**Discriminant analysis**

**Description of the dataset**

The data set "Wholesale customers data.csv" refers to clients of a wholesale distributor.

It includes the annual spending in monetary units (m.u.) on diverse product categories.

It contains the following variables:

1) channel: customers' Channel - Horeca (Hotel/Restaurant/Café) (1) or Retail channel (2) (Nominal);

2) region: customers' Region- Lisbon (1), Oporto (2) or Other (3) (Nominal);

3) fresh: annual spending (m.u.) on fresh products (Continuous);

4) milk: annual spending (m.u.) on milk products (Continuous);

5) grocery: annual spending (m.u.) on grocery products (Continuous);

6) frozen: annual spending (m.u.)on frozen products (Continuous);

7) detergents\_paper: annual spending (m.u.) on detergents and paper products (Continuous)

8) delicatessen: annual spending (m.u.) on and delicatessen products (Continuous);

**Overview of the strategy:**

Suppose our client is a Wholesale Distributor that has information on current and past clients expenditure on different products categories. We are required to design a tool to discriminate between smaller clients (i.e., the “Horeca” channel) and bigger clients (i.e., the “Retail” channel) depending on the expenditure per category of products.

First, we present a general overview of the theoretical aspects of Discriminant Analysis techniques.

Then we present a general overview of the structure of the data. We show the frequency of observations across channels and the distributions of the variables of interest. We study the assumptions of the model (namely normality and homoskedasticity between classes), comment on them, and transform the data accordingly for the analysis.

After that, we perform discriminant analysis using Linear Discriminant Analysis, Quadratic Discriminant Analysis and Bayes Discriminant Analysis. We compare different measures of accuracy across models and recommend one of them based on the results.

**Descriptive statistics:**

Table \*\*\* contains the descriptipe statistics of the dataset. [\*\*\* general comment on the min max means variances]. The categorical variables are class and region, both of them denoting the type of client (i.e., if the client is a hotel, restaurant, or café (“horeca”)) and the region where the client is based (i.e., Lisbon, Oporto, or other region). The covariates contain information about expenditure in monetary units across 6 different categories:

1. Fresh: annual spending on fresh products.
2. Milk: annual spending on milk products.
3. Grocery: annual spending on grocery products.
4. Frozen: annual spending on frozen products.
5. Detergents\_paper: annual spending on detergents and paper products.
6. Delicatessen: annual spending (m.u.) on and delicatessen products.

Calendar

Description automatically generated

Table \*\*\* displays the frequency of observations across the distribution channel and the region. 67% of the clients belong to the “Horeca” category, and 32% belong to the retail channel. 17% of the clients are located in Lisbon, 11 percent of them are located in Oporto, and the remaining 72% are located in a different location. Figure \*\*\* visualizes this data.

Text, letter

Description automatically generated

Table

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated

Figure \*\*\* displays the distributions of each of the covariates. In general, most clients spend between 0k and 30k monetary units a year on all categories. Some clients spend much more than the average, with annual expenditures ranging between 30k and 100k sin some categories. Expenditure on detergents and/or paper, frozen products, and delicatessen products ranges from 0k to 10k approximately. This is expected as they are either luxury goods or goods which frequency is not as high as basic needs food such as fresh products, milk and grocery products in general. Figure \*\*\* condenses the distributional information across covariates for ease of interpretation.

Chart

Description automatically generated

Chart

Description automatically generated

**Testing the assumptions of the model on the data.**

LDA assumes that the observations come from a normally distributed DGP with constant variance between classes. This generates decision rules that are linear in the covariates (i.e., it is possible to discriminate observations by a line or a plane, depending on the number of classes). If the homoskedasticity assumption is violated, then it is better to apply QDA. QDA generates decision rules by weighting \*\*\*. Bayes \*\*\*

We first explore variance equality between individual covariates using Bartlett's test for equal variance across samples for individual variables. Then we perform Cai TT, Ma Z (2013\*\*\*) two sample test of equality of covariance matrices. Results are in table \*\*\*. Bartlett's test strongly rejects the null hypothesis of equality of variance between groups for each of the covariates. Results are in figure \*\*\*. This gives us a hit that for each of the covariates, the distribution of the data is quite different across classes. For the Cai and Ma test we strongly reject the null hypothesis that the covariance matrices between classes is equal.

Text

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

Next, we explore the normality assumptions. First, we compute the Shapiro-Wilk test of normality for each of the covariates individually. For all the individual covariates we strongly reject the null hypothesis of normality. Results are in Table \*\*\*. We then perform a Generalized Shapiro-Wilk test for Multivariate Normality. We again strongly reject the null hypothesis of joint normality in the sample. Results in Table \*\*\*.

Text

Description automatically generated with low confidence

Graphical user interface, text, application, email

Description automatically generated

We decide to apply a log transformation and standardize the data for ease of manipulation and to comply with the model assumptions. It is worth emphasizing that QDA does not require the assumption of equality of covariances between classes, and that Bayes Discriminant rule does not require the assumption of normality in the data.

As expected, once transformed and standardized, the mean and variance of the transformed covariates are equal to 0 and 1 respectively (see Table \*\*\*). Figure \*\*\* and Figure \*\*\* show that the distribution of transformed variables now resembles a normal distribution.

Table

Description automatically generated with medium confidence

Box and whisker chart

Description automatically generated with medium confidence

Chart

Description automatically generated

Nonetheless, the same normality tests as before are performed on the transformed data.

The proportion of observations in each class should be considered when creating the priors. If not and the probabilities of classification do not match the class proportions, the classification procedure will be erroneous. Even though for two out of the six variables the hypothesis of homoskedasticity is still dejected, the generalized two-sample test of equality of covariances strongly rejects the null hypothesis of homoskedasticity.

Text

Description automatically generated with medium confidence A picture containing text

Description automatically generated

A similar story is told for the normality tests. The hypothesis of normality, both at the individual variable level and as a joint distribution of the covariates is strongly rejected. In a nutshell, the data is neither normally distributed nor homoscedastic across classes.

Text

Description automatically generated with medium confidence Graphical user interface, text, application, email

Description automatically generated

**Discriminant Analysis:**

The total number of observations are split into two groups: the "training" group and the "testing” group using a split ratio of 0.65. The training subsample is used to create the discriminant rules while the testing subset is used to determine the different accuracy measures. After splitting the data, it is verified that the proportions of observations in the two classes are maintained both in the training set and in the testing set. The decomposition proportions are presented in Table \*\*\*.

A picture containing text

Description automatically generated

The assumptions analysis in the previous code suggested that Bayes DA was more suited for analyzing the data and that QDA better than LDA. Table \*\*\* presents accuracy measures for each of the three methods for discriminating between classes. Note that these accuracy measures use the training subset. The discriminating metho with the highest accuracy rate is LDA (95%) while QDA and BAYES have accuracy rates of 93.7% and 94.4% respectively. The false positive rate is quite low for all discriminating methods, all range between 3% and 5%. Note, however, that because the accuracy is based on the same observations that were used to compute the discriminant rules, this method of evaluation is naively optimistic.

A picture containing table

Description automatically generated

The cross-validation criterion uses the training subset we had previously defined. Table \*\*\* presents the accuracy measures under the testing subset. The accuracy rate of LDA is around 86% while QDA and BAYES are the same at 87%. The misclassification rates range around 14% to126% being the highest for the LDA. In general, the measures of accuracy are very close across classification methods which suggest that means between classes are quite different while variance within classes is low. Note also that the similarity of accuracy measures between QDA and BAYES could Indicate that the estimates of the variance/covariance matrices are accurate enough.

A picture containing table

Description automatically generated

A picture containing graphical user interface

Description automatically generated

**Additional analysis**

Compute how does the rules perform under untransformed data. It changes almost nothing \*\*\*.

Table

Description automatically generated with low confidence

Useful links:

<https://stats.stackexchange.com/questions/31366/linear-discriminant-analysis-and-bayes-rule-classification>

<https://cran.r-project.org/web/packages/covTestR/covTestR.pdf>

https://www.analytixlabs.co.in/blog/factor-analysis-vs-pca/#:~:text=PCA%20is%20used%20to%20decompose,variables%20in%20the%20dataset%20data.