Machine learning 1

Linear and quadratic discriminant analysis

Piotr Wójcik
pwojcik@wne.uw.edu.pl

academic year 2017/2018

UNIWERSYTET WARSZAWSKI

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Linear discriminant analysis – motivation

- the objective of linear discriminant analysis (**LDA**) is to find the function that mostly differentiates the groups of observations in terms of the **average value** of a variable (or many variables)
- if the means of a variable are significantly different between groups of observations, then we can say that this variable **discriminates** these groups
- when the groups in the data are well separated, parameter estimates of logistic regression will be unstable – in LDA this problem does not occur
- LDA is also very popular when there are more than two groups in the data

Fisher's linear discriminant analysis

- for two groups LDA is very similar to linear regression
- LDA for two groups is also called **Fisher linear discriminant discriminant**
- if groups are coded in the data as values 1 and 2, the use of linear regression will give analogous results like LDA
- in case of two groups, we estimate a **discriminant function**, which can be written with the following linear equation:

$$group = \beta_0 + \beta_2 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- interpretation of the results is analogous to linear regression
- variables that have the highest **standardized** regression coefficients best **discriminate** between groups

Linear discriminant analysis – more groups

- generally for k groups one needs to estimate $k-1$ **discriminant functions**, e.g.:
  - 1. discriminant function between group 1 and other groups $2-k$
  - 2. discriminant function between group 2 and other groups $3-k$
  - 3. ...
  - $k-1$. discriminant function between group $k-1$ and k
- coefficients in these discriminant functions can be interpreted in the same way as for Fisher's LDA

Linear discriminant analysis – classification

- the purpose of applying the LDA is usually not only to indicate which variables discriminate between different groups, but also **predicting classification into groups**
- here one uses **classification functions**, which should not be confused with discriminant functions
- there are as many **classification functions** as many as groups in the data
- classification functions allow to calculate **classification values** (**discriminant scores**) for **each observation in each group**

LDA – classification functions

- for a single explanatory variable X **classification function** is given by the formula:

$$\hat{\delta}_k(x_i) = x_i \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k)$$

- where
  - $\hat{\delta}_k(x)$ is the **discriminant score** for observation i, for which variable $X = x_i$ and group k
  - $\hat{\mu}_k$ is the average value of X in group k
  - $\hat{\sigma}^2$ is the weighted average of group variances for all k groups
  - $\hat{\pi}_k$ is the *a-priori* probability, that observation belongs to group k – one can assume $1/k$ or take empirical frequency observed for group k

LDA – classification functions – cont'd

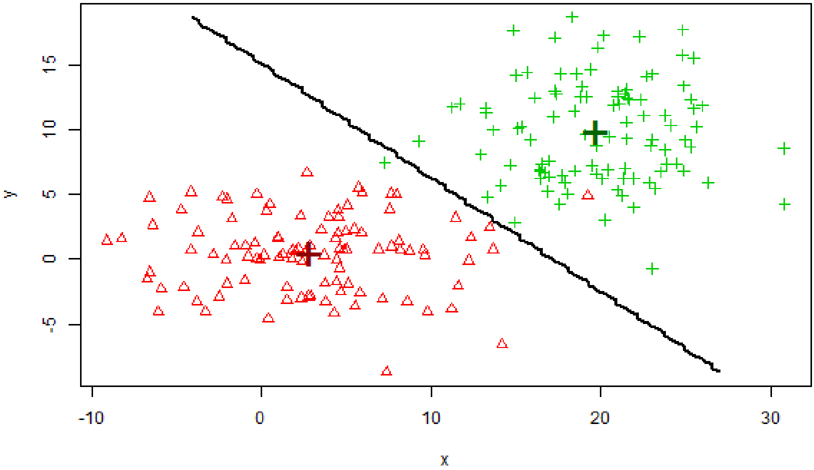
- with a larger number of explanatory variables, the LDA assumes that in each of the k groups they come from a **multivariate normal distribution** with the μ_k mean vector and a variance-covariance matrix Σ (**identical** for all groups): $N(\mu_k, \Sigma)$
- **discriminant score** is then calculated as:

$$\hat{\delta_k}(x_i) = x_i^T \Sigma^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \Sigma^{-1} - \hat{\mu}_k + log(\hat{\pi}_k)$$

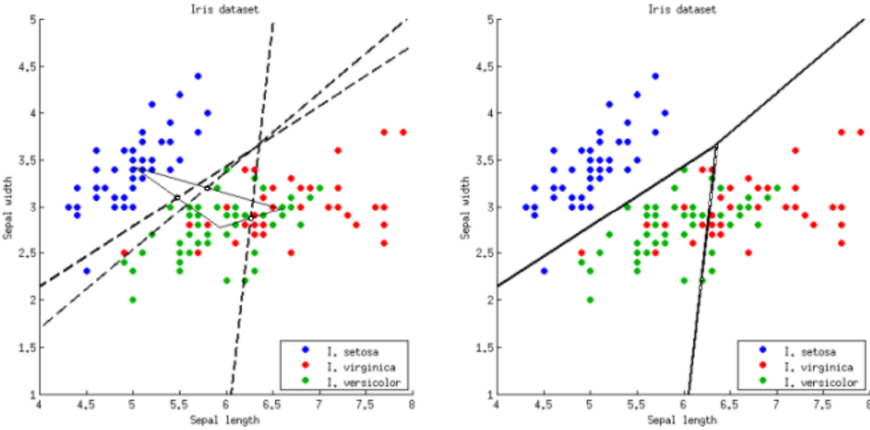
LDA – classification functions and decision boundaries

- **discriminant score** of each observation i is calculated for each of k groups
- the observation is finally classified into the group for which the **score is the highest**
- for **each pair of groups** one can set **linear decision boundary** between them
- in LDA for any two groups i and j , the decision boundary will be **a straight line** passing through a point lying halfway between the centroids of both groups $(\hat{\mu}_i + \hat{\mu}_j)/2$ and perpendicular to $\Sigma^{-1}(\hat{\mu}_i - \hat{\mu}_j)$
- for more than two groups **all** decision boundaries will intersect **at the same point** and on their basis **linear boundaries** can be determined between particular groups

Sample linear decision boundaries – 2 groups



Sample linear decision boundaries – 3 groups



Quadratic discriminant function

- LDA assumes **identical** variance-covariance matrix in all k groups, which is a **quite restrictive** assumption
- quadratic discriminant analysis (**QDA**) also assumes multivariate normal distribution, but with **different variance-covariance matrices** in different groups
- in other words – we release the assumption that individual variables have the same variances and mutual correlations in each group
- in addition, in the quadratic discriminant analysis, discriminant functions are **nonlinear** – they take into account relations of the second order (quadratic)
- also **decision boundaries** separating individual groups are functions of the second order – they have **parabolic shape**

Quadratic discriminant function

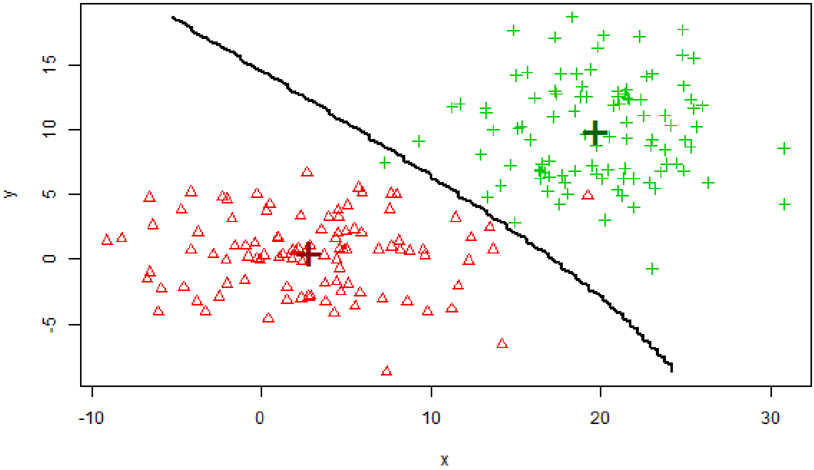
- from a mathematical point of view **QDA** assumes that observation from k-th group comes from a multivariate normal distribution with an average of μ_k and a variance-covariance matrix Σ_k
- **classification function** for observation i and group k in this case has the following form:

$$\hat{\delta}_k(x_i) = -\frac{1}{2}{x_i}^T\Sigma_k^{-1}x_i + {x_i}^T\Sigma_k^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_i^T\Sigma_k^{-1}\hat{\mu}_k - \frac{1}{2}log|\Sigma_k| + log(\hat{\pi}_k)$$

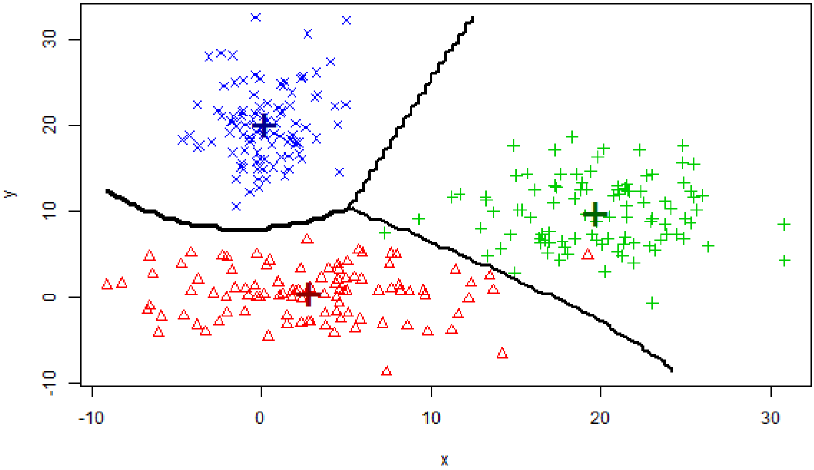
(symbols have the same meaning as discussed earlier)

- similar to LDA, the observation is finally classified into the group for which the **discriminant score is the highest**

Sample quadratic decision boundaries – 2 groups



Sample quadratic decision boundaries – 3 groups



a priori probabilities

- when classifying the discriminant analysis takes into account ***a priori* probabilities of belonging to groups**
- the simplest assumption could be that these probabilities are equal to **empirical frequencies** observed for groups – usually analyzed groups do not have equal frequencies
- this assumption will be incorrect if the distribution of groups in the sample **does not correspond** to the structure of the population (might be the result of the imperfect sample selection procedure)
- *a-priori* probabilities have an effect on classification results – if we have additional knowledge about them, **this should be included in the analysis**
- if we are not sure if the distribution in the sample is a reliable reflection of the real probabilities, you can assume **equal *a priori* probabilities** for all groups $(1/k)$

Summary

- LDA is a method similar to logistic regression
- both methods often give similar classification results
- LDA, however, has **more restrictive assumptions**: normality of variables distribution and homogeneity of variance in all groups
- if these assumptions are (approximately) fulfilled, LDA can give better results than logistic regression
- in the case of failure to meet these assumptions, logistic regression may be better
- QDA is less restrictive and allows to distinguish between groups well if the boundaries between them are not linear
- in small samples LDA and QDA will often give better results than logistic regression
- LDA and QDA are easy to generalize to a situation with more than two groups

Thank you for your attention