IIT ROORKEE

NPTEL ONLINE
CERTIFICATION COURSE

Discriminant Analysis Part-2

LECTURE 60

DR. GAURAV DIXIT
DEPARTMENT OF MANAGEMENT STUDIES

Discriminant Analysis

- Statistical technique
  - Used for classification and profiling tasks
  - Model-based approach
  - Idea is
    - To find a separating line or hyperplane equidistant from centroids of different classes

      Or
    - To find a separating line or hyperplane that is best at discriminating the records into different classes
  - Classification procedure is based on distance based metrics
    - Based on the distance of a record from each class

Discriminant Analysis

- Classification
  - Best separation between items is found by measuring their distance from each class
  - An item is classified to the closest class
- Euclidean distance metric
  - Distance of a record (x_1, …, x_p) from centroid ($\bar{x}_1$, …, $\bar{x}_p$) of a class is computed

$$D_{eu}(x, \bar{x}) = \sqrt{(x_1 - \bar{x}_1)^2 + \cdots + (x_p - \bar{x}_p)^2}$$

Where centroid $\bar{x}$ is a vector of means of p predictors

Discriminant Analysis

- Issues with Euclidean distance metric
  - Distance values depend on the unit of a measurement
  - Based on mean and doesn’t account for variance
    - Variability plays an important role in determining the closeness of a record to a particular class
  - Distance should be computed using std. dev. (z-scores) instead of unit of measurement
  - Correlation between variables is ignored

Discriminant Analysis

- “Statistical distance” (or Mahalanobis distance) can be used to overcome issues with Euclidean distance metric

$$D_{ml}(x, \bar{x}) = [x - \bar{x}]' \, S^{-1} \, [x - \bar{x}]$$

Where $[x - \bar{x}]'$ is transpose matrix of $[x - \bar{x}]$

- Column vectors are turned into row vectors

and S^{-1} is inverse matrix of S (covariance matrix between p predictors)

- Can be considered as p-dimensional extension of division operation

Discriminant Analysis

- Linear Classification Functions
  - Used as basis for separation of records into classes
    - Compute classification score measuring closeness of a record to each class
    - Highest classification score is equivalent of smallest statistical distance
  - Main idea is
    - To find linear functions of predictors that maximize ratio of between-class variability to within-class variability
- Open RStudio

Discriminant Analysis

- Assumptions and other issues
  - Predictors follow multivariate normal distribution for all classes
    - Given adequate sample points for all classes, relatively robust to violations of normality assumption
  - Correlation structure between predictors for each class should be same
  - Sensitive to outliers

Discriminant Analysis

- Further Comments on discriminant analysis
  - Application and performance aspects are similar to multiple linear regression
  - In discriminant analysis, coefficients of linear discriminant are optimized w.r.t class separation
    - In linear regression, coefficients are optimized w.r.t outcome variable
  - Estimation technique is least squares
    - Same as linear regression

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…