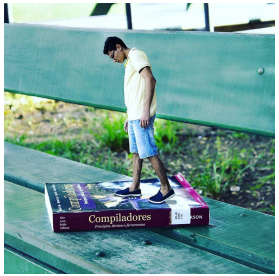
Applying Factor Analysis with cluster-then-label Semi-supervised Learning Approach in Classification Problems



Kaggle Days Meetup Delhi
NCR
05th December, Delhi

Caesar Lupum, Crislânio Macêdo

A brief overview ...

1. Semi-Supervised learning (SSL)
   a. Basic Concept
   b. Cluster-then-label Semi-supervised Learning Approach

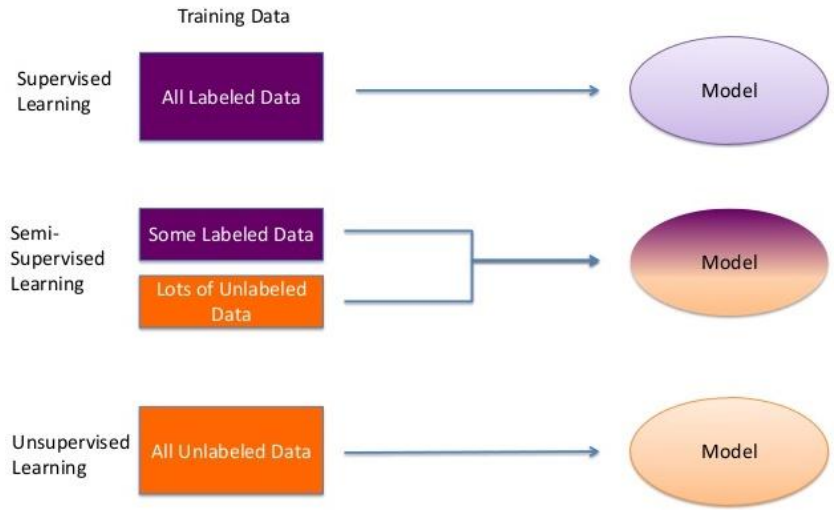
2. Factor Analysis (FA)
   a. First Notable mention
   b. Basic Concept
   c. Applications
   d. Factor Analysis Model
   e. Statistics Associated with Factor Analysis
   f. Conducting Factor Analysis
3. Demo
   a. Problem
   b. Pipeline modeling FA
   c. Pipeline modeling FA+SSL

WHOIAM?

> Crislânio Macêdo
> Majored in Computer Science from Universidade Federal do Ceará and Mastering Degree from Universidade Estadual do Ceará.
> Kaggle Notebook Master

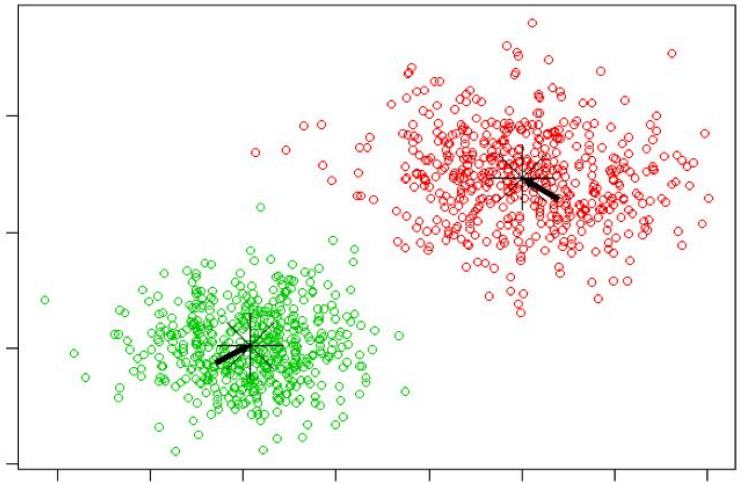
How to reach me:
LinkedIn:/crislanio
Kaggle:/caesarlupum
GitHub: @crislanio
Medium: @crislanio.ufc
Twitter: @crs_macedo

Semi-Supervised learning (SSL)-Basic Concept



Semi-supervised learning is a class of machine learning tasks and techniques that also make use of **unlabeled data** for training – typically a small amount of labeled data with a large amount of unlabeled data

Semi-Supervised learning (SSL)- Cluster-then-label Semi-supervised Learning Approach

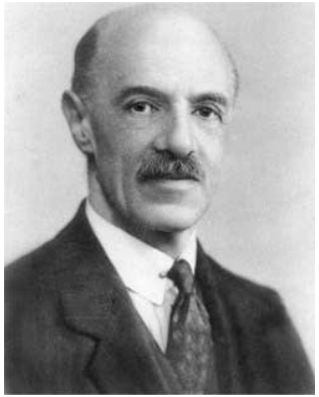


Intra-cluster distances are minimized.

Inter-cluster distances are maximized.

Factor Analysis - First Notable mention

Charles Edward Spearmen was known for his seminal work on testing and measuring of HUMAN INTELLIGENCE by using the FACTOR ANALYSIS during World War I.



CHARLES EDWARD SPEARMEN (BRITISH PSYCHOLOGIST)

Factor analysis is usually dated from Charles Spearman's paper 'General Intelligence' Objectively Determined and Measured published in the American Journal of Psychology in 1904

Factor Analysis - Basic Concept

What is factor analysis?

- Factor analysis is used:
  - To identify underlying dimensions, or factors, that explain the correlations among a set of variables.
  - To identify a new, smaller, set of uncorrelated variables to replace the original set of correlated variables.

Factor Analysis - Basic Concept

Advantages

Disadvantages

Factor Analysis - Basic Concept
Why Factor Analysis?

- Testing of theory
  - Explain covariation among multiple observed variables by Mapping variables to latent constructs (called “factors”)

- Understanding the structure underlying a set of measures
  - Gain insight to dimensions
  - Construct validation (e.g., convergent validity)

Factor Analysis - Applications

In physical and biological sciences

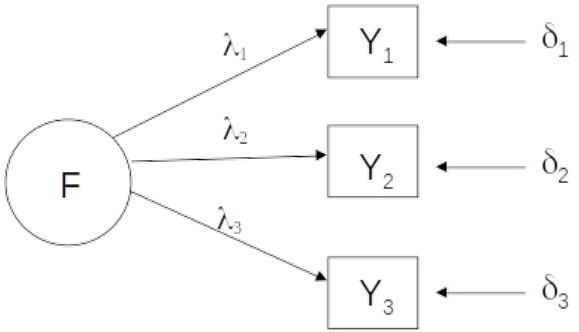
- geochemistry, hydrochemistry, ecology, molecular biology

Factor analysis can be used for summarizing high-density **oligonucleotide(**short DNA or RNA molecules**)** DNA microarrays

- Market Segmentation
- Achievement in education
- Diagnostic criteria in mental health
- Personality and cognition in psychology

source: https://en.wikipedia.org/wiki/Factor_analysis

Factor Analysis - Factor Analysis Model



$$Y_1 = \lambda_1 F + \delta_1$$

$$Y_2 = \lambda_2 F + \delta_2$$

$$Y_3 = \lambda_3 F + \delta_3$$

- The factor F is not observed; only Y_1, Y_2, Y_3 are observed
- δ_i represent variability in the Y_i NOT explained by F
- Y_i is a linear function of F and δ_i

Statistics Associated with Factor Analysis

Communality. Amount of variance a variable shares with all the other variables. This is the proportion of variance explained by the common factors.

Eigenvalue. Represents the total variance explained by each factor.

Factor loadings. Correlations between the variables and the factors.

Factor matrix. A factor matrix contains the factor loadings of all the variables on all the factors
Factor scores. Factor scores are composite scores estimated for each respondent on the derived factors.
Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. Used to examine the appropriateness of factor analysis.

Statistics Associated with Factor Analysis

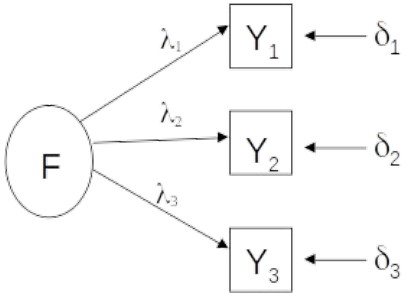
Bartlett's test of sphericity. Bartlett's test of sphericity is used to test the hypothesis that the variables are uncorrelated in the population (i.e., the population corr matrix is an identity matrix)

Correlation matrix. A correlation matrix is a lower triangle matrix showing the simple correlations, r, between all possible pairs of variables included in the analysis. The diagonal elements are all 1.

Percentage of variance. The percentage of the total variance attributed to each factor.

Scree plot. A scree plot is a plot of the Eigenvalues against the number of factors in order of extraction.

Factor Analysis - Factor Analysis Model



$$Y_1 = \lambda_1 F + \delta_1$$
$$Y_2 = \lambda_2 F + \delta_2$$
$$Y_3 = \lambda_3 F + \delta_3$$

Given all variables in standardized form, i.e.
$var(Y_i)=var(F)=1$

- Factor loadings: λ_i
  $\lambda_i = corr(Y_i,F)$

- Communality of Y_i: h_i^2
  $h_i^2 = \lambda_i^2 = [corr(Y_i,F)]^2$
  =% variance of Y_i explained by F

- Uniqueness of Y_i: $1-h_i^2$
  = residual variance of Y_i

- Degree of factorial determination:
  $=\Sigma \lambda_i^2/n$, where n=# observed variables Y

Factor Analysis - Matrix Notation

with n variables and m factors

$$\mathbf{Y}_{nx1} = \mathbf{\Lambda}_{nxm}\mathbf{F}_{mx1} + \boldsymbol{\delta}_{nx1}$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \cdots & \cdots & \lambda_{1m} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \lambda_{n1} & \cdots & \cdots & \lambda_{nm} \end{bmatrix}_{n\times m} \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix}_{m\times 1} + \begin{bmatrix} \delta_1 \\ \vdots \\ \vdots \\ \delta_n \end{bmatrix}_{n\times 1}$$

Demo - RETAIL CASE

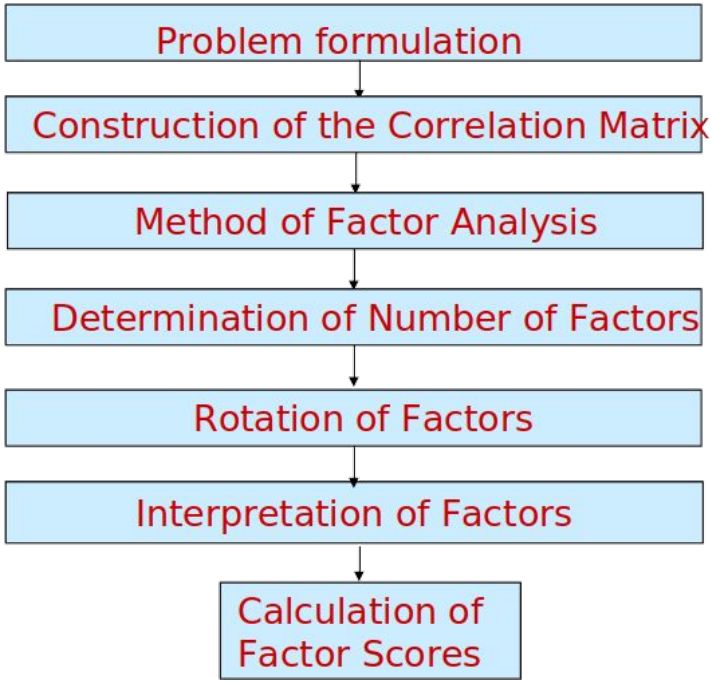
source

You are responsible for the analysis that will serve as a foundation for the strategy of entering the Brazilian market of a large multinational retailer in the supermarket sector.

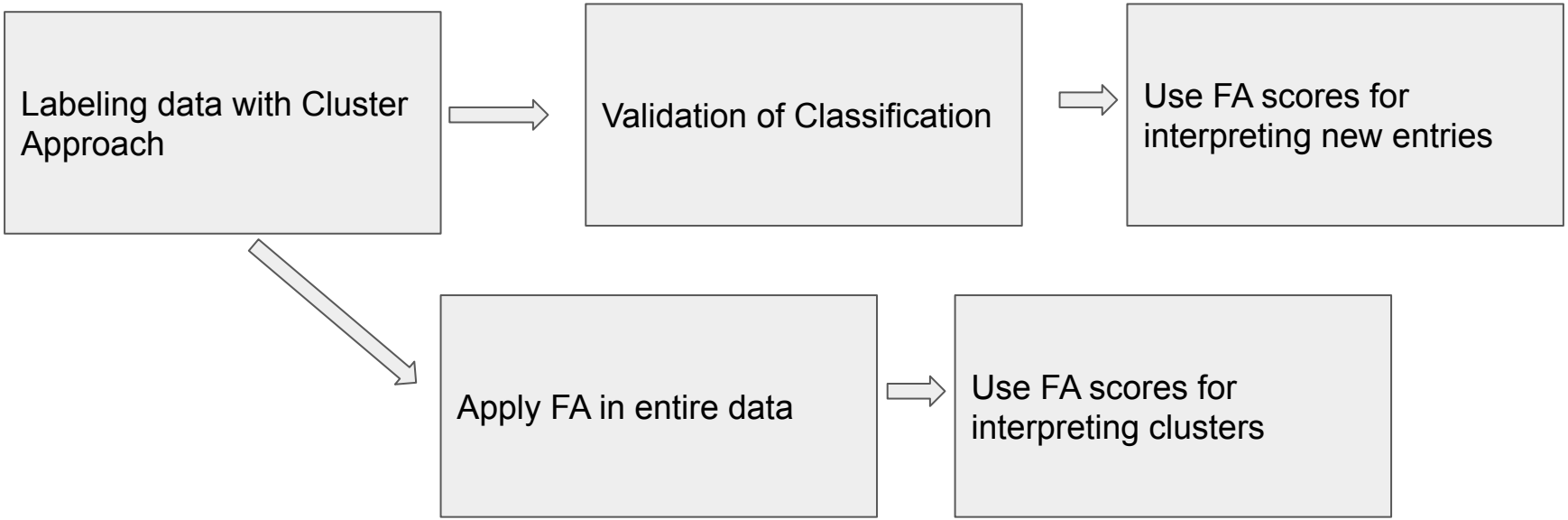
task 1: Classify Brazilian municipalities based on the available information
task 2: Develop a classification model to calculate the probability that a given municipality belongs to one of the groups created.
Which groups of municipalities should be the gateway to a company in the country? Why?

🛍️ Retail Case - SSL+FA: https://www.kaggle.com/caesarlupum/retail-case-ssl-fa/

Demo - Conducting Factor Analysis



Demo - SSL+ FA


Labeling data with Cluster Approach
Validation of Classification
Use FA scores for interpreting new entries
Apply FA in entire data
Use FA scores for interpreting clusters

References and resources

Types of learning
Factor analysis
FA in psychometrics
Retail Case SSL +Factor Analysis

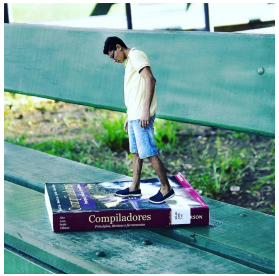
kaggle DAYS MEETUP

Q&A section

LOGICAI

kaggle

Applying Factor Analysis with cluster-then-label Semi-supervised Learning Approach in Classification Problems



Kaggle Days Meetup Delhi
NCR
05th December, Delhi

Caesar Lupum, Crislânio Macêdo