Applications of Pinciple Component Analysis and Linear Discriminant Analysis on Consulting Survey Data

Ruinan(Victor) Zhang

March 24, 2017

1 Abstract

This paper address two statistical techiniques to analyze surveying data and applications on how to visualize those data for business analytics.

2 Motivation

The project is inspired by company, Smari Inc. The challenge was to provide statistical evidence for consulting business decisions by visualizing collected survey data. One specific question is how to evaluate competitors' performance on specific features. For example, cellphone brands may interested to know what do the customers think of them in terms of design, user-friendness, system-smoothiness. The goal of this project is to build an open-source online app for consulting companies to upload their dataset and construct visulizations to help them to answer such consulting questions.

3 Statistics Concepts Involved

3.1 Principal Component Analysis

Principal Component Analysis (PCA for the rest of the document) is a statistical technique to summarize features into algebraic combinations to capture the most correlation between the features and reponses. PCA is usually used to capture the most informations of provided features meanwhile reduce the number of features for modeling complexity.

Suppose I want to do PCA on a set of random vectors $\vec{X}$s with size equals to p. PCA generates a coefficient matrix E with size typically smaller than the length of $\vec{X}$ (otherwise, PCA loses its goal of dimentional reduction). Let Σ denotes as variance-covariance matrix of $\vec{X}$. Then the eigenvalues of Σ: λ_1 through λ_p. These eigenvalues are ordered so that λ_1 is the largest eigenvalue and λ_p is the smallest. Then we also calculate the eigenvectors of Σ: $\vec{e_1}$ through $\vec{e_p}$. It turns out the elements for these eigenvectors are the coefficients of principle components. Let the ith principle component denotes as Y_i, we can get variance of Y_i as following:

$$var(Y_i) = var(e_{i1}X_1 + ... + e_{ip}X_p) = \lambda_i$$

λ_1 being the biggest eigenvalue indicates that Y_i captures the most variance. The proportion of variance caputured by i th principle components can be calculated as $\lambda_i / \sum \lambda$. With the coefficients, the ith principal components can be easily calculated as following:

$$Y_i = e_{i1}X_1 + ... + e_{ip}X_p$$

[1]

In general, to visualize the PCA result on a 2-dimmension graph, the first two principle components are chosen as the x and y axis, and coordinates to represent X_i can be found by solving the matrix $Y = \vec{e} \times X$ with Y already known.

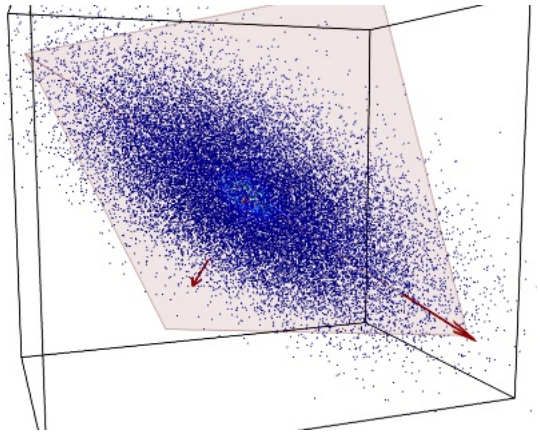

Figure 1: Summarize 3D data into 2 principle components

Figure 1 is a good visualization of summarizing 3-dimmensional data into 2 principle components:

3.2 Linear Discriminant Analysis

Linear discriminant analysis(LDA) is a method used in statistics to find a linear combination of features that characterizes or separates two or more classes of objects or event. The idea of LDA is to find a linear combination of variables to model the difference between the classes of a particular set of variable (usually categorical variables). LDA technique is frequently used as dimensionality reduction before classifications. LDA can be visualized similar to PCA as they both generate set of linear coefficients.

The idea behind LDA is to find a classification criterion and use it to minimize total error of classification. To make an example, suppose there is a dataset with two groups and k correspondant features for each group. Let μ_1 and μ_2 be the esponse variable mean of the first group and second group. You can start with some apriori probabilities of p_1 and p_2 to have this equation: $\mu_0 = p_1 \times \mu_1 + p_2 \times \mu_2$. In LDA, within-class ($S_w$) and ($S_b$) scatter are used to formulate

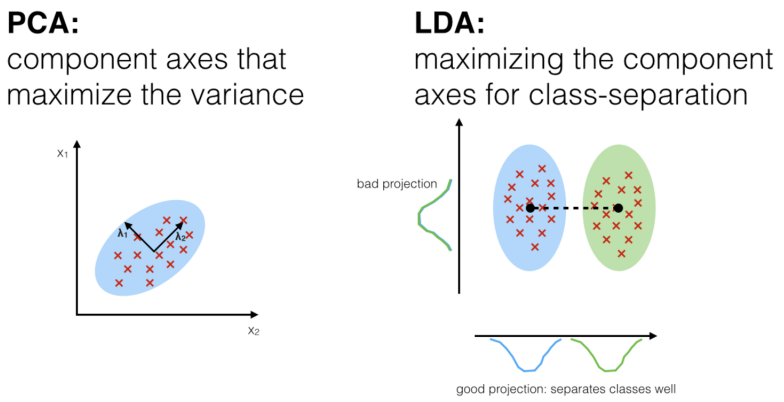

Figure 2: PCA vs. LDA

[3]

criteria for class seperation. Within-class scatter is the expected variance of each of the classes. $S_w = \sum_i (p_i + \Sigma_i)$ where Σ_i is the covariance matrix for ith group. The between-class scatter is computed using: $S_b = \sum_i (\mu_i - \mu_0) \times (\mu_j - \mu_0)^T$. Note that S_b can be thought of as the covariance of data set whose members are the mean vectors of each class. For the class independent transform, the optimizing criterion C is computed as $C = S_w^{-1} \times S_b$. Once C is optimized, we can apply it as the transformation matrix and obtain the linear-discriminantly transformed data. [2]

3.3 PCA vs. LDA

PCA and LDA are similar in the way they both generate linear combination of variables which best explains data. LDA explicitly attempts to model the difference between the classes of data while PCA models the similarities between the classes. Here is a good visulization on difference between PCA and LDA

4 Online App on ShinyIO

References

[1] https://onlinecourses.science.psu.edu/stat505/node/49

[2] S. Balakrishnama, A. Ganapathiraju. *Linear Discriminant Analysis*
https://www.isip.piconepress.com/publications/reports/1998/isip/lda/lda_theory.pdf

[3] Sebastian Raschka
http://sebastianraschka.com/Articles/2014_python_lda.html