Lecture 12: Logit vs LDA

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

February 12, 2024

Plan

1) Logit vs LDA

2) QDA

3) LDA vs PCA

4) Modeling approach

Logistic regression

Suppose we work with binary outputs, i.e., $y_i \in \{0, 1\}$.

Linear regression may not be the best model.

- $x^T \beta \in \mathbb{R}$ not in $\{0, 1\}$.

- Linearity may not be appropriate. Does doubling the predictor doubles the probability of $Y = 1$? (e.g. probability of going to the beach vs outdoors temperature).

Logistic regression: Different perspective. Instead of modelling the $\{0, 1\}$ output, we model the probability that $Y = 0, 1$.

Idea: We model $P(Y = 1 \mid X = x)$.

- Now: $P(Y = 1 \mid X = x) \in [0, 1]$ instead of $\{0, 1\}$.

- We want to relate that probability to $x^T \beta$.

Logistic regression

We assume the log-odds

$$\begin{aligned}\operatorname{logit}(P(Y=1 \mid X=x)) &= \log \frac{P(Y=1 \mid X=x)}{1-P(Y=1 \mid X=x)} \\ &= \log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = x^T\beta.\end{aligned}$$

Equivalently,

$$\begin{aligned} P(Y=1 \mid X=x) &= \frac{e^{x^T\beta}}{1+e^{x^T\beta}} \\ P(Y=0 \mid X=x) &= 1-P(Y=1 \mid X=x) = \frac{1}{1+e^{x^T\beta}}\end{aligned}$$

The logistic function is given by the inverse-logit:

$$\operatorname{logit}^{-1}(x) = \operatorname{logistic}(x) = \frac{e^x}{(1+e^x)} = \frac{1}{(1+e^{-x})}$$

Hence, the logit function is the quantile function associated with the standard logistic distribution.

Logistic regression

In summary, we are assuming:

- $Y \mid X = x \sim \text{Bernoulli}(p)$.

- $\text{logit}(p) = \text{logit}(E(Y \mid X = x)) = x^T\beta$.

More generally, one can use a generalized linear model (GLM). A GLM consists of:

- A probability distribution for $Y \mid X = x$ from the exponential family.

- A linear predictor $\eta = x^T\beta$.

- A link function g such that $g(E(Y \mid X = x)) = \eta$.

Estimating the parameters

In logistic regression, we are assuming a model for Y. We typically estimate the parameter β using maximum likelihood.

Recall: If $Y \sim \text{Bernoulli}(p)$, then

$$P(Y = y) = p^y (1-p)^{1-y}, \quad y \in \{0, 1\}.$$

Thus, $L(p) = \prod_{i=1}^n p^{y_i} (1-p)^{1-y_i}$.

Here $p = p(x_i, \beta) = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$. Therefore,

$$L(\beta) = \prod_{i=1}^n p(x_i, \beta)^{y_i} (1 - p(x_i, \beta))^{1-y_i}.$$

Taking the logarithm, we obtain

$$\begin{aligned} l(\beta) &= \sum_{i=1}^n y_i \log p(x_i, \beta) + (1-y_i) \log (1 - p(x_i, \beta)) \\ &= \sum_{i=1}^n y_i \left(x_i^T \beta - \log \left(1 + x_i^T \beta\right)\right) - (1 - y_i) \log \left(1 + e^{x_i^T \beta}\right) \\ &= \sum_{i=1}^n \left[y_i x_i^T \beta - \log \left(1 + e^{x_i^T \beta}\right)\right]. \end{aligned}$$

Logistic regression

Taking the derivative:

$$\frac{\partial}{\partial \beta_j} l(\beta) = \sum_{i=1}^{n} \left[y_i x_{ij} - x_{ij} \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right].$$

Needs to be solved using numerical methods (e.g. Newton-Raphson).

Logistic regression often performs well in applications. As before, penalties can be added to regularize the problem or induce sparsity. For example,

$$\min_{\beta} -l(\beta) + \alpha \|\beta\|_1$$

$$\min_{\beta} -l(\beta) + \alpha \|\beta\|_2.$$

Logistic regression

Example

If we analyze the case of a pairwise relationship (i.e. the case when the probability of an event occurring depends on a single factor $\boldsymbol{x}$), then the logit model can be written as follows:

$$P\left(y_i = 1\right) = F\left(z_i\right) = \frac{1}{1 + e^{-z_i}},$$

where $z_i = \beta_1 + \beta_2 x_i$;

Say, we have estimated the coefficients and the estimated probability of passing the test on hours of study:

$$\hat{P}\left(y_i = 1\right) = \frac{1}{1 + e^{-(-9 + 0.5 x_i)}}.$$

How can such results be interpreted?

– Estimate the probability of the occurrence under certain conditions.

$$\hat{P}\left(y_i = 1 \mid x_i = 15\right) = 0.18.$$

That is, for a student who prepared for 15 hours, the probability of passing the test is 18%.

Logistic regression

– Interpret the results in terms of the change in the dependent variable resulting from a subtle change in the regressor, i.e. marginal effect

To do this, we calculate the derivative of the probability with respect to x:

$$\frac{dP\left(y_i = 1\right)}{dx} = \frac{e^{-(\beta_1+\beta_2 x)}}{\left(1 + e^{-(\beta_1+\beta_2 x)}\right)^2} \cdot \beta_2$$

From an applied point of view, the inconstancy of the marginal effect gives rise to some complexity: it is not very clear at what point to calculate it.

In practice,

1) marginal effect for the sample average.

E.g. marginal effect at point $\bar{x}$ – the average sample preparation time for the test

2) average marginal effect: calculate the marginal effect for each student, then calculate the average of the n marginal effects.

Logistic regression with more than 2 classes

- Suppose now the response can take any of $\{1, \ldots, K\}$ values.

- Can still use logistic regression.

- We use the categorical distribution instead of the Bernoulli distribution.

- $P(Y = i \mid X = x) = p_i, 0 \leq p_i \leq 1, \sum_{i=1}^{K} p_i = 1.$

- Each category has its own set of coefficients:

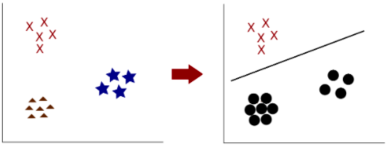
$$P(Y = i \mid X = x) = \frac{e^{x^T \beta^{(i)}}}{\sum_{i=1}^{K} e^{x^T \beta^{(i)}}}.$$

- Estimation can be done using maximum likelihood as for the binary case.

Multiple classes of data

Other popular approaches to classify data from multiple categories.

- One versus all:(or one versus the rest) Fit the model to separate each class against the remaining classes. Label a new point x according to the model for which $x^T\beta + \beta_0$ is the largest.



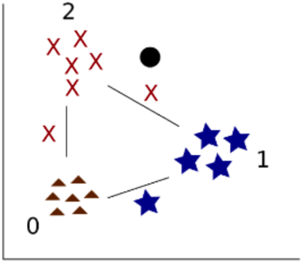
Need to fit the model K times.

Multiple classes of data

- One versus one:

(1) Train a classifier for each possible pair of classes. Note: There are $\binom{K}{2} = K(K-1)/2$ such pairs.

(2) Classify a new points according to a majority vote: count the number of times the new point is assign to a given class, and pick the class with the largest number.



Need to fit the model $\binom{K}{2}$ times (computationally intensive).

Linear discriminant analysis (LDA)

- Categorical data Y. Predictors $X_1, \ldots, X_p$.

- We saw how logistic regression can be used to predict Y by modelling the log-odds

$$\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = x^T \beta.$$

- More now examine other models for $P(Y=i \mid X=x)$.

Recall: Bayes' theorem. Given two events A, B :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Using Bayes' theorem

- $P(Y=i \mid X=x)$ harder to model.

- $P(X=x \mid Y=i)$ easier to model.

Going back to our prediction using Bayes' theorem:

$$P(Y=i \mid X=x) = \frac{P(X=x \mid Y=i)P(Y=i)}{P(X=x)}$$

Linear discriminant analysis (LDA)

More precisely, suppose

- $Y \in \{1, \dots, k\}$.
- $P(Y = i) = \pi_i \quad (i = 1, \dots, k)$.
- $P(X = x \mid Y = i) \sim f_i(x) \quad (i = 1, \dots, k)$.

Then

$$P(Y = i \mid X = x) = \frac{P(X = x \mid Y = i)P(Y = i)}{P(X = x)} = \frac{P(X = x \mid Y = i)P(Y = i)}{\sum_{j=1}^{k} P(X = x \mid Y = j)P(Y = j)} = \frac{f_i(x)\pi_i}{\sum_{j=1}^{k} f_j(x)\pi_j}.$$

- We can easily estimate π_i using the proportion of observations in category i.
- We need a model for $f_i(x)$.

Linear discriminant analysis (LDA)

A natural model for the f_j s is the multivariate Gaussian distribution:

$$f_j(x) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma_j}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}.$$

Linear discriminant analysis (LDA): We assume $\Sigma_j = \Sigma$ for all $j = 1, \ldots, k$.

Quadratic discriminant analysis (QDA): general case, i.e., Σ_j can be distinct.

Note: When p is large, using QDA instead of LDA can dramatically increase the number of parameters to estimate.

In order to use LDA or QDA, we need:

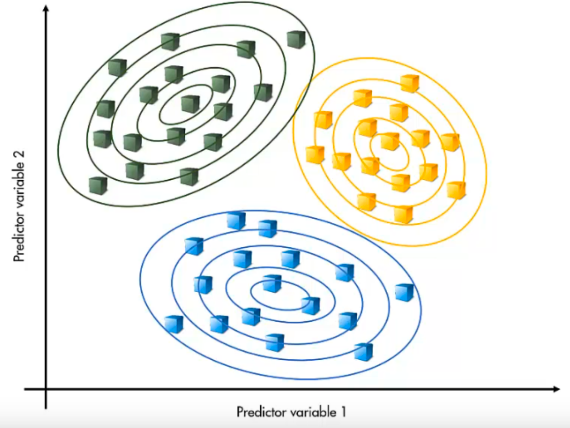
- An estimate of the class probabilities π_j.

- An estimate of the mean vectors μ_j.

- An estimate of the covariance matrices Σ_j (or Σ for LDA).

Linear discriminant analysis (LDA)

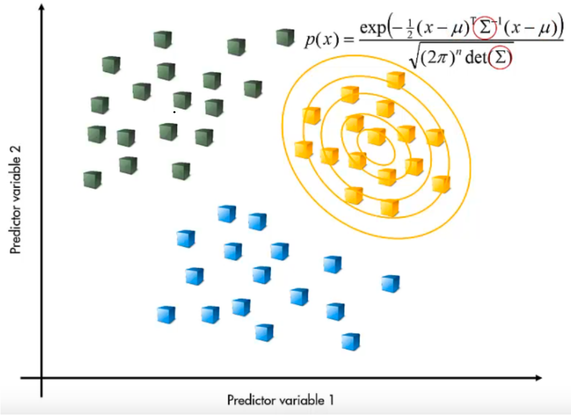
Consider samples from a multi-dimensional normal distribution



Linear discriminant analysis (LDA)

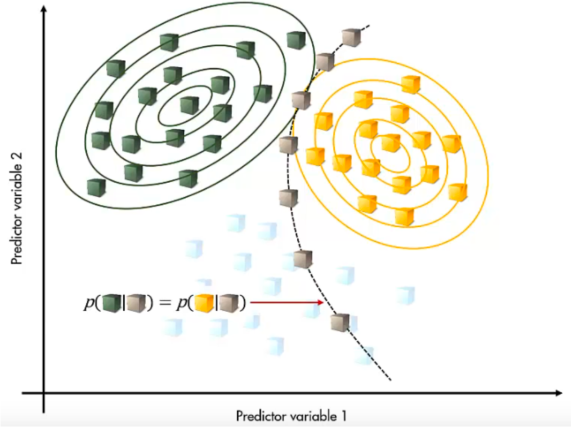
Theoretically we could fit an *n*-dimensional normal distribution to the observations in each class

This involves calculating the mean vector and covariance matrix for each class, which determine the center and shape of the distribution respectively



Linear discriminant analysis (LDA)

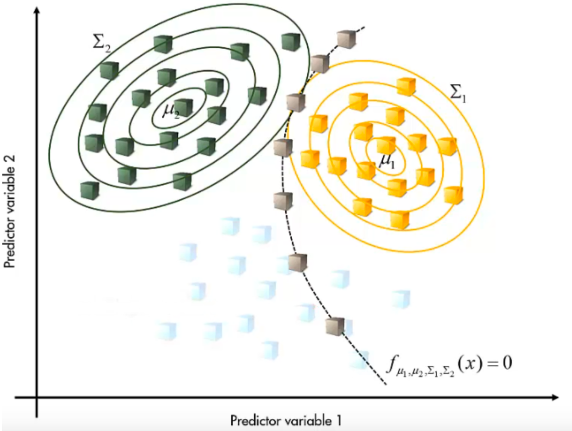
Having fitted the distributions we could draw a boundary between the classes – the set of points where the probabilities are equal



Observations on one side of the boundary would be classified as one class observations on the other side as the other class

Linear discriminant analysis (LDA)

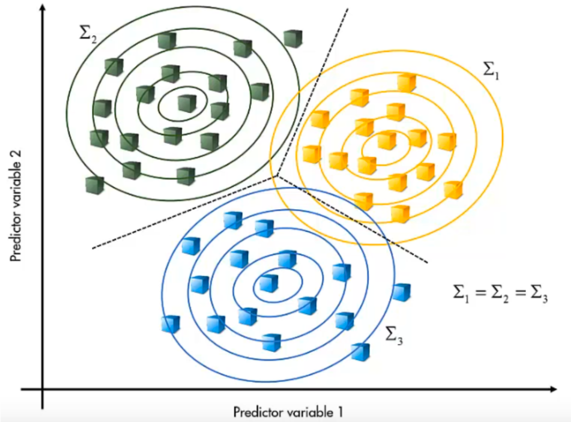
We can do all this theoretically resulting in an equation for the boundary that depends on the parameters of the fitted distribution



=> we just need to calculate the means and covariances and apply the formula for the boundary

Linear discriminant analysis (LDA)

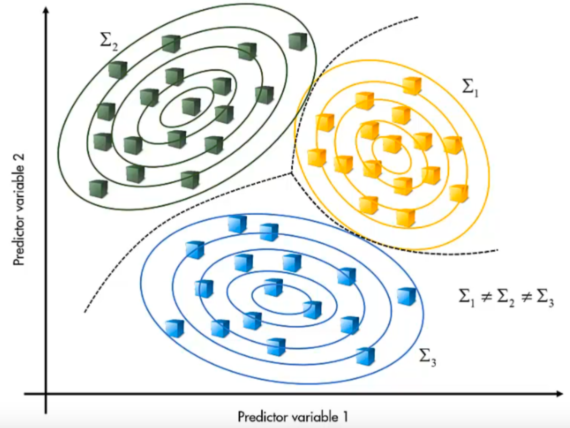
Assumption: the distributions all have the same shape, i.e. same covariance matrices



=> the boundaries turn out to be linear and the coefficients of these linear boundaries are easily calculated from the individual class means and shared covariance matrix of the observations

Linear discriminant analysis (LDA)

IF the assumption about the same covariance matrices are the same for all classes is unreasonable



=> the boundaries turn out to be quadratic and again the coefficients are determined by the mean vectors and covariance matrices of the observed classes

Linear discriminant analysis (LDA)

Estimating the parameters

LDA: Suppose we have N observations, and N_j of these observations belong to the j category ($j = 1, \dots, k$). We use

- $\hat{\pi}_j = N_j / N$.

- $\hat{\mu}_j = \frac{1}{N_j} \sum_{y_i = j} x_i$ (average of x over each category).

- $\hat{\Sigma} = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{y_i=j} \left(x_i - \hat{\mu}_j\right) \left(x_i - \hat{\mu}_j\right)^T$. (Pooled variance.)

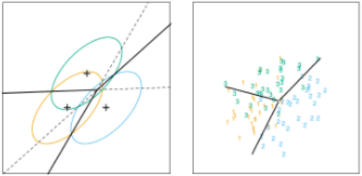


Figure: The left panel shows three Gaussian distributions, with the same covariance and different means. Contours of constant density (95% prob.).
Broken straight lines – The Bayes decision boundaries between each pair of classes Solid lines – Bayes decision boundaries.
Right panel a sample of 30 obs., and the fitted LDA decision boundaries.

LDA: linearity of the decision boundary In the previous figure, we saw that the decision boundary is linear. Indeed, examining the log-odds:

$$
\begin{aligned}
&\log \frac{P(Y=l \mid X=x)}{P(Y=m \mid X=x)} = \log \frac{f_l(x)}{f_m(x)} + \log \frac{\pi_l}{\pi_m} \\
&= \log \frac{\pi_l}{\pi_m} - \frac{1}{2}\left(\mu_l + \mu_m\right)^T \Sigma^{-1}\left(\mu_l - \mu_m\right) + x^T \Sigma^{-1}\left(\mu_l - \mu_m\right) \\
&= \beta_0 + x^T \beta.
\end{aligned}
$$

Note that the previous expression is linear in x. Recall that for logistic regression, we model

$$
\log \frac{P(Y=l \mid X=x)}{P(Y=m \mid X=x)} = \beta_0 + x^T \beta.
$$

How is this different from LDA?

- In LDA, the parameters are more constrained and are not estimated the same way.

- Can lead to smaller variance if the Gaussian model is correct.

- In practice, logistic regression is considered safer and more robust.

- LDA and logistic regression often return similar results.

QDA: quadratic decision boundary

Let us now examing the log-odds for QDA: in that case no simplification occurs as before

$$
\begin{aligned}
&\log \frac{P(Y=l \mid X=x)}{P(Y=m \mid X=x)} \\
&= \log \frac{\pi_l}{\pi_m} + \frac{1}{2} \log \frac{\det \Sigma_m}{\det \Sigma_l} \\
&- \frac{1}{2} (x-\mu_l)^T \Sigma_l^{-1} (x-\mu_l) - \frac{1}{2} (x-\mu_m)^T \Sigma_l^{-1} (x-\mu_m) .
\end{aligned}
$$

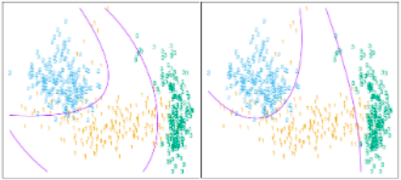


Figure: Two methods for fitting quadratic boundaries.
The left plot the quadratic decision boundaries obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$.
The right plot shows the quadratic decision boundaries found by QDA.
The differences are small, as is usually the case.

QDA: quadratic decision boundary

- Despite their simplicity, LDA and QDA often perform very well.

- Both techniques are widely used.

Problems when $n < p$:

- Estimating covariance matrices when n is small compared to p is challenging.

- The sample covariance (MLE for Gaussian) $S = \frac{1}{n-1} \sum_{j=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$ has rank at most $\min(n, p)$ so is singular when $n < p$.

- This is a problem since Σ needs to be inverted in LDA and QDA.

Many strategies exist to obtain better estimates of Σ (or Σ_j).

Among them:

- Regularization methods. E.g. $\hat{\Sigma}(\lambda) = \hat{\Sigma} + \lambda I$.

- Graphical modelling (discussed later during the course).

Linear Discriminant Analysis uses Bayes's theorem to estimate the probabilities.

Assume we have three classes class 0 , class 1 , class 2 in the data set.

Step 1. LDA calculates the prior probabilities of each of the classes $P(y=0), P(y=1), P(y=2)$ of the data set.

Step 2. Consider an observation x.

$P(x \mid y=0), P(x \mid y=1), P(x \mid y=2)$ represent the likelihood functions.

Step 3. Calculates the Posterior probabilities to make predictions.

$$P(y=0 \mid x) = P(x \mid y=0) * P(y=0)/P(x)$$
$$P(y=1 \mid x) = P(x \mid y=1) * P(y=1)/P(x)$$
$$P(y=2 \mid x) = P(x \mid y=2) * P(y=2)/P(x)$$

In general: $P(y=y_i \mid x) = P(x \mid y=y_i) * P(y_i)/P(x)$

The posterior probability = Likelihood*Prior/Evidence.

LDA vs PCA

Example: differentiate drug performance based on gene X and gene Y

Reducing a 2-D graph to a 1-D graph

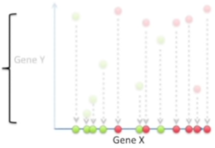


Figure: Bad option would be to ignore Gene Y

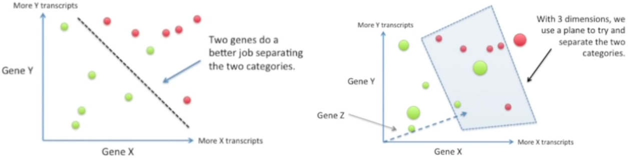


Figure: Using 2 or 3 genes to decide

LDA vs PCA

4 or more genes:

Same problems as with PCA:

1. Dimensionality Reduction.

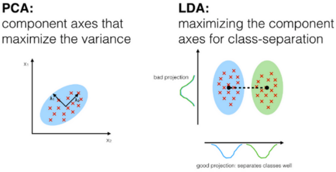
2. Visualize Classes.

- PCA reduces dimensions by focusing on the genes with the most variation.

- PCA is useful for plotting data with a lot of dimensions (or a lot of genes) onto a simple X/Y plot.

NOT interested in the genes with the most variation, BUT in maximizing the seperatibility between the two groups so we can make the best decisions.

- Linear Discriminant Analsys (LDA) is like PCA, but it focuses on maximzing the seperatibility among known categories.



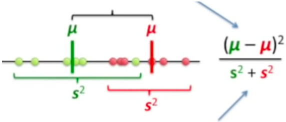
LDA vs PCA

LDA vs PCA

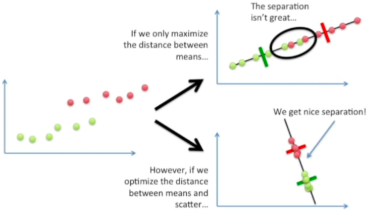
How LDA creates a new axis

The new axis is created according to two criteria (considered simultaneously):

1) Maximize the distance between means.



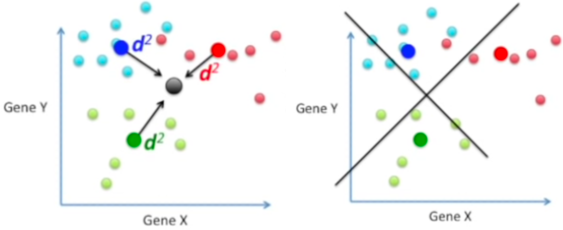
2) Minimize the variation (which LDA calls "scatter" and is represented by s^2) within each category.



LDA for 3 categories

Criteria for maximize the distance between each category and the central point while minimizing the scatter for each category:

$$\frac{d_1^2 + d_2^2 + d_3^2}{s_1^2 + s_2^2 + s_3^2}$$



LDA creates 2 axes to separate the data: the 3 central points for each category define a plane => two lines to optimize separation.

– for 2 genes? Plot doesn't change much.

– for 10,000 genes? 2 axes that maximize the separation of three categories.

Assumptions of Linear Discriminant Analysis

1) Multivariate normality within groups

2) Homoscedascity: the independent variables have equal variances and covariances across all the categories.

This assumption helps the Linear Discriminant Analysis to create the linear decision boundary between the categories.

When this assumption fails => Quadratic Discriminant Analysis: the mathematical function which separates the categories will now be quadratic.

Check: Bartlett's test

3) No multicollinearity

The performance of prediction can decrease with the increased correlation between the independent variables.

Note: Studies show that LDA is robust to slight violations of these assumptions.

4) Linearity among all pairs of variables.

Applications of LDA

1. LDA is most widely used in pattern recognition tasks. For example, analyzing customer behavior patterns based on attributes.

2. Image recognition, LDA can distinguish between categories. For example Faces and not faces, objects and not objects.

3. In the medical field, To classify patients into different groups based on symptoms for a particular disease.

Discriminant analysis can be used to answer following Market Research questions

1. In terms of demographic characteristics, how do customers who exhibit bank loyalty differ from those who do not?

2. Do heavy, medium and light users of bottled beer differ in terms of their consumption of frozen foods?

3. What psychographic characteristics help differentiate between price-sensitive and nonprice-sensitive buyers of electronic equipment?

4. What are the distinguishing characteristics of consumers who respond to direct mail solicitations?

Modeling approach

The main objectives of discriminant analysis are to:

1. develop linear combinations of the predictor variables

2. test the existence of significant differences among the groups in terms of the predictor variables

3. identify the predictor variables which contribute most to the inter-group differences

4. classify cases to one of the groups based on the values of the predictor variables

5. evaluate the accuracy of classification.

In LDA we take a different approach and impose grouping on the objects with a view to asking one or or more of three key questions:

1. Is it possible to differentiate among groups on the basis of differences in a set of shared variables

2. Which variables are most important in differentiating among groups?

3. What are the chances that the observed differences are not due to random variation?

Modeling approach

Goal: derive a variate, the linear combination of two (or more) independent variables that will discriminate best between a-priori defined groups.

Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance.

The linear combination for a discriminant analysis, also known as the discriminant function, is derived from an equation that takes the following form:

$$Z_{ik} = b_{0i} + b_{1i}X_{1k} + \ldots + b_{ji}X_{Jk}$$

Z_{ik} ... discriminant score of discriminant function i for object k

$Z_{ik} = b_{0i} + b_{1i}X_{1k} + \ldots + b_{ji}X_{Jk}$

Z_{ik} ... discriminant score of discriminant function i for object $k, i = 1, \ldots, G-1$

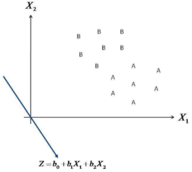
X_{jk} ... independent variable j for object $k, j = 1, 2, \ldots, J$

b_{ji} ... discriminant weight for independent variable j and discriminant function i

b_{0i} ... constant of discriminant function i

Note: Different kinds of specifications for DA functions are available.

Modeling approach



Estimation of the DA function(s)

$$\max_b \gamma = \frac{\sum_{g=1}^{G} I_g \left(\bar{Z}_g - \bar{Z}\right)^2}{\sum_{g=1}^{G} \sum_{i=1}^{I_g} \left(\bar{Z}_{gi} - \bar{Z}_g\right)^2} = \frac{SS_b}{SS_w}$$

γ – discriminant criteria (eigen value)

I_g – size of group g

Z_{gi} – i-th discriminant value of group g

SS_b – sum of squared deviations between groups, explained deviation

SS_w – sum of squared deviations within groups, remaining/unexplained deviations

Note: in this sence LDA is an extension of MANOVA.

Modeling approach

Wilks' lambda evaluates the statistical significance of the discriminatory power of the discriminant function.

As γ gives the maximal value of the discriminant criteria, a high value of γ indicates high quality. However, γ has no upper limit. Therefore appropriate transformations of γ are used:

$$\frac{\gamma}{1+\gamma} = \frac{SS_b}{SS_b + SS_w} = \frac{\text{explained variation}}{\text{total variation}}$$
$$\frac{1}{1+\gamma} = \frac{SS_w}{SS_b + SS_w} = \frac{\text{unexplained variation}}{\text{total variation}}$$

$\sqrt{\frac{\gamma}{1+\gamma}}$ – canonical correlation coefficient c.

$\frac{1}{1+\gamma} \in [0; 1]$ is called Wilks' Lambda Λ.

Λ also shows whether the group means are equal.

– Large values (near 1) ndicate that the group means may be similar.

– Small values (near 0) indicate that the group means may be different.

Modeling approach

Wilks' Lambda (Λ) is an inverse quality criterium.

If K DA functions are computed the characteristics γ, c, and Λ are computed separately for each DA function.

In order to analyze the dissimilarity of the groups multivariate Wilks' Lambda is calculated:

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1 + \gamma_k}$$

γ_k denotes the eigen value of the k-th DA function.

Modeling approach

A suitable transformation of Λ allows a significance test regarding the DA function:

$$\chi_B^2 = -\left(N - \frac{J+G}{2} - 1\right)\ln(\Lambda) \sim \chi^2_{J\cdot(G-1)}$$

N – number of observation units

J – number of variables

G – number of groups

Λ – multivariate Wilks' Lambda

Hypothesis:

H_0 : The groups are not different from each other.

H_1 : There are different groups.

Modeling approach

In order to test whether an additional DA function is necessary given k DA functions are already estimated Wilks' Lambda for the residual discriminant value can be used:

$$\Lambda_k = \prod_{q=k+1}^{K} \frac{1}{1+\gamma_q} \sim \chi^2_{(J-k)\cdot(G-k-1)}$$

Note:

– If one or more functions are deemed not statistically significant, the discriminant model should be re-estimated with the number of functions to be derived limited to the number of significant functions.

– Assessment of predictive accuracy and the interpretation of the discriminant functions will be based only on significant functions.

– It can be useful not to use all significant DA functions. Generally, 2-3 DA functions are sufficient.

Assessing Group Membership Prediction Accuracy

Hit Ratio – the correctly classified observation units divided by the number of observation units.

$$HR = \frac{c_{11} + c_{22}}{c_{11} + c_{12} + c_{21} + c_{22}}$$

The results are summarized in a classification matrix:

true class membership	predicted class membership	
	Group A	Group B
Group A	C_{11}	C_{12}
Group B	C_{21}	C_{22}

- The hit ratio must be compared with the a-priori hit ratio or a random assignment.

Modeling approach

Importance of the Independent Variables

– If the discriminant function is statistically significant

– and the classification accuracy is acceptable,

the focus lies on making substantive interpretations of the findings.

Three methods of determining the relative importance have been proposed:

(1) Standardized discriminant weights.

(2) Discriminant loadings (structure correlation).

(3) Partial F-values.

Modeling approach

(1) Standardized discriminant weights.

Goal: examine the sign and magnitude of the standardized discriminant weight (discriminant coefficient) assigned to each variable in computing the discriminant functions.

$$b_j^* = b_j \cdot s_j$$

b_j ... discriminant coefficient of variable j s_j^2 ... within group variance of variable j

– When the sign is ignored, each weight represents the relative contribution of its associated variable to that function.

– Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights.

For more than one DA function mean standardized discriminant weight for each variable is calculated:

$$\bar{b}_j = \sum_{k=1}^{K} \left| b_{jk}^* \right| \cdot EP_k$$

b_{jk}^* ... standardized discriminant weight for variable j and discriminant function k
EP_k ... Eigenvalue proportion of DA function k

Modeling approach

(2) Discriminant loadings (structure correlation) – measures of simple correlations between each predictor variable and the discriminant function(s).

– represents the variance which the predictor variable shares with the discriminant function(s)

– assessing the relative contribution of each independent variable to the DA function.

(3) Partial F-values – the absolute sizes of the significant F values.

Large F values indicate greater discriminatory power .

– In practice, rankings using the F-values approach are the same as the ranking derived from using discriminant weights, but the F values indicate the associated level of significance for each variable.

Summary

1) LDA vs Logit

The model of LDA satisfies the assumption of the linear logistic model.

The difference

a) Linear logistic regression and LDA is that the linear logistic model only specifies the conditional distribution $\Pr(G = k \mid X = x)$.

No assumption is made about $\Pr(X)$; while the LDA model specifies the joint distribution of X and $G \cdot Pr(X)$ is a mixture of Gaussians.

b) LR is solved by maximizing the conditional likelihood of G given $X : \Pr(G = k \mid X = x)$; while LDA maximizes the joint likelihood of G and $X : \Pr(X = x, G = k)$.

2) LDA vs QDA

When assumption about within-group Homogeneity fails => Quadratic Discriminant Analysis: the mathematical function which separates the categories will now be quadratic.

Summary

3) LDA vs PCA

Both allow

1. Dimensionality Reduction.

2. Visualize Classes.

Difference

- PCA reduces dimensions by focusing on the components with the most variation

- LDA maximizing the seperatibility between the two groups so we can make the best decisions.

- Recall: CFA focuses on interpretability

4) Modelling approach

- Relation to MANOVA via maximization of between to within ratio

- Analysis steps

1. Evaluate the statistical significance with **Wilks' lambda**

2. Assess Prediction Accuracy with Hit Ratio

3. Determine the relative importance of factors: Standardized discriminant weights, Discriminant loadings (structure correlation), Partial F-values.