Machine learning and statistical learning

3. Classification

Ewen Gallic
ewen.gallic@gmail.com

MASTER in Economics - Track EBDS - 2nd Year

2020-2021

This chapter discusses a second category of supervised learning methods: **classification**.

In the regresison case, the response variable y was real-valued.

We consider now the case in which it is **qualitative** (or **categorical**).

We will cover three classifiers in this chapter:

- the logistic regression
- the k-nearest neighbors
- the linear discriminant analysis

Note that trees, random forests, bagging and boosting will be covered in Pierre Michel's course.

Some references

- Berk (2008). Statistical learning from a regression perspective, volume 14. Springer.
- James et al. (2013). An introduction to statistical learning, volume 112. Springer.
- Scott Long (1997). Regression models for categorical and limited dependent variables. Advancedquantitative techniques in the social sciences, 7.

1. Introduction

Introduction

In classification problems, we want to assign a class to a **quantitative** response.

We saw in the introduction of the course that it may be a good idea to estimate a probability for each of the categories and then to assign the class, based on the mode, for example.

Why not fitting a linear regression

In the case in which the response variable is categorical, the linear regression may not be appropriate.

Let us consider a simple case in which we are trying to predict the occupation of individuals, based on their characteristics. Let us suppose there are three classes for the response variable:

The response variable could be:

$$y = \begin{cases} 1 & \text{blue-collar jobs} \\ 2 & \text{white-collar jobs} \\ 3 & \text{professional jobs} \end{cases}$$

Why not fitting a linear regression

Technically, it is possible to fit a linear model using this response variable. But this implies :

- an **ordering** on the outcome, putting white-collar jobs between blue-collar and professional jobs
- the difference between `blue-collar` and `white-collar` is the same as the difference between `white-collar` and `professional`.

We could code the response variable differently, and the resutls would be completely differents.

Why not fitting a linear regression

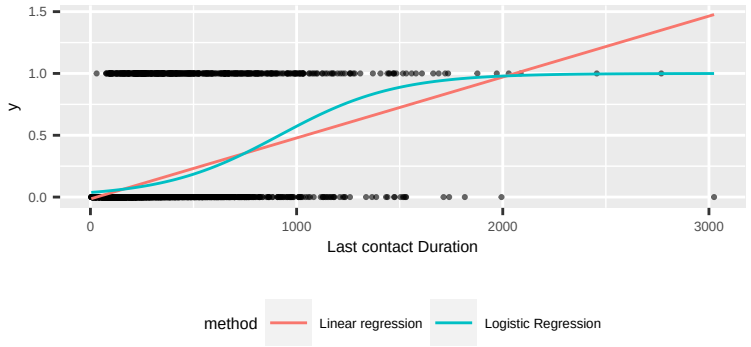
However, if the response variable is categorical, but the variables **can be ordered**, and **if we think that the gap between each category is similar**, a linear regression can be envisaged. For example, if the response variable is some age class:

- [18-25]
- [26-35]
- [36-45]
- [46-55]

If the **response variable is binary**, fitting a linear regression is less problematic. If our response variable is coded using 0 and 1 values:

- as we saw in the previous chapter, flipping the coding of the variable does not change the prediction
- but the prediction may lie outside the $[0, 1]$ interval, making them hard to interpret as probabilities.

Why not fitting a linear regression



2. Logistic regression

Dataset

Let us consider some data on which we will build some examples: Bank Marketing Data Set

Source : Moro et al. (2014) A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31.

To download the data: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Dataset

The data concerns results from direct marketing campaigns (phone calls) of a Portuguese banking institution.

The aim is to predict whether the client will subscribe from a term deposit (variable y).

The original dataset has $45,211$ training observations. We will use the provided random subset using only only 10% of observations here.

Dataset

- bank client data:
  - age (numeric)
  - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
  - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
  - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course', 'university.degree','unknown')
  - default: has credit in default? (categorical: 'no','yes','unknown')
  - balance: average yearly balance, in euros (numeric)
  - housing: has housing loan? (categorical: 'no','yes','unknown')
  - loan: has personal loan? (categorical: 'no','yes','unknown')

Dataset

- related with the last contact of the current campaign:
  - contact: contact communication type (categorical: 'cellular','telephone')
  - day: last contact day of the month (numeric: 1 to 31)
  - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
  - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- other attributes:
  - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
  - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
  - previous: number of contacts performed before this campaign and for this client (numeric)
  - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

Dataset

The response variable is coded as follows:

$$y = \begin{cases} 1 & \text{if the client has subscribed to a term deposit} \\ 0 & \text{if the client has not subscribed to a term deposit} \end{cases}$$

4000 (88.48%) clients have not subscribed while 521 (11.52%) have.

Logistic regression

Let us consider a simple example in which we try to model the class of the response variable y using the variable `balance` as the sole predictor (which corresponds to the average yearly balance, in euros).

The logistic regression models the **probability** that the client will subscribe to a long term deposit or not. The probability of subscribing given `balance` writes:

$$\mathbb{P}(y = \text{yes} \mid balance)$$

We can rely on this probability to assign a class (yes or no) to the observation.

- For example, we can assign the class `yes` for all observations where $\mathbb{P}(y = \text{yes} \mid balance) > 0$.
- But we can also select a **different threshold**.

2.1 Model

2.1.1 The Linear Probability Model

A little detour

Before turning to the logistic model, let us a little detour to see what happens when we fit a linear model to the relationship between our predictor $\mathbf{x}$ and our response variable y.

The structural model writes, for $in \in 1, \ldots, n$:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i,$$

where ε_i is an error term with zero mean and variance σ^2.

If the predictor is a binary outcome, the model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

The conditional expectation of y given x is:

$$\mathbb{E}(y \mid x) = \beta_0 + \beta_1 x.$$

Linear regression model for binary outcome

As we consider the case in which y is a binary variable, the **unconditional expectation** of y is the probability that the event occurs:

$$\mathbb{E}(y_i) = [1 \times \mathbb{P}(y_i = 1)] + [0 \times \mathbb{P}(y_i = 0)] = \mathbb{P}(y_i = 1).$$

For the regression model:

$$\mathbb{E}(y_i \mid \mathbf{x}_i) = [1 \times \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)] + [0 \times \mathbb{P}(y_i = 0 \mid \mathbf{x}_i)] = \mathbb{P}(y_i = 1 \mid \mathbf{x}_i).$$

The expected value of y given $\mathbf{x}$ is the probability that $y = 1$ given $\mathbf{x}$. We therefore have:

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$$

Problems with the linear regression model and binary outcomes

Multiple problems arise when fitting a linear regression model to a binary outcome response.

First, let us consider the issue regarding **heteroscedasticity**.

If y is a binary outcome random variable, with mean μ, then:

$$\mathbb{E}(Y) = 0 \cdot \mathbb{P}(Y = 0) + 1 \cdot \mathbb{P}(Y = 1) \Rightarrow \mathbb{P}(Y = 1) = \mu.$$

As $\mathbb{E}(Y^2) = 0^2 \cdot \mathbb{P}(Y = 0) + 1^2 \cdot \mathbb{P}(Y = 1) = \mathbb{P}(Y = 1)$, the variance of Y writes:

$$\mathbb{V}ar(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \mu(1 - \mu).$$

Problems with the linear regression model and binary outcomes

We know that $\mathbb{E}(y \mid \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, hence the conditional variance of y depends on $\mathbf{x}$:

$$\mathbb{V}ar(y \mid \mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x})\left[1 - \mathbb{P}(y = 1 \mid \mathbf{x})\right] = \mathbf{x}\boldsymbol{\beta}(1 - \mathbf{x}\boldsymbol{\beta}).$$

Hence, the variance of the errors depends on the values of $\mathbf{x}$ and is not constant.

As a consequence, the least squares estimate of $\boldsymbol{\beta}$ is inneficient and the standard errors are biased.

Problems with the linear regression model and binary outcomes

Let us now turn to the **normality assumption**.

As the response variable only takes values between 0 and 1, the error at point x_0 must be equal to:

$$\begin{cases} 1 - \mathbb{E}(y \mid x_0) & \text{if } y = 1 \\ 0 - \mathbb{E}(y \mid x_0) & \text{if } y = 0 \end{cases}$$

Hence, the errors cannot be normally distributed, which leads to biased estimates.

Another problem arises, as we saw previously, that of **nonsensical predictions**: some predictions may be negative or greater than one.

This is problematic, as we are predicting a probability.

2.1.2 A latent model for binary variables

A latent model for binary variables

To overcome the mentioned issues, $\mathbb{P}(y \mid \mathbf{x})$ must be modeled using a function that gives outputs between 0 and 1 for all values of $\mathbf{x}$.

Let us suppose there is an **unobserved** (or **latent**) variable $y^\star$ ranging from $-\infty$ to $+\infty$ that generates the observed values of y:

- the large values of $y^\star$ are observed as 1
- the small values of $y^\star$ are observed as 0.

The latent variable

For example, let us consider the clients who subscribe to a term deposit (the observed response variable y). This variable can be observed in two states:

- the client has subscribed
- the client has not subscribed.

Some clients may be closed to the decision of subscribing, other may be very far from it. In both cases, we observe $y = 0$.

The idead behind the **latent** $y^\star$ is that there exists an underlying **propensity to subscribe** that generates the observed state.

We **cannot observe** $y^\star$, but at some point, a **change in** $y^\star$ **results in a change in what we observe**:

- for example, as the balance increases, the propensity to subscribe to a term deposit increases too.

The latent variable

The latent $y^\star$ is assumed to be linearly related to the observed predictors:

$$y_i^\star = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i.$$

The observed variable y writes:

$$y_i = \begin{cases} 1 & \text{if } y_i^\star > \tau \\ 0 & \text{if } y_i^\star \leq \tau \end{cases} \tag{1}$$

where τ is called the **threshold** (or **cutpoint**), assumed to be equal to zero for now.

Estimating the model: intuition

The dependant variable is not observed. The model cannot be estimated using least squares.

We will see in a next section how to estimate the model using **maximum likelihood** (ML).

The ML estimation requires some assumptions about the distribution of the errors ε.

For the **logit model**, we assume that the **distribution of the errors is logistic**.

We also assume that:

- $\mathbb{E}(\varepsilon \mid \mathbf{x}) = 0$
- $\mathbb{V}ar(\varepsilon \mid \mathbf{x}) = \pi^2/3$

Note: if we assume the errors are $\mathcal{N}(0, 1)$, it results in the probit model.

Standard logistic distribution

The probability distribution function of the standard logistic distribution writes:

$$\lambda(\varepsilon) = \frac{\exp(\varepsilon)}{[1+\exp(\varepsilon)]^2}.$$

The cumulative distribution function writes:

$$\Lambda(\varepsilon) = \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}$$

Standard logistic distribution

The logistic distribution can be rescaled to have a unit variance, resulting in the **standardized logistic distribution**.

The probability distribution becomes:

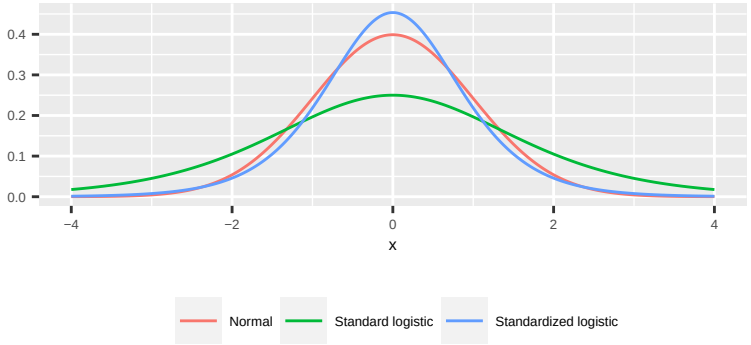
$$\lambda^S(\varepsilon) = \frac{\gamma \exp(\gamma\varepsilon)}{[1+\exp(\gamma\varepsilon)]^2}.$$

The cumulative distribution function writes:

$$\Lambda^S(\varepsilon) = \frac{\exp(\gamma\varepsilon)}{1+\exp(\gamma\varepsilon)},$$

where $\gamma = \pi/\sqrt{3}$

Standard logistic distribution



Estimating the model: intuition

Assuming the distribution of the errors is useful to compute the probability of $y = 1$ for a given $\mathbf{x}$.

Indeed:

$$\begin{aligned}\mathbb{P}(y = 1 \mid \mathbf{x}) &= \mathbb{P}(y^\star > 0 \mid \mathbf{x}) \\ &= \mathbb{P}(\mathbf{x}\boldsymbol{\beta} + \varepsilon > 0 \mid \mathbf{x}) \\ &= \mathbb{P}(\varepsilon > -\mathbf{x}\boldsymbol{\beta} \mid \mathbf{x}) \\ &\mathbb{P}(\varepsilon \leq \mathbf{x}\boldsymbol{\beta} \mid \mathbf{x}) \quad \text{(symmetric cdf)} \\ &= F(\mathbf{x}\boldsymbol{\beta}) = \Lambda(\mathbf{x}\boldsymbol{\beta}).\end{aligned}$$

Hence the probability of observing an event given $\mathbf{x}$ is the cummulative density evaluated at $\mathbf{x}\boldsymbol{\beta}$.

Deriving the model without appealing to a latent variable

Another way of deriving the logit model without appealing to a latent variable is to specify a nonlinear model relating the predictors to the probability of an event.

In a first step, the probability of an event is **transformed into the odds**:

$$\frac{\mathbb{P}(y=1 \mid \mathbf{x})}{\mathbb{P}(y=0 \mid \mathbf{x})} = \frac{\mathbb{P}(y=1 \mid \mathbf{x})}{1-\mathbb{P}(y=1 \mid \mathbf{x})}$$

The **odds** are used to evaluate how often an event happens relative to how often is does not.

The odds range from 0 (when $\mathbb{P}(y=1 \mid \mathbf{x})=0$) to ∞ (when $\mathbb{P}(y=1 \mid \mathbf{x})=1$).

Deriving the model without appealing to a latent variable

The log of the odds, as known as the **logit** ranges from $-\infty$ to ∞.

This suggests a model that is linear in the logit:

$$\ln\left[\frac{\mathbb{P}(y=1 \mid \mathbf{x})}{1-\mathbb{P}(y=1 \mid \mathbf{x})}\right] = \mathbf{x}\boldsymbol{\beta} \tag{2}$$

Hence, a marginal change in $\mathbf{x}$ changes the log odds by the coefficient β associated with $\mathbf{x}$.

We can easily show that Eq. 2 equivalent to the **logistic function**:

$$\mathbb{P}(y=1 \mid \mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1+\exp(\mathbf{x}\boldsymbol{\beta})}. \tag{3}$$

Deriving the model without appealing to a latent variable

From Eq. 3, it can easily be shown that:

$$\frac{\mathbb{P}(y=1 \mid \mathbf{x})}{1-\mathbb{P}(y=1 \mid \mathbf{x})} = \exp(\mathbf{x}\boldsymbol{\beta}) \tag{4}$$

Hence, we see that a marginal increase in $\mathbf{x}$ can be equivalently interpreted as follows: it multiplies the odds by the coefficient β associated with $\mathbf{x}$

2.2 Estimation

Estimation

In the case of a binary response variable and the logit model, we wan to estimate the coefficients β_0 and β_1 from the logistic function:

$$\mathbb{P}(y = 1 \mid \mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 \mathbf{x})}{1 + \exp(\beta_0 + \beta_1 \mathbf{x})}.$$

First, let us define p as the probability of observing the value of y:

$$p_i = \begin{cases} \mathbb{P}(y_i = 1 \mid \mathbf{x}_i) & \text{if } y_i = 1 \text{ is observed} \\ 1 - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i) & \text{if } y_i = 0 \text{ is observed} \end{cases} \tag{5}$$

Assuming the observations are independant, the likelihood equation writes:

$$\mathcal{L}(\boldsymbol{\beta} \mid y, \mathbf{x}) = \prod_{i=1}^{n} p_i \tag{6}$$

Estimation

Hence, combining Eq. 5 and Eq. 6:

$$\mathcal{L}(\boldsymbol{\beta} \mid y, \mathbf{x}) = \prod_{y=1} \mathbb{P}(y_i = 1 \mid \mathbf{x}_i) \prod_{y=0} [1 - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)] \tag{7}$$

We have seen that $\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i\boldsymbol{\beta})$, *i.e.*, $\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \Lambda(\mathbf{x}_i\boldsymbol{\beta})$ in the case of the logit model.

Hence, the likelihood equations also writes:

$$\mathcal{L}(\boldsymbol{\beta} \mid y, \mathbf{x}) = \prod_{y=1} \Lambda(\mathbf{x}_i\boldsymbol{\beta}) \prod_{y=0} [1 - \Lambda(\mathbf{x}_i\boldsymbol{\beta})] \tag{8}$$

Estimation

The log likelihood thus writes:

$$\ell(\boldsymbol{\beta} \mid y, \mathbf{x}) \coloneqq \ln \mathcal{L}(\boldsymbol{\beta} \mid y, \mathbf{x}) = \sum_{y=1} \ln \Lambda(\mathbf{x}_i \boldsymbol{\beta}) + \sum_{y=0} \ln \left[1 - \Lambda(\mathbf{x}_i \boldsymbol{\beta})\right] \tag{9}$$

$$= \sum_{i=1}^{n} \mathbb{1}_{(y_i=1)} \ln \Lambda(\mathbf{x}_i \boldsymbol{\beta}) + (1 - \mathbb{1}_{y_i=1}) \ln \left[1 - \Lambda(\mathbf{x}_i \boldsymbol{\beta})\right] \tag{10}$$

The estimates of β are chosen to **maximize** the likelihood function.

Effect of a variable on the odds

Let us get back to our clients data.

	Model 1
(Intercept)	-3.25593456^{***}
	(0.08457673)
duration	0.00354955^{***}
	(0.00017136)
AIC	2705.75264185
BIC	2718.58561882
Log Likelihood	-1350.87632092
Deviance	2701.75264185
Num. obs.	4521

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Statistical models

We see that $\beta_1 = 0.00354955$:

- a unit increase in the duration of the last call is associated with an increase in the probability of subscribing to a term deposit
- a unit increase in the duration of the last call is associated with an increase in the log odds of the response by 0.00354955.

2.2.1 Numerical methods

Numerical methods

Let us look at how to use numerical methods to estimate the coefficients.

We will look at different ways of doing so, using iterative solutions.

First, we need to begin with an initial guess that we will denote $\boldsymbol{\beta}_0$: the **start values**.

At each iteration, we will try to improve the previous guess by adding a vector ζ_n of adjustments:

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + \zeta_n.$$

The iterations stops as soon as a convergence criterion is met:

- roughly, when the gradient of the log likelihood gets close to 0
- or when the estimates do not change from one step to the next.

Sometimes, there is no convergence, and we do not get the ML estimates.

Numerical methods

ζ_n can be expressed as:

$$\zeta_n = \boldsymbol{D}_n \boldsymbol{\gamma}_n,$$

- $\boldsymbol{\gamma}_n$: **gradient vector**:
  - $\boldsymbol{\gamma}_n = \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}_n}$
  - it indicates the direction of the change in the log likelihood for a change in the parameters
- $\boldsymbol{D}_n$: **direction matrix**
  - it reflects the curvature of the likelihood function (how rapidly the gradient is changing)

The Newton-Raphson Method

The rate of change in the slope of $\ln \mathcal{L}$ is indicated by the second derivatives, contained in the Hessian matrix.

In the case of a single predictor, we have two parameters to estimate, β_0 and β_1. The Hessian matrix is thus:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \begin{pmatrix} \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ln \mathcal{L}}{\partial \beta_1 \partial \beta_1} \end{pmatrix}$$

If $\frac{\partial^2 \ln \mathcal{L}}{\partial \beta_0 \partial \beta_0}$ is large relative to $\frac{\partial^2 \ln \mathcal{L}}{\partial \beta_1 \partial \beta_1}$:

- the gradient is changing more rapidly as β_0 changes than β_1 changes.

In that case, smaller adjustments to the estimates of β_0 would be indicated.

The Newton-Raphson Method

Recall that:

$$\ell(\boldsymbol{\beta} \mid y, \mathbf{x}) = \sum_{i=1}^{n} \mathbb{1}_{(y_i=1)} \ln \Lambda(\mathbf{x}_i\boldsymbol{\beta}) + (1 - \mathbb{1}_{y_i=1}) \ln \left[1 - \Lambda(\mathbf{x}_i\boldsymbol{\beta})\right]$$

Hence, it can easily be shown that, for all $j = 0, 1, \ldots, p$:

$$\frac{\partial \ell(\boldsymbol{\beta} \mid y, \mathbf{x})}{\partial \beta_j} = \sum_{i=1}^{n} \mathbf{x}_{ij} \left[\mathbb{1}_{(y_i=1)} - \Lambda(\mathbf{x}_i\boldsymbol{\beta})\right]$$

In matrix form:

$$\frac{\partial \ell(\boldsymbol{\beta} \mid y, \mathbf{x})}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_{i} \left[\mathbb{1}_{(y_i=1)} - \Lambda(\mathbf{x}_i\boldsymbol{\beta})\right]$$

The Newton-Raphson Method

The Newton-Raphson algorithm requires the Hessian matrix:

$$\frac{\partial^2 \ell(\boldsymbol{\beta} \mid y, \mathbf{x})}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \left[\Lambda(\mathbf{x}_i \boldsymbol{\beta})(1 - \Lambda(\mathbf{x}_i \boldsymbol{\beta}))\right]$$

The following equation is used by the Newton-Raphson algorithm at each iteration:

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n - \left(\frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^\top}\right)^{-1} \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}_n}$$

The Newton-Raphson Method

To illutsrate the iterative process of the Newton-Raphson algorithm, let us go through an example, using matrix notations.

- $\boldsymbol{y}$: column vector of the response variable, of dimenstion $n \times 1$
- $\boldsymbol{X}$: matrix of predictors, including the constant, of dimension $n \times (p+1)$
- $\boldsymbol{p}_n$: vector of fitted probabilities at iteration n, of dimension $n \times 1$
- $\boldsymbol{\Omega}$: diagonal matrix of weights, of dimension $n \times n$
  - the ith element of the diagonal is: $\Lambda(\boldsymbol{X}_i\boldsymbol{\beta}_{n-1})(1-\Lambda(\boldsymbol{X}_i\boldsymbol{\beta}_{n-1}))$

The Newton-Raphson Method

Using these notations, we can write the gradient and the Hessian as follows:

$$\frac{\partial \ell(\boldsymbol{\beta} \mid y, \mathbf{x})}{\partial \beta} = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}_{n-1})$$
$$\frac{\partial^2 \ell(\boldsymbol{\beta} \mid y, \mathbf{x})}{\partial \beta \partial \beta^\top} = -\boldsymbol{X}^\top \boldsymbol{\Omega}_{n-1} \boldsymbol{X}$$

The Newton-Raphson Method

At each step:

$$\begin{aligned}
\boldsymbol{\beta}_n &= \boldsymbol{\beta}_{n-1} + \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}_{n-1}) \\
&= \underbrace{\left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right)}_{I_n} \boldsymbol{\beta}_{n-1} + \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}_{n-1}) \\
&= \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \left[\boldsymbol{X} \boldsymbol{\beta}_{n-1} + \Omega^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}_{n-1})\right] \\
&= \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{z}
\end{aligned}$$

where $\boldsymbol{z} = \left[\boldsymbol{X} \boldsymbol{\beta}_{n-1} + \Omega^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}_{n-1})\right]$.

Written this way, $\boldsymbol{z}$ can be viewed as the response of the model, $\boldsymbol{X}$ the predictors and $\boldsymbol{\beta}_n$ the coefficients of a weighted least squares regression.

The Newton-Raphson Method

See Lab exercise.

2.3 Predictions

Predictions

Is is easy to use the estimates to perform some predictions.

For example, in our example of clients subscribing to term deposit:

$$y_i = \beta_0 + \beta_1 \text{duration}_i + \varepsilon_i$$

We estimated that $\hat{\beta}_0 = -3.25593456$ and $\hat{\beta}_1 = 0.00354955$.

Hence, the probability of subscribing to a term deposit for an individual whose last call duration was 250 seconds is given by:

$$\begin{aligned}\hat{\mathbb{P}}(y = 1 \mid \text{duration} = 250) &= \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \times 250\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \times 250\right)} \\ &= \frac{\exp\left(-3.25593456 + 0.00354955 \times 250\right)}{1 + \exp\left(-3.25593456 + 0.00354955 \times 250\right)} \\ &= 0.0856028\end{aligned}$$

2.4 Interpretation

2.4.1 The effect of the Parameters

Binary Regression Model

Let us consider the binary response model with a single predictor $\mathbf{x}$. The model writes:

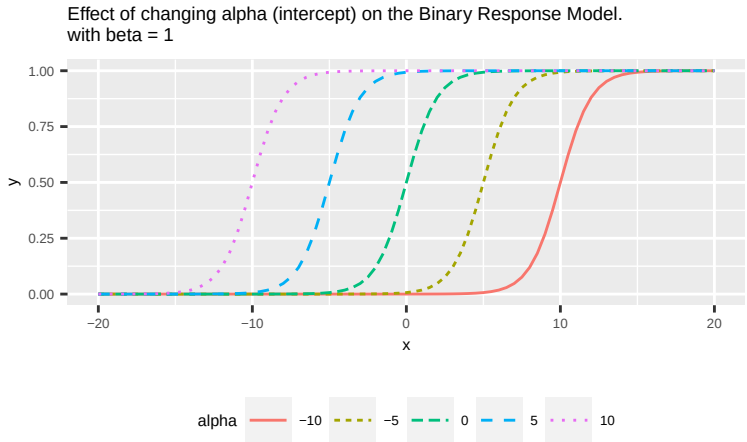
$$\mathbb{P}(y = 1 \mid \mathbf{x}) = F(\alpha + \beta\mathbf{x}),$$

i.e., for a logit model:

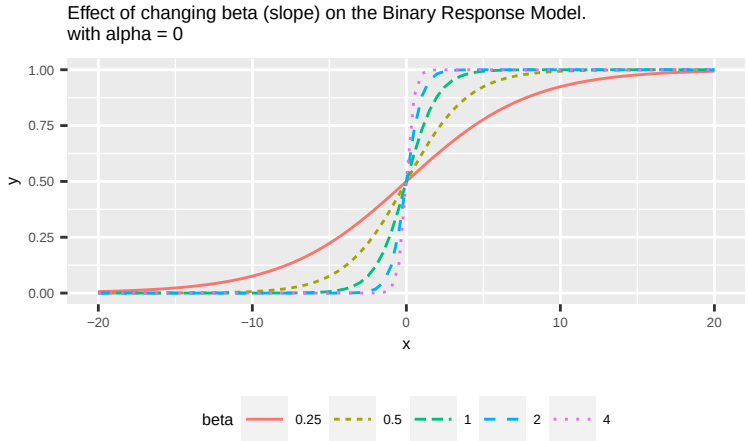
$$\mathbb{P}(y = 1 \mid \mathbf{x}) = \frac{\exp(\alpha + \beta\mathbf{x})}{1 + \exp(\alpha + \beta\mathbf{x})}$$

Let us vary α and β to observe how it affects the probability.

Varying β



Varying α



3. Binary Response Model with Multiple Predictors

Binary Response Model with Multiple Predictors

Let us now consider multiple regressors (only two, for illustrative purposes):

$$\mathbb{P}(y = 1 \mid \mathbf{x}) = F(\alpha + \beta_1 x_1 + \beta_2 x_2),$$

i.e., for a logit model:

$$\mathbb{P}(y = 1 \mid \mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}$$

Let us consider the following values:

- $\alpha = 1$
- $\beta_1 = 1$
- $\beta_2 = .75$

Let us vary α and β_1 and β_2 to observe how it affects the probability.

Varying α

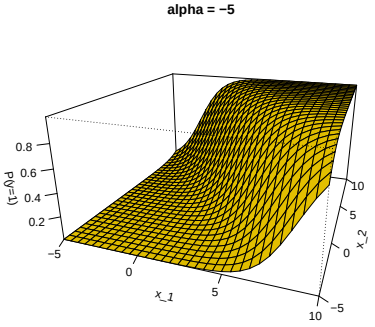


Figure 1: Effects of changing α on the Logit Model $\mathbb{P}(y = 1 \mid x_1, x_2)$, with $\beta_1 = 1$ and $\beta_2 = .75$

Varying β_1

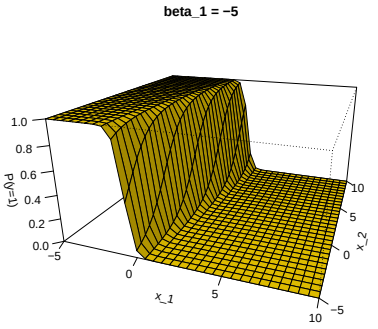


Figure 2: Effects of changing α on the Logit Model $\mathbb{P}(y = 1 \mid x_1, x_2)$, with $\alpha = 1$ and $\beta_2 = .75$

Varying β_2

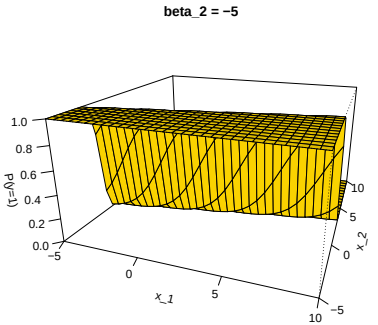


Figure 3: Effects of changing α on the Logit Model $\mathbb{P}(y = 1 \mid x_1, x_2)$, with $\alpha = 1$ and $\beta_1 = 1$

Using Predicted Probabilities

To know the effect of a covariate on the probability of the event, we can look at how this probability changes when varying a covariate.

However, where there are more than two predictors, the surface response cannot be plotted as it was done previously when facing a single covariate.

In such cases, the interpretation depends on whether or not **the relationship between the response and the predictors can be considered as linear**. To that end, we need to compute the **range of probabilities**.

Using Predicted Probabilities

The predicted probability of an event given some values for the predictors $\mathbf{x}$ for the ith individual writes:

$$\widehat{\mathbb{P}}(y = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i \widehat{\boldsymbol{\beta}})$$

The minimum and maximum probabilities **in the sample** write, respectively:

$$\min \widehat{\mathbb{P}}(y = 1 \mid \mathbf{x}_i) = \min_i F(\mathbf{x}_i \widehat{\boldsymbol{\beta}})$$

$$\max \widehat{\mathbb{P}}(y = 1 \mid \mathbf{x}_i) = \max_i F(\mathbf{x}_i \widehat{\boldsymbol{\beta}})$$

- If the range of probabilities $(\max \widehat{\mathbb{P}}(y = 1 \mid \mathbf{x}_i) - \min \widehat{\mathbb{P}}(y = 1 \mid \mathbf{x}_i))$ is **between 0.2 and 0.8**, the relationship between $\mathbf{x}_i$ and y can be considered as linear:
  - thus, the marginal effect of $\mathbf{x}_i$ can be obtained using simple measures
- Otherwise, other methods need to be used.

The effect of Each Variable on the Predicted Probability

To assess the effect of a (numerical) variable on the predicted probability, we can look at how the predicted probability changes as the variables varies **from its minimum to its maximum value**.

To do so, for the kth variable:

- the other variables are set to their average value ($\overline{\mathbf{x}}_{(-k)}$)
- we compute the predicted probability when $\mathbf{x}_k$ is at its maximum value ($\widehat{\mathbb{P}}(y = 1 \mid \overline{\mathbf{x}}_{(-k)}, \min \mathbf{x}_k)$)
- we compute the predicted probability when $\mathbf{x}_k$ is at its minimum value ($\widehat{\mathbb{P}}(y = 1 \mid \overline{\mathbf{x}}_{(-k)}, \max \mathbf{x}_k)$)

The **predicted change in the probability** as $\mathbf{x}_k$ varies from its min to its max is then computed as:

$$\widehat{\mathbb{P}}(y = 1 \mid \overline{\mathbf{x}}_{(-k)}, \min \mathbf{x}_k) - \widehat{\mathbb{P}}(y = 1 \mid \overline{\mathbf{x}}_{(-k)}, \max \mathbf{x}_k)$$

The effect of Each Variable on the Predicted Probability

If, among the $(-k)$ covariates, some are non numerical:

- it is possible to set these variables to the mode of the distribution
- or different combinations of the levels can be envisaged.

Discrete Change

Let us now consider the **partial change in** y.

Let $\mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_k)$ be the probability of an event given some values for $\mathbf{x}_{(-k)}$ and a specific value for $\mathbf{x}_k$.

Let $\mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_{k+\delta})$ be the probability of an event after a variation of δ in $\mathbf{x}_k$, keeping all other variables unchanged.

The **discrete change** in the probability for a change of δ in x_k is:

$$\mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_{k+\delta}) - \mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_k)$$

Discrete Change

$$\mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_{k+\delta}) - \mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_k)$$

Following a variation of δ in $\mathbf{x}_k$, keeping all the other variables unchanged, the predicted probability of an event is changed by $\mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_{k+\delta}) - \mathbb{P}(y = 1 \mid \mathbf{x}_{(-k)}, \mathbf{x}_k)$.

Some usual values are picked for δ:

- **a (centered) unit change**: increasing $\overline{\mathbf{x}}_k$ to $\overline{\mathbf{x}}_k + 1$
- **a standard deviation change**:
  - $$\mathbb{P}\left(y = 1 \mid \overline{\mathbf{x}}_{(-k)}, \overline{\mathbf{x}}_k + \frac{std(\mathbf{x}_k)}{2}\right) - \mathbb{P}\left(y = 1 \mid \overline{\mathbf{x}}_{(-k)}, \overline{\mathbf{x}}_k - \frac{std(\mathbf{x}_k)}{2}\right)$$
- **a change from 0 to 1 for dummy variables**

4. K-nearest neighbors

4.1 The Bayes classifier

The Bayes classifier

In a two-class problem where the response variable Y can take only two distinct values (*e.g.*, class 1 and class 2), the **Bayes classifier** corresponds to

- predicting class 1 if $\mathbb{P}(Y = 1 \mid X = x_0) \geq 0.5$ for a given observation $X = x_0$
- predicting class 2 otherwise

Formally, this corresponds to assign an observation x_0 the class k for which $\mathbb{P}(Y = k \mid X = x_0)$ is the highest.

Example

To illustrate this, let us consider an example of a response variable y that takes two values : blue and orange, depending on the values of two predictors x_1 and x_2, such that:

$$y = \begin{cases} \text{blue} & \text{if } x_1^2 + x_2^2 > 60^2 \\ \text{orange} & \text{if } x_1^2 + x_2^2 \leq 60^2 \end{cases}$$

We randomly generate $n = 100$ observations for x_1 and x_2 from a $\mathcal{U}[0, 100]$, and assign the *true class* to y. From these observations, we fit a given classifier.

Then, we randomly generate $n = 100$ new observations for x_1 and x_2 from the same uniform distribution and use the classifier to predict the probability of belonging to each of the two classes.

$$\hat{y} = \begin{cases} \text{blue} & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) \geq 0.5 \\ \text{orange} & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) < 0.5 \end{cases}$$

Example

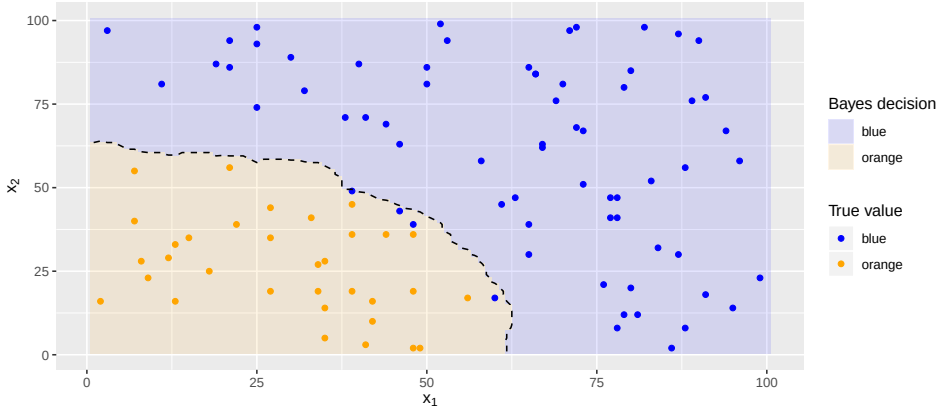


Figure 4: Bayes' decision boundary (dashed line).

Error rate

The error rate of the Bayes classifier at point $X = x_0$ will always be $1 - \max_k \mathbb{P}(Y = k \mid X = x_0)$.

The overall Bayes error rate is given by:

$$1 - \mathbb{E}(\max_k \mathbb{P}(Y = k \mid X)), \tag{11}$$

where the expectation averages the probability over all possible values of X.

4.2 K-Nearest Neighbors

K-Nearest Neighbors

When facing real data, we do not know the conditional distribution of Y given X. Hence, computing the Bayes classifier is not possible.

In this section, we look at a classifier that estimates the conditional distribution of Y given X, namely the **K-nearest neighbors** (KNN) classifier.

K-Nearest Neighbors

The basic idea of the KNN classifier is, as follows:

- from a given positive integer K and a test observation x_0, identify the K points in the training data that are **closest** to x_0, represented by $\mathcal{N}_0$
- estimate the conditional probability for class k as the fraction of points in $\mathcal{N}_0$ whose response values equal k:

$$\mathbb{P}(Y = k \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{1}_{(y_i = k)}. \tag{12}$$

- once the conditional probabilities for each of the K classes are estimated, apply Bayes rule and assign x_0 to the class with the highest probability.

Example

Let us consider again a response variable that can take two values: either “blue” or “orange”.

For the example, we draw 100 points in a unit square and randomly assign them a class.

Then, we consider a point at coordinates $(0.75, 0.5)$ and try to predict the class for this point using a KNN classifier.

We vary the number of nearest neighbors to consider: $K = \{3, 5, 10\}$.

Example

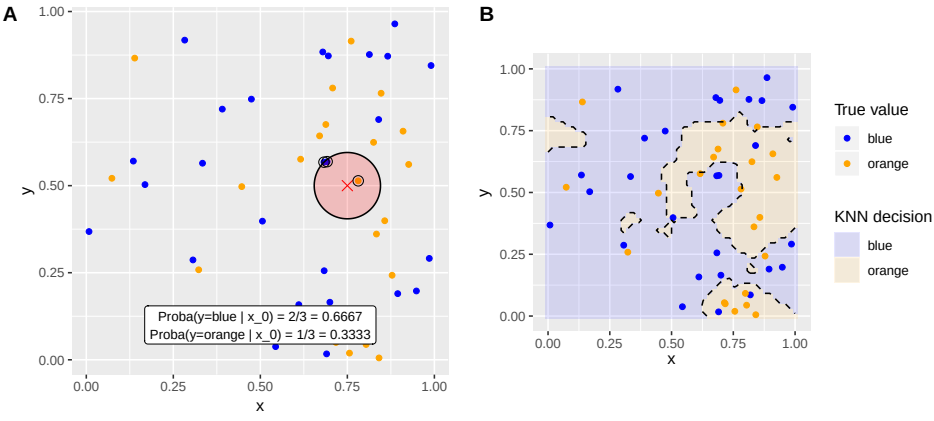


Figure 5: KNN approach, varying K.

Exercise

Lab exercise.

5. Linear Discriminant Analysis

Linear Discriminant Analysis

In the first part of this chapter, we have looked at a way of fitting $\mathbb{P}(y = k \mid \mathbf{x})$ using the logistic function, with $k = \{0, 1\}$.

In this section, we will consider a categorical response variable that can take more than two classes.

In a nutshell:

- we will model the distribution of the predictors X given y
- rely on Bayes' theorem to derive $\mathbb{P}(y = k \mid \mathbf{x})$

Bayes' theorem for classification

Let Y be a qualitative variable that can take K possible unordered values.

Let π_k be the overall **prior probability** that a randomly chosen observation comes from the kth class, $k = 1, \ldots, K$.

The **density function** of $\mathbf{x}$ for an observation from the kth class writes:

$$f_k(x) \equiv \mathbb{P}(\mathbf{x} = x \mid Y = k)$$

Baye's theorem states that:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{13}$$

Bayes' theorem for classification

Getting an estimation of the *prior* π_k is easy: we can simply compute the fraction of Y in the training sample that belong to the kth class.

To estimate $f_k(X)$ is harder, unless we assume simple forms for the densities.

Let $p_k(x) = \mathbb{P}(Y = k \mid X = x)$ be the **posterior probability** that an observation $X = x$ belongs to the kth class (given the predictor value for that observation).

5.1 Liminar discriminant analysis for p=1

Liminar discriminant analysis for p=1

Let us assume first that we only have only $p = 1$ predictor.

We want to obtain an estimate of $f_k(x)$ to be able to predict $\mathbb{P}(Y = k \mid \mathbf{x} = x)$ using Eq. 13.

Based on that prediction, we will classify an observation to the class for which $\mathbb{P}(Y = k \mid \mathbf{x} = x)$ is the highest.

Gaussian density

Let us assume that the density $f_k(x)$ is Gaussian:

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right), \tag{14}$$

where μ_k and σ_k^2 are the mean and variance parameters, respectively, for the kth class.

Let us assume that the variance is equal accross all K classes: $\sigma_1^2 = \ldots = \sigma_K^2$

Gaussian density

Under the Gaussian hypothesis regarding the density $f_k(x)$, Eq. 13 becomes:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_l)^2}{2\sigma^2}\right)}. \tag{15}$$

Using Bayes' classifier, the class assigned to the observation $X = x$ is the one for which $\mathbb{P}(Y = k \mid \mathbf{x} = x)$ is the highest.

Let us show that this is equivalent to assign the observation $X = x$ the class for which the following value is the largest:

$$\delta_k(x) = x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

Gaussian density

Indeed, as the denominator of Eq. 15 does not depend on k, we can write:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = C\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right),$$

where $C = \sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_l)^2}{2\sigma^2}\right)$

We can further write:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = C_2\pi_k \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right) \tag{16}$$

where $C_2 = C \cdot \frac{1}{\sqrt{2\pi\sigma^2}}$

Gaussian density

Taking the log of Eq. 16 leads to:

$$\log\left(\mathbb{P}(Y=k \mid \mathbf{x}=x)\right) = \log(C_2) + \underbrace{\log(\pi_k) - \frac{(x-\mu_k)^2}{2\sigma^2}}_{\text{depends on k}} \tag{17}$$

We therefore want to maximize over k the following expression:

$$\log(\pi_k) - \frac{(x-\mu_k)^2}{2\sigma^2} \tag{18}$$

Gaussian density

That is, we wan to maximize over k the following expression:

$$\begin{aligned}
&\log(\pi_k) - \frac{(x-\mu_k)^2}{2\sigma^2} \\
=&\log(\pi_k) - \frac{x^2 - 2\mu_k x + \mu_k^2}{2\sigma^2} \\
=&C_3 + \log(\pi_k) - + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}.
\end{aligned} \tag{19}$$

where $C_3 = \frac{x^2}{2\sigma^2}$. The objective is thus to find the maximum value of:

$$\delta_k(x) = x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

Hence, for a point $X = x$, the predicted class is the one with the highest $\delta_k(x)$.

Gaussian density

Let us take an example, with $K = 2$, and $\pi_1 = \pi_2$.

The Bayes' classifier assigns an observation:

- to class 1 if $\delta_1(x) > \delta_2(x)$, *i.e.*, if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$
- to class 2 otherwise.

The **Bayes decision boundary** corresponds to the point where $\delta_1(x) = \delta_2(x)$, *i.e.*, where:

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}. \tag{20}$$

Example

Let us take an example to illustrate these concepts.

Let us suppose that we randomly pick $n_1 = 25$ observations from a $\mathcal{N}(-1, 1)$ with distribution function $f_1(x)$: the drawn observations are assigned the class 1.

Let us randomly pick $n_2 = 25$ more observations from a $\mathcal{N}(1, 1)$ with distribution function $f_2(x)$: the drawn observations are assigned the class 2.

As the two distributions overlap, there is uncertainty regarding which class an observation belongs.

Let us assume that an observation is equally likely to come from either class (so $\pi_1 = \pi_2$):

- according to Eq. 20, the Bayes' classifier assigns x to class 1 if $x < 0$.

Example

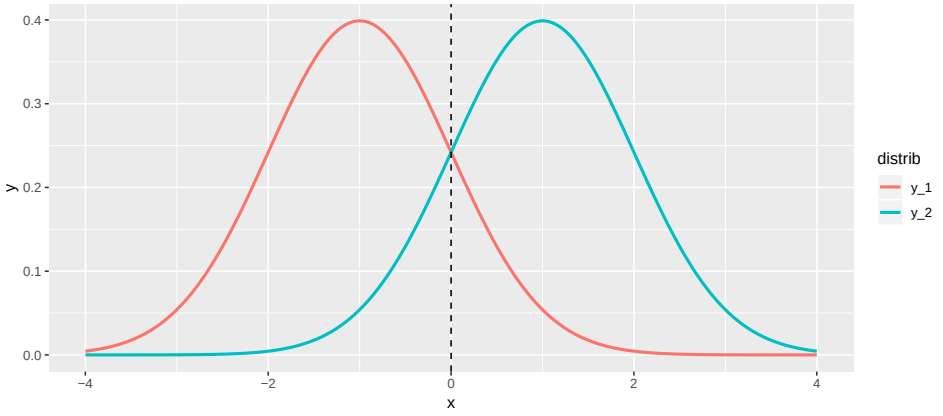


Figure 6: Example of LDA with 2 classes defined by the draw of 2 Normal distributions.

When the mean and variances are not known

In the previous example, not only we know that the classes are drawn from Normal distributions, but we also know the parameters of the distributions.

In such a situation, it is possible to compute the Bayes decision boundary.

In practice, **assuming that the classes are drawn from Normal distributions**, we still need to estimate the parameters $\mu_1, \ldots, \mu_k$ and $\sigma_1, \ldots, \sigma_k$.

The **linear discriminant analysis** (LDA) method approximates the Bayes classifier by estimating first these parameters.

When the mean and variances are not known

To estimate the **mean of the distribution** of class k, the sample mean is used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{21}$$

And to estimate the **variance**, a weighted average of sample variances is used:

$$\hat{\sigma}_k^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \tag{22}$$

When the probability π_k is unknown

We may not have information regarding π_k, the probability of class membership.

In that case, the LDA estimates it as the proportion of observations that belong the the kth class:

$$\hat{\pi}_k = \frac{n_k}{n}. \tag{23}$$

The LDA classifier

Using the estimate of μ_k given in Eq. 21, the estimate of σ^2 given in Eq. 22 and the estimate of π_k given in Eq. 23, the predicted class is the one with the highest estimate of $\delta_k(x)$, *i.e.*, the one with the highest:

$$\hat{\delta}_k(x) = x \times \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k). \tag{24}$$

Example

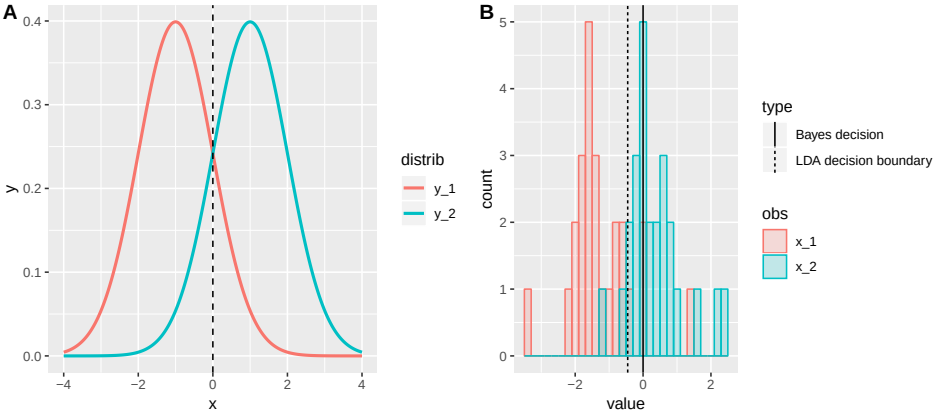


Figure 7: Example of LDA with 2 classes defined by the draw of 2 Normal distributions.

5.2 Linear discriminant analysis for p>1

Linear discriminant analysis for p>1

Now let us turn to the case in which $p > 1$, *i.e.*, when there are multiple predictors.

Let us assume that $X = (X_1, \ldots, X_p)$ is drawn from a **multivariate Normal** distribution, with a class-specific mean vector and a common covariance matrix:

$$X \sim \mathcal{N}(\mu, \mathbf{\Sigma}),$$

where μ is a vector with p comenents and $\mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X.

The probability density function of X writes:

$$f(x) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \mathbf{\Sigma}^{-1}(x-\mu)\right). \tag{25}$$

Linear discriminant analysis for p>1

Recall that:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

Pluging the density function for the kth class; *i.e.*, $f_k(X = x)$ into this equation writes:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \mathbf{\Sigma}^{-1}(x-\mu_k)\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_l)^\top \mathbf{\Sigma}^{-1}(x-\mu_l)\right)}. \quad (26)$$

Linear discriminant analysis for p>1

Using Bayes' classifier, the class assigned to the observation $X = x$ is the one for which $\mathbb{P}(Y = k \mid \mathbf{x} = x)$ is the highest.

Let us show that this is equivalent to assign the observation $X = x$ the class for which the following value is the largest:

$$\delta_k(x) = \delta_k(x) = -\frac{1}{2}\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k + x^\top \mathbf{\Sigma}^{-1}\mu_k + \log(\pi_k).$$

Linear discriminant analysis for p>1

As in the case where $p = 1$, the denominator of Eq. 26 does not depend on k. Eq. 26 writes:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = \frac{C\pi_k}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \mathbf{\Sigma}^{-1}(x - \mu_k)\right). \tag{27}$$

where $C = \frac{1}{\sum_{l=1}^{K} \pi_l \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_l)^\top \mathbf{\Sigma}^{-1}(x-\mu_l)\right)}$.

We can also write:

$$\mathbb{P}(Y = k \mid \mathbf{x} = x) = C_2\pi_k \exp\left(-\frac{1}{2}(x - \mu_k)^\top \mathbf{\Sigma}^{-1}(x - \mu_k)\right). \tag{28}$$

where $C_2 = \frac{C\pi_k}{(2\pi)^{p/2}|\Sigma|^{1/2}}$.

Linear discriminant analysis for p>1

Taking the logarithm of Eq. 28 results in:

$$\log \mathbb{P}(Y = k \mid \mathbf{x} = x) = \log(C_2) + \log(\pi_k) - \frac{1}{2}(x - \mu_k)^\top \mathbf{\Sigma}^{-1}(x - \mu_k). \tag{29}$$

We therefore aim at **maximizing over k** the following expression:

$$\log(\pi_k) - \frac{1}{2}(x - \mu_k)^\top \mathbf{\Sigma}^{-1}(x - \mu_k) \tag{30}$$

$$= \log(\pi_k) - \frac{1}{2}\left[x^\top \mathbf{\Sigma}^{-1} x - x^\top \mathbf{\Sigma}^{-1}\mu_k - \mu_k^\top \mathbf{\Sigma}^{-1} x + \mu_k^\top \mathbf{\Sigma}^{-1}\mu_k\right] \tag{31}$$

$$= \log(\pi_k) - \frac{1}{2}\left[x^\top \mathbf{\Sigma}^{-1} x + \mu_k^\top \mathbf{\Sigma}^{-1}\mu_k\right] + x^\top \mathbf{\Sigma}^{-1}\mu_k \tag{32}$$

$$= C_3 + \log(\pi_k) - \frac{1}{2}\left[\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k\right] + x^\top \mathbf{\Sigma}^{-1}\mu_k \tag{33}$$

where $C_3 = -\frac{1}{2}\left[x^\top \mathbf{\Sigma}^{-1} x\right]$

Linear discriminant analysis for p>1

The objectif is to find the maximum value of:

$$\delta_k(x) = -\frac{1}{2}\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k + x^\top \mathbf{\Sigma}^{-1}\mu_k + \log(\pi_k).$$

Hence, for a point $X = x$, the predicted class is the one with the highest $\delta_k(x)$.

Example

Let us take an example to illustrate this. Let us consider three equally-sized Gaussian classes drawn from normal bivariate distributions with identical covariance matrix but with class-specific mean vectors.

In that case, the Baye's decision boundaries are set according to:

$$-\frac{1}{2}\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k + x^\top \mathbf{\Sigma}^{-1}\mu_k + \log(\pi_k) = -\frac{1}{2}\mu_l^\top \mathbf{\Sigma}^{-1}\mu_l + x^\top \mathbf{\Sigma}^{-1}\mu_l + \log(\pi_l)$$
$$\Leftrightarrow \quad -\frac{1}{2}\mu_k^\top \mathbf{\Sigma}^{-1}\mu_k + x^\top \mathbf{\Sigma}^{-1}\mu_k = -\frac{1}{2}\mu_l^\top \mathbf{\Sigma}^{-1}\mu_l + x^\top \mathbf{\Sigma}^{-1}\mu_l$$

as $\log(\pi_k) = \log(\pi_l)$ in our example as the classes are equally-sized.

Example

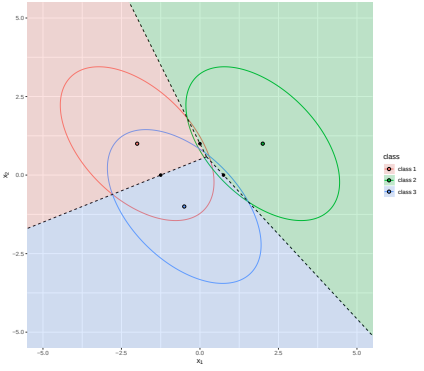


Figure 8: Classes drawn from three bivariate Normal distributions with identical covariance matrix but with class-specific mean vectors.

Example when the parameters are unknown

Now, let us consider that we only know that the classes are drawn from three bivariate Normal distributions, with the same covariance matrix but that we do not know:

- the mean vector for each Gaussian: μ_1, μ_2, μ_3
- the values of the covariance matrix: $\mathbf{\Sigma}$
- the probabilities that an observation belongs to a given class: π_1, π_2, π_3.

As in the univariate case, we can estimate these parameters from the observations.

Example when the parameters are unknown

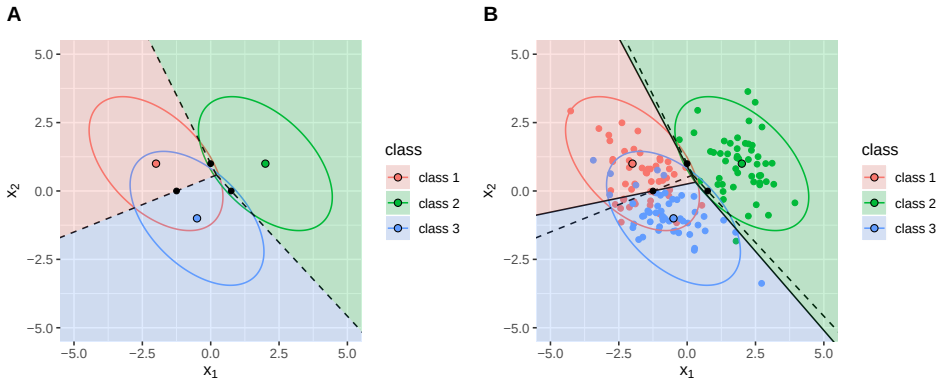


Figure 9: Classes drawn from three bivariate Normal distributions with identical covariance matrix but with class-specific mean vectors.

6. Assessing the quality of classification

Assessing the quality of classification

When fitting a classifier on sample data and testing it on a test sample, we can compare the predictions with the *observed values* and compute some metrics to assess goodness of fit.

6.1 Confusion tables

Confusion tables

It is possible to use **confusion tables** to compare the predictions of the fitted model to the actual classes. These tables cross-tabulate the **observed classes** against the classes that the **classifier assigns**.

Let us consider that we trained a model on a train dataset to be able to predict a categorical response variable with two classes: "yes" and "no".

		Predicted class		
		no predicted	yes predicted	Model error
True class	no	a	b	$b/(a+b)$
	yes	c	d	$c/(c+d)$
	Use error	$c/(a+c)$	$b/(b+d)$	Overall error $= \frac{(b+c)}{(a+b+c+d)}$

Table 2: A confusion table.

Confusion tables

We can create these tables for both predictions made on the training and the testing samples.

Four kinds of performance assessment can be made from confusion tables.

1. Looking at the overall proportion of cases **incorrectly classified**

		Predicted class		
		no predicted	yes predicted	Model error
True class	no	a	b	$b/(a+b)$
	yes	c	d	$c/(c+d)$
	Use error	$c/(a+c)$	$b/(b+d)$	Overall error $= \frac{(b+c)}{(a+b+c+d)}$

Confusion tables

2. Sometimes, we may want to be **more accurate for one class than for another** (*e.g.*, if we are trying to detect a cancer). Looking at **false positive** and **false negative**.

		Predicted class		
		no predicted	yes predicted	Model error
True class	no	a	b	$b/(a+b)$
	yes	c	d	$c/(c+d)$
	Use error	$c/(a+c)$	$b/(b+d)$	Overall error $= \frac{(b+c)}{(a+b+c+d)}$

Confusion tables

3. The **column proportions** help evaluate how useful the classifier results are likely to be if put to work: what happens when forecasting.

		Predicted class		
		no predicted	yes predicted	Model error
True class	no	a	b	$b/(a+b)$
	yes	c	d	$c/(c+d)$
	Use error	$c/(a+c)$	$b/(b+d)$	Overall error $= \frac{(b+c)}{(a+b+c+d)}$

Here the number of false positive is b.

Confusion tables

4. The **ratio of the number of false negative to the number of false positives** shows how the results are trading one kind of error for the other:
   - here c/b : if b is 5 times larger than c, there are 5 false positives for every false negative: the classifier produces results in which false negatives are five times more important than false positives.

Example

Let us go back to our logistic regression classification, on the data of client who subscribe or not to a term deposit. We model the probability of subscribing using the information on duration, education, campaign.

We obtain the following confusion matrix:

Table 3: Confusion matrix for the logistic regression.

y	No	Yes
No	3940	60
Yes	434	87

The **overall error** (we only consider a training sample here) is $\frac{434+60}{4,521} = 0.1092679$.

Example

Table 4: Confusion matrix for the logistic regression.

y	No	Yes
No	3940	60
Yes	434	87

- **false positive**: the classifier predicted that the client would subscribe but he actually did not
  - $\frac{60}{3,940+60} = 0.015$
- **false negative**: the classifier predicted that the client would not subscribe but he actually did
  - $\frac{434}{434+87} = 0.8330134$

We notice that the classifier is more accurate to classify the class "No" (when clients do not subscribe).

Example

Table 5: Confusion matrix for the logistic regression.

y	No	Yes
No	3940	60
Yes	434	87

Let us look at the ratio of the number of false negative to the number of false positives: $\frac{434}{60} = 7.233333$.

Hence, for every false positive (predicting that the client will subscribe), there are 7.23 false negative (predicting that the client will not subscribe when in fact he would).

6.2 The ROC curve

Costs

In the point mentioned in the previous slide, there is an underlying idea of the existence of **some costs**.

Recall that c is the number of false negative in our table, and b is the number of false positives.

The **off-diagonal cells** of the confusion table give us the number of **false positive** and **false negatives**.

The ratio of the cells c/b tells us that for every false positive, there are c/b false negatives:

- one false positive is "worth" c/b false negatives
- the cost is c/b times greater: **it is c/b times more costly to missclassify a false positive than to misclassify a false negative**.

Costs

So here, in the example of the clients who subscribe or not to a term deposit, for every false positive (predicting that the client will subscribe), there are 7.23 false negative (predicting that the client will not subscribe when in fact he would):

- this may not be acceptable for a business industry...
- we may want to be able to sacrifice some accuracy in predicting false positive in favor of better predict false negative.

Threshold of 0.5

So far, to assign a class to an observation x_0 in the case of a binary categorical response variable, we have used a threshold of 0.5 (either in the logistic regression, the Bayes classifier, or LDA):

- predicting default class if $\mathbb{P}(Y = 1 \mid X = x_0) \geq 0.5$ for a given observation $X = x_0$
- predicting alternative class otherwise

I.e., this corresponds to assign an observation x_0 the class k for which $\mathbb{P}(Y = k \mid X = x_0)$ is the highest.

Back to the example

We may want to change the value for the **threshold**.

Let us go back to the example of the two-class response variable y that takes two values: blue and orange, depending on the values of two predictors x_1 and x_2, such that:

$$y = \begin{cases} \text{blue} & \text{if } x_1^2 + x_2^2 > 60^2 \\ \text{orange} & \text{if } x_1^2 + x_2^2 \leq 60^2 \end{cases}$$

We randomly generate $n = 100$ observations for x_1 and x_2 from a $\mathcal{U}[0, 100]$, and assign the *true class* to y. From these observations, we fit a given classifier.

Then, we randomly generate $n = 100$ new observations for x_1 and x_2 from the same uniform distribution and use the classifier (k-nearest neihgbors with $K = 25$) to predict the probability of belonging to each of the two classes.

Changing the threshold value

What we previously did was to assign the class to each observed value according to:

$$\hat{y} = \begin{cases} \text{blue} & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) \geq 0.5 \\ \text{orange} & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) < 0.5 \end{cases}$$

Let us look at how we can change the **threshold** τ:

$$\hat{y} = \begin{cases} \text{blue} & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) \geq \tau \\ \text{orange} & \text{if } \hat{\mathbb{P}}(Y = 1 \mid X = x_0) < \tau \end{cases}$$

Changing the threshold value

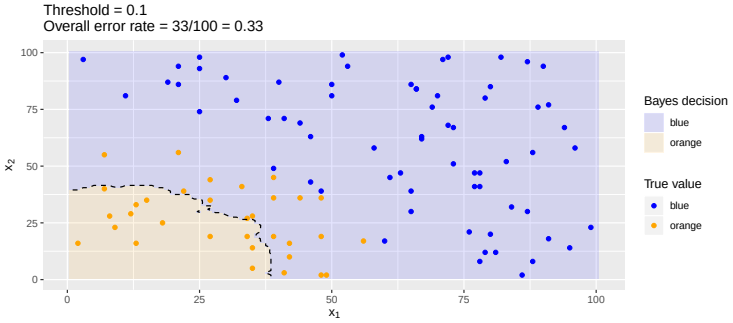


Figure 10: Decision boundary when varying the threshold.

Reminder

		Predicted class		
		$-$	$+$	Model error
True class	$-$	TN	FP	$FPR = FP/(TN+FP)$
	$+$	FN	TP	$FNR = FN/(FN+TP)$
	Use error	$FOR = FN/(TN+FN)$	$FDR = FP/(FP+TP)$	Overall error $ACC = \frac{(FP+FN)}{(TN+FP+FN+TP)}$

Table 6: A confusion table.

Changing the threshold value

Now, let us consider the bank data again and our logistic regression.

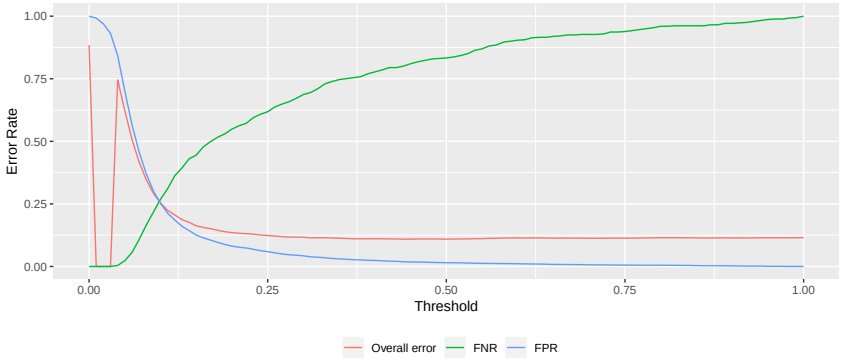


Figure 11: Varying the threshold on the logistic regression of the bank example. The red line represents the overall error. The blue line represents the fraction of misclassified clients among the non-subscribers. The green line represents the fraction of misclassified clients among the subscribers.

Changing the threshold value

In the previous graph, we can note that:

- a threshold of 0.5 minimizes the overall error rate
- reducing the value of the threshold:
  - diminishes the error rate among the individuals who subscribed ($FNR = FN/(FN + TP)$)
  - but increases the error rate among the individuals who did not ($FPR = FP/(FP + TN)$.

The ROC curve

For a binary categorical response variable, it is possible to draw a **ROC curve** which allows us to display the two types of error for different thresholds.

ROC means *receiver operating characteristics*.

The ROC curve is obtained by plotting the true positive rate agains false positive rate, at different values of the threshold.

The **area under the ROC curve (AUC)** allows us to assess the overall performance of a classifier:

- a value of 0.5 corresponds the the AUC for a random classifier
- the closer the AUC is to 1, the better

This is therefore a metric that can be used to select between different classifiers...

The ROC curve

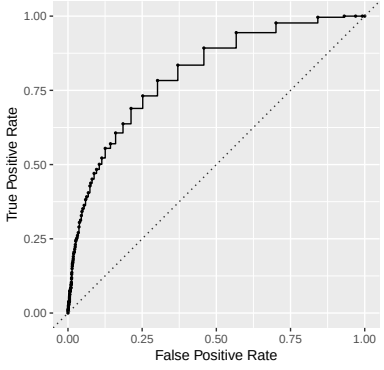


Figure 12: ROC curve for the logistic regression on the bank data.

In this example, the AUC is 0.8221.

Exercise

Lab exercise.

Berk, R. A. (2008). *Statistical learning from a regression perspective*, volume 14. Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, 7.