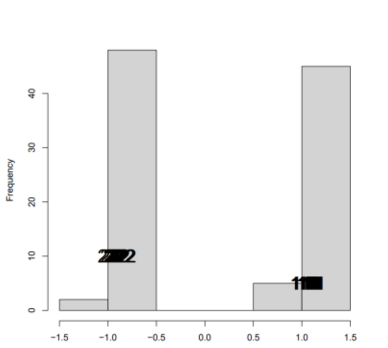
Distribution from feature 1 to 999 for both Class one and two: $X \sim Norm(0,1)$

Distribution on feature 1000 in Class one: $X \sim Norm(2.2,1)$

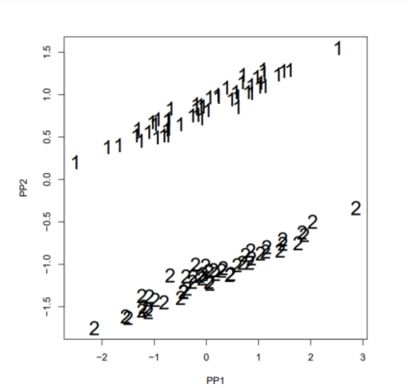
Distribution on feature 1000 in Class two: $X \sim Norm(-2.2,1)$

Example Projections

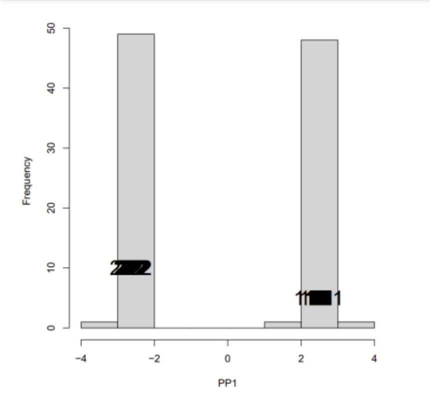
LDA



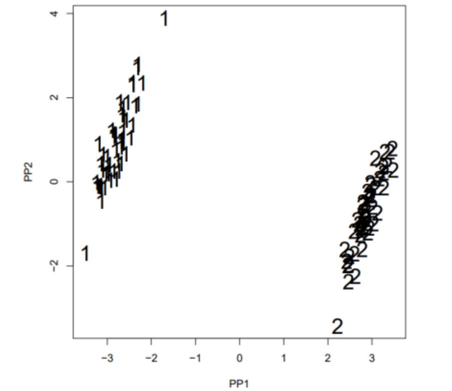
LDA



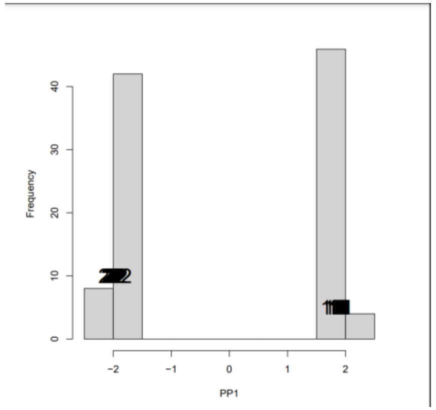
PDA, lambda = 0.5



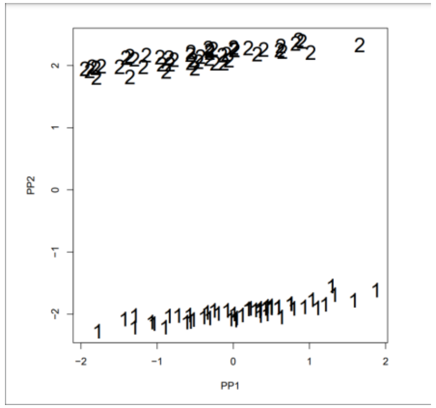
PDA, lambda = 0.5



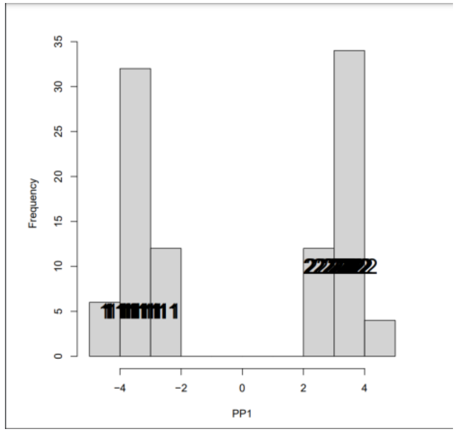
PDA, lambda = 0.1



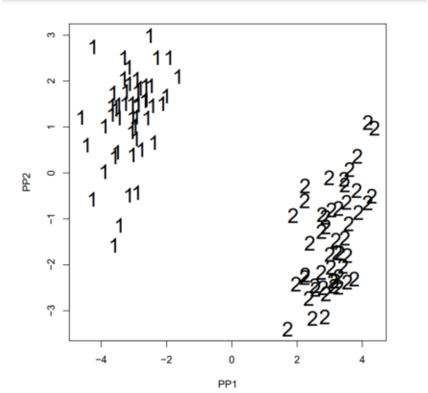
PDA, lambda = 0.1



PDA, lambda = 0.9



PDA, lambda = 0.9



Reference

FreeCodeCamp.org. “An Illustrative Introduction to Fisher’s Linear Discriminant.” freeCodeCamp.org, February 8, 2021.

Srivani, Karpuram Dhanalakshmi. “Iris Flowers Classification Using Machine Learning.” Analytics Vidhya, April 5, 2023. https://www.analyticsvidhya.com/blog/2022/06/iris-flowers-classification-using-machine-learning/.

Robert I. Kabacoff. *R in Action, Second Edition*. Manning Publications, 2015.

Johnson, Richard A. (Richard Arnold), and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th ed., Pearson Prentice Hall, 2007.

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.