Chapter 8
Credit Scoring

Credit scoring provides a statistical assessment of a borrower's creditworthiness that helps financial institutions in making decisions on loan applications. In this chapter, we review the uses of discriminant analysis, and binomial logistic regression, with application to credit scoring. We also cover the properties of receiver operating characteristics curves.

8.1 Discriminant Analysis

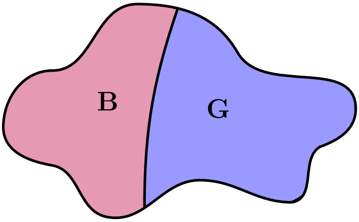
Consider a set Ω of credit applicants which is partitioned as

$$\Omega = B \cup G$$

into a subset G of "good" (or solvent) applicants, and a subset B of "bad" (or insolvent) applicants, with $B \cap G = \emptyset$. Applicants are selected at random within the set Ω according to a probability distribution $\mathbb{P}$, so that

$$\mathbb{P}(B) + \mathbb{P}(G) = 1.$$

This version: May 3, 2024
https://personal.ntu.edu.sg/nprivault/indext.html



Here, $\mathbb{P}(B)$ represents the probability that an applicant chosen at random may default, while $\mathbb{P}(G)$ represents the probability that a randomly selected applicant is solvent. In addition, each applicant $\omega \in \Omega$ is assigned a real-valued rating (or score) $X(\omega)$ via a *random variable*[*]

$$\begin{aligned} X : \Omega &\longrightarrow \mathbb{R} \\ \omega &\longmapsto X(\omega) \end{aligned}$$

with probability density function $f_X : \mathbb{R} \to \mathbb{R}_+$.

Definition 8.1. *1) The function*

$$\begin{aligned} \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto \mathbb{P}(B \mid X = x) \end{aligned}$$

is respectively called the probability default curve*.*

2) The function

$$\begin{aligned} \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto \mathbb{P}(G \mid X = x) \end{aligned}$$

is called the probability acceptance *curve.*

Denoting by $f_X(x \mid B)$, resp. $f_X(x \mid G)$, the probability density function of X given B, resp. X given G, we have

$$f_X(x) = f_X(x \mid G)\mathbb{P}(G) + f_X(x \mid B)\mathbb{P}(B), \quad x \in \mathbb{R},$$

and the Bayes formula yields

$$\mathbb{P}(B \mid X = x) = f_X(x \mid B)\frac{\mathbb{P}(B)}{f_X(x)}$$

[*] See (MOE and UCLES 2022, page 14 lines 4-5) and (MOE and UCLES 2020, page 19 lines 4-5).

$$= \frac{\mathbb{P}(B) f_X(x \mid B)}{\mathbb{P}(G) f_X(x \mid G) + \mathbb{P}(B) f_X(x \mid B)}, \tag{8.1}$$

and

$$\mathbb{P}(G \mid X = x) = \frac{\mathbb{P}(G) f_X(x \mid G)}{\mathbb{P}(G) f_X(x \mid G) + \mathbb{P}(B) f_X(x \mid B)}.$$

Definition 8.2. *1) The True Positive Rate (TPR) is given by the tail distribution function*

$$\overline{F}_G(x) := \mathbb{P}(X > x \mid G) = \int_x^\infty f_X(y \mid G) dy, \qquad x \in \mathbb{R}.$$

2) The False Positive Rate (FPR) is given by the tail distribution function

$$\overline{F}_B(x) := \mathbb{P}(X > x \mid B) = \int_x^\infty f_X(y \mid B) dy, \qquad x \in \mathbb{R}.$$

Example. In case X is Gaussian distributed given $\{G, B\}$ with the conditional densities

$$f_X(x \mid G) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)} \tag{8.2}$$

and

$$f_X(x \mid B) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)}, \tag{8.3}$$

with $\mu_B < \mu_G$, we have

$$f_X(x) = \frac{\mathbb{P}(G)}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)} + \frac{\mathbb{P}(B)}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)}, \quad x \in \mathbb{R},$$

and by (8.1) we find the logistic probability default curve

$$\begin{aligned} \mathbb{P}(B \mid X = x) &= \frac{\mathbb{P}(B) f_X(x \mid B)}{\mathbb{P}(G) f_X(x \mid G) + \mathbb{P}(B) f_X(x \mid B)} \\ &= \frac{\mathbb{P}(B) \mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)}}{\mathbb{P}(G) \mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)} + \mathbb{P}(B) \mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)}} \\ &= \frac{1}{1 + \mathrm{e}^{\alpha + \beta x}}, \qquad x \in \mathbb{R}, \end{aligned} \tag{8.4}$$

where we let

$$\beta := \frac{\mu_G - \mu_B}{\sigma^2} > 0$$

and

$$\alpha := -\beta \frac{\mu_G + \mu_B}{2} + \log \frac{\mathbb{P}(G)}{\mathbb{P}(B)}.$$

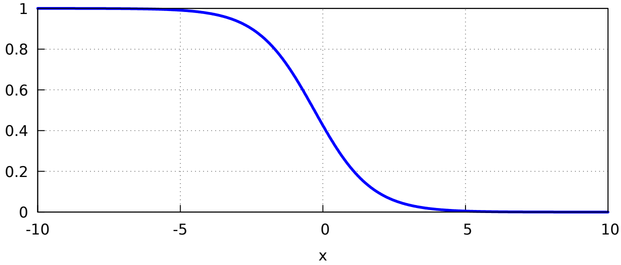


Fig. 8.1: Probability default curve $x \mapsto \mathbb{P}(B \mid X = x) = 1/(1 + \mathrm{e}^{a+x})$.

Similarly, the probability acceptance curve is given by

$$\begin{aligned}\mathbb{P}(G \mid X = x) &= \frac{\mathbb{P}(G) f_X(x \mid G)}{\mathbb{P}(G) f_X(x \mid G) + (1 - \mathbb{P}(G)) f_X(x \mid B)} \\ &= \frac{\mathbb{P}(G) \mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)}}{\mathbb{P}(G) \mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)} + \mathbb{P}(B) \mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)}} \\ &= \frac{1}{1 + \mathrm{e}^{-\alpha-\beta x}} \\ &= \frac{\mathrm{e}^{\alpha+\beta x}}{1 + \mathrm{e}^{\alpha+\beta x}}, \qquad x \in \mathbb{R}. \end{aligned} \tag{8.5}$$

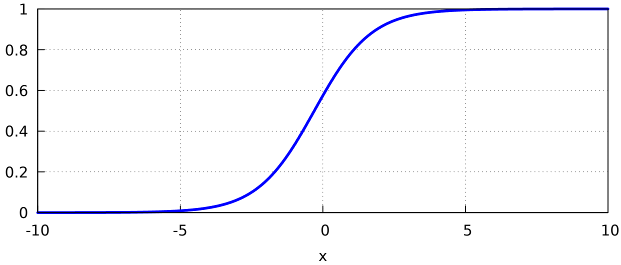


Fig. 8.2: Probability acceptance curve $x \mapsto \mathbb{P}(G \mid X = x) = \mathrm{e}^{a+x}/(1 + \mathrm{e}^{a+x})$.

8.2 Decision Rule

We decide to accept the applicants whose score $X(\omega)$ belongs to a decision (or acceptance) set $\mathscr{A} \subset \mathbb{R}$, and to reject those whose score $X(\omega)$ belongs to the rejection set $\mathscr{A}^c = \mathbb{R} \setminus \mathscr{A}$ which is the complement of $\mathscr{A}$ in $\mathbb{R}$.

Definition 8.3. *Let*

- $D(x)$ *represents the cost incurred by the default of an applicant with score* $x \in \mathscr{A}$, *and*
- $L(x)$ *represents the loss (or missed earnings) incurred by the rejection of an applicant with score* $x \in \mathscr{A}^c$.

The cost associated to this decision rule is defined as

$$\underbrace{D(X)\mathbb{1}_{\{X\in\mathscr{A}\}\cap B}}_{\text{Cost of accepting a "bad" applicant}} + \underbrace{L(X)\mathbb{1}_{\{X\in\mathscr{A}^c\}\cap G}}_{\text{Loss from rejecting a "good" applicant}}.$$

Theorem 8.4. *The optimal acceptance set* $\mathscr{A}^* \subset \mathbb{R}$ *that minimizes the expected cost*

$$\mathbb{E}\left[D(X)\mathbb{1}_{\{X\in\mathscr{A}\}\cap B} + L(X)\mathbb{1}_{\{X\in\mathscr{A}^c\}\cap G}\right]$$

is given by

$$\mathscr{A}^* = \left\{x \in \mathbb{R} \ : \ \lambda(x) \geqslant \frac{D(x)}{L(x)}\frac{\mathbb{P}(B)}{\mathbb{P}(G)}\right\},$$

where $\lambda(x)$ *is the* likelihood ratio

$$\lambda(x) := \frac{f_X(x \mid G)}{f_X(x \mid B)} = \frac{\mathbb{P}(G \mid X = x)}{\mathbb{P}(B \mid X = x)}\frac{\mathbb{P}(B)}{\mathbb{P}(G)}, \qquad x \in \mathbb{R}.$$

Proof. The expected cost corresponding to an acceptance set $\mathscr{A} \subset \mathbb{R}$ can be written as

$$\begin{aligned}
&\mathbb{E}\left[D(X)\mathbb{1}_{\{X\in\mathscr{A}\}\cap B} + L(X)\mathbb{1}_{\{X\in\mathscr{A}^c\}\cap G}\right] \\
&= \int_{\mathscr{A}} D(x)\mathbb{P}(B \cap \{X \in dx\}) + \int_{\mathscr{A}^c} L(x)\mathbb{P}(G \cap \{X \in dx\}) \\
&= \int_{\mathscr{A}} D(x)\mathbb{P}(B \mid X = x)f_X(x)dx + \int_{\mathscr{A}^c} L(x)\mathbb{P}(G \mid X = x)f_X(x)dx \\
&= \int_{-\infty}^{\infty} \left(\mathbb{1}_{\mathscr{A}}(x)D(x)\mathbb{P}(B \mid X = x) + \mathbb{1}_{\mathscr{A}^c}(x)L(x)\mathbb{P}(G \mid X = x)\right)f_X(x)dx.
\end{aligned}$$

The expected cost can be minimized pointwise by finding the set $\mathscr{A}$ that minimizes the conditional expected cost

$$x \longmapsto \mathbb{E}\left[D(X)\mathbb{1}_{\{X\in\mathscr{A}\}\cap B} + L(X)\mathbb{1}_{\{X\in\mathscr{A}^c\}\cap G} \,\middle|\, X = x\right]$$

$$= \mathbb{1}_{\mathscr{A}}(x)D(x)\mathbb{P}(B \mid X = x) + \mathbb{1}_{\mathscr{A}^c}(x)L(x)\mathbb{P}(G \mid X = x)$$

given that $X = x$. For this, we note that

$$\begin{aligned}
&\mathbb{1}_{\mathscr{A}}(x)D(x)\mathbb{P}(B \mid X = x) + \mathbb{1}_{\mathscr{A}^c}(x)L(x)\mathbb{P}(G \mid X = x) \\
&\quad \geqslant \min(D(x)\mathbb{P}(B \mid X = x), L(x)\mathbb{P}(G \mid X = x)) \\
&\quad = D(x)\mathbb{P}(B \mid X = x)\mathbb{1}_{\{D(x)\mathbb{P}(B|X=x)\leqslant L(x)\mathbb{P}(G|X=x)\}} \\
&\qquad + L(x)\mathbb{P}(G \mid X = x)\mathbb{1}_{\{L(x)\mathbb{P}(G|X=x)<D(x)\mathbb{P}(B|X=x)\}} \\
&\quad = D(x)\mathbb{P}(B \mid X = x)\mathbb{1}_{\mathscr{A}^*}(x) + L(x)\mathbb{P}(G \mid X = x)\mathbb{1}_{(\mathscr{A}^*)^c}(x),
\end{aligned}$$

where the set $\mathscr{A}^*$ which achieves equality in the above inequality is given by

$$\begin{aligned}
\mathscr{A}^* &:= \{D(x)\mathbb{P}(B \mid X = x) \leqslant L(x)\mathbb{P}(G \mid X = x)\} \\
&= \left\{ x \in \mathbb{R} \ : \ \frac{D(x)}{L(x)} \leqslant \frac{\mathbb{P}(G \mid X = x)}{\mathbb{P}(B \mid X = x)} \right\}.
\end{aligned}$$

In addition, the optimal acceptance set $\mathscr{A}^*$ can be rewritten in terms of the *likelihood ratio* function

$$\lambda(x) := \frac{f_X(x \mid G)}{f_X(x \mid B)} = \frac{\mathbb{P}(G \mid X = x)}{\mathbb{P}(B \mid X = x)} \frac{\mathbb{P}(B)}{\mathbb{P}(G)}, \qquad x \in \mathbb{R},$$

as

$$\mathscr{A}^* = \left\{ x \in \mathbb{R} \ : \ \lambda(x) \geqslant \frac{D(x)}{L(x)} \frac{\mathbb{P}(B)}{\mathbb{P}(G)} \right\}.$$

□

For simplicity, in the sequel we assume that $D = D(x)$ and $L = L(x)$ are constant in $x \in \mathbb{R}$, in which case we have

$$\mathscr{A}^* = \left\{ x \in \mathbb{R} \ : \ \lambda(x) \geqslant \frac{D}{L} \frac{\mathbb{P}(B)}{\mathbb{P}(G)} \right\}.$$

Proposition 8.5. *In the Gaussian example* (8.2)-(8.3) *with* $\mu_B < \mu_G$, *the optimal acceptance set* $\mathscr{A}^*$ *is given by*

$$\mathscr{A}^* = \left[\frac{\mu_G + \mu_B}{2} + \frac{1}{\beta} \log\left(\frac{D}{L} \frac{\mathbb{P}(B)}{\mathbb{P}(G)} \right), \infty \right), \tag{8.6}$$

where

$$\beta := \frac{\mu_G - \mu_B}{\sigma^2} > 0,$$

under the condition

$$\mathbb{E}[X \mid B] = \mu_B < \mu_G = \mathbb{E}[X \mid G].$$

Proof. In the Gaussian example (8.2)-(8.3), the likelihood ratio is given by

$$\begin{aligned}\lambda(x) &= \frac{f_X(x \mid G)}{f_X(x \mid B)} \\ &= \mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)+(x-\mu_B)^2/(2\sigma^2)} \\ &= \mathrm{e}^{-(\mu_G^2-\mu_B^2-2x(\mu_G-\mu_B))/(2\sigma^2)} \\ &= \mathrm{e}^{\beta x-(\mu_G^2-\mu_B^2)/(2\sigma^2)}, \qquad x \in \mathbb{R},\end{aligned}$$

hence the condition

$$\lambda(x) = \mathrm{e}^{\beta x-(\mu_G^2-\mu_B^2)/(2\sigma^2)} \geqslant \frac{D}{L}\frac{\mathbb{P}(B)}{\mathbb{P}(G)}$$

is equivalent to

$$\begin{aligned}x &\geqslant \frac{\mu_G^2-\mu_B^2}{2\beta\sigma^2} + \frac{1}{\beta}\log\left(\frac{D}{L}\frac{\mathbb{P}(B)}{\mathbb{P}(G)}\right) \\ &\geqslant \frac{\mu_G^2-\mu_B^2}{2(\mu_G-\mu_B)} + \frac{1}{\beta}\log\left(\frac{D}{L}\frac{\mathbb{P}(B)}{\mathbb{P}(G)}\right) \\ &\geqslant \frac{\mu_G+\mu_B}{2} + \frac{1}{\beta}\log\left(\frac{D}{L}\frac{\mathbb{P}(B)}{\mathbb{P}(G)}\right) \\ &=: x^*,\end{aligned}$$

hence $\mathscr{A}^* = [x^*, \infty)$, provided that

$$\beta := \frac{\mu_G - \mu_B}{\sigma^2} > 0,$$

which yields (8.6). □

We note that the optimal boundary point x^* satisfies the relation

$$\lambda(x^*) = \frac{\mathbb{P}(X = x^* \mid G)}{\mathbb{P}(X = x^* \mid B)} = \frac{\mathbb{P}(G \mid X = x^*)}{\mathbb{P}(B \mid X = x^*)}\frac{\mathbb{P}(B)}{\mathbb{P}(G)} = \frac{D}{L}\frac{\mathbb{P}(B)}{\mathbb{P}(G)},$$

i.e.

$$\frac{\mathbb{P}(G \mid X = x^*)}{\mathbb{P}(B \mid X = x^*)} = \frac{D}{L}. \tag{8.7}$$

Figure 8.3 illustrates the optimal decision rule by taking $L = D = \$1$ and using the default and acceptance curves (8.4)-(8.5).

This version: May 3, 2024
https://personal.ntu.edu.sg/nprivault/indext.html

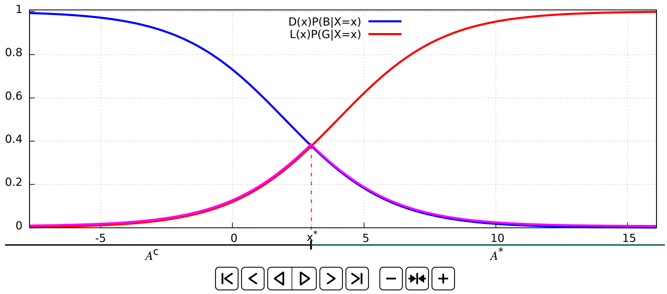


Fig. 8.3: Animated graph of optimal decision rule.*

In Figure 8.3, the conditional expected cost function

$$x \longmapsto \mathbb{1}_{\mathscr{A}}(x)D(x)\mathbb{P}(B \mid X = x) + \mathbb{1}_{\mathscr{A}^c}(x)L(x)\mathbb{P}(G \mid X = x)$$

is represented by the purple curve for a given set $\mathscr{A}$. Its uniform minimum over different threshold values is obtained for $\mathscr{A}^*$ of the form $\mathscr{A}^* = [x^*, \infty)$ where x^* lies at the intersection of the curves $x \mapsto D(x)\mathbb{P}(B \mid X = x)$ and $x \mapsto L(x)\mathbb{P}(G \mid X = x)$ as in (8.7).

The acceptance rate, or probability that an applicant is accepted according to the rule $\mathscr{A}$, is given by

$$\begin{aligned}\mathbb{P}(X \in \mathscr{A}) &= \mathbb{P}(\{X \in \mathscr{A}\} \cap G) + \mathbb{P}(\{X \in \mathscr{A}\} \cap B) \\ &= \mathbb{P}(X \in \mathscr{A} \mid G)\mathbb{P}(G) + \mathbb{P}(X \in \mathscr{A} \mid B)\mathbb{P}(B),\end{aligned}$$

where

$$\mathbb{P}(\{X \in \mathscr{A}\} \cap B) = \mathbb{P}(X \in \mathscr{A} \mid B)\mathbb{P}(B)$$

is the default rate, or probability that an applicant accepted according to the rule $\mathscr{A}$ will default.

We can also minimize the default rate $\mathbb{P}(\{X \in \mathscr{A}\} \cap B)$ subject to a given acceptance rate $\mathbb{P}(X \in \mathscr{A}) = a$.

8.3 Logistic Regression

In this section, we address the problem of constructing the random score variable $X : \Omega \to \mathbb{R}$ in a concrete setting. For this, consider a set of m

* The animation works in Acrobat Reader on the entire pdf file.

This version: May 3, 2024
https://personal.ntu.edu.sg/nprivault/indext.html

financial criteria or indicators $(x_{i,j})_{j=1,2,\ldots,m}$ applying to each of n credit applicants $i = 1, 2, \ldots, n$, with $\mathbb{P}(G) = 0.7$ and $\mathbb{P}(B) = 0.3$ in this data set.

```
install.packages("caret")
library(caret); data(GermanCredit); head(GermanCredit)
ggplot(GermanCredit, aes(x = Class)) + geom_bar(aes(y = (..count..)/sum(..count..)),fill=c(
      "red","darkgreen")) + labs(y = "prob.") + theme_bw()
```

j	Age	ForeignWorker	Property.RealEstate	Housing.Own	CreditHistory	Class
$x_{i,j}$	$x_{i,1} = 67$	$x_{i,2} = 1$	$x_{i,3} = 0$	$x_{i,4} = 1$	$x_{i,5} = 0$	$c_i = 1$

The credit scoring class ("good" or "bad") of applicant n°i is denoted by $c_i \in \{0, 1\}$ depending on his status, *i.e.* $c_i = 1$ for "good" applicants and $c_i = 0$ for "bad" applicants.

Linear regression

The score z_i of a given credit applicant in row n^o i is modeled as

$$z_i = F\left(\gamma_0 + \sum_{j=1}^{m} \gamma_j x_{i,j}\right), \qquad i = 1, 2, \ldots, n.$$

where $(\gamma_j)_{j=0,1,\ldots,m}$ is a family of linear coefficients, where $p_i := 1 - z_i$ represents the probability that applicant n^o i may default. In a linear regression model we would take $F(z) := z$, hence the system of equations

$$c_i = \gamma_0 + \sum_{j=1}^{m} \gamma_j x_{i,j}, \qquad i = 1, 2, \ldots, n,$$

would be used to estimate the coefficients $(\gamma_j)_{j=0,1,\ldots,m}$.

Binomial logistic regression

The shortcoming of linear models is that $F(z_i)$, which is assumed to represent a probability value, may exit the interval $[0, 1]$. Logistic regression models address this issue by replacing $F(x) = x$ with the logistic CDF F_L defined as

$$F_L(x) := \frac{\mathrm{e}^x}{1 + \mathrm{e}^x}, \qquad x \in \mathbb{R},$$

see also the Gaussian probability acceptance curve (8.5).

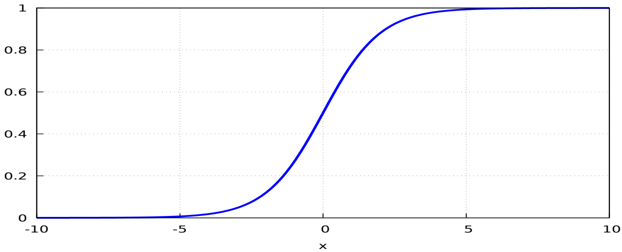


Fig. 8.4: Logistic CDF $x \mapsto F_L(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$.

We model the probability $p_i = \mathbb{P}(i \in G)$ that applicant n^o i is rated "good" as

$$p_i = \mathbb{P}(i \in G) = \mathbb{P}\left(G \,\middle|\, X = \gamma_0 + \sum_{j=1}^m \gamma_j x_{i,j}\right) = F_L\left(\gamma_0 + \sum_{j=1}^m \gamma_j x_{i,j}\right)$$

as in (8.5), $1 \leqslant i \leqslant n$. The probability of sampling a given $\{0,1\}$-valued sequence $(c_i)_{1\leqslant i\leqslant n}$ of applicant classifications is

$$\prod_{\substack{1\leqslant i\leqslant n\\ c_i=0}} \mathbb{P}(i \in B) \prod_{\substack{1\leqslant i\leqslant n\\ c_i=1}} \mathbb{P}(i \in G) = \prod_{i=1}^n \left((\mathbb{P}(i \in B))^{1-c_i}(\mathbb{P}(i \in G))^{c_i}\right).$$

In order to estimate the sequence of coefficients $(\gamma_j)_{j=0,1,\ldots,m}$, we aim at maximizing the log-likelihood ratio

$$\begin{aligned}
\log L(\beta|x) &:= \log \prod_{i=1}^n \left((\mathbb{P}(i \in B))^{1-c_i}(\mathbb{P}(i \in G))^{c_i}\right) \\
&= \log \prod_{i=1}^n \left(\left(\overline{F}_L\left(\gamma_0 + \sum_{j=1}^m \gamma_j x_{i,j}\right)\right)^{c_i}\left(F_L\left(\gamma_0 + \sum_{j=1}^m \gamma_j x_{i,j}\right)\right)^{1-c_i}\right) \\
&= \log \prod_{i=1}^n \left(\left(\frac{\mathrm{e}^{\gamma_0+\sum_{j=1}^m \gamma_j x_{i,j}}}{1+\mathrm{e}^{\gamma_0+\sum_{j=1}^m \gamma_j x_{i,j}}}\right)^{c_i}\left(\frac{1}{1+\mathrm{e}^{\gamma_0+\sum_{j=1}^m \gamma_j x_{i,j}}}\right)^{1-c_i}\right) \\
&= \sum_{i=1}^n c_i \log \frac{\mathrm{e}^{\gamma_0+\sum_{j=1}^m \gamma_j x_{i,j}}}{1+\mathrm{e}^{\gamma_0+\sum_{j=1}^m \gamma_j x_{i,j}}} + \sum_{i=1}^n (1-c_i) \log \frac{1}{1+\mathrm{e}^{\gamma_0+\sum_{j=1}^m \gamma_j x_{i,j}}}
\end{aligned}$$

This version: May 3, 2024
https://personal.ntu.edu.sg/nprivault/indext.html

over $\beta = (\gamma_j)_{j=0,1,\ldots,m}$ on the training set. The default probabilities are then estimated on the testing set from

$$1 - p_i = \overline{F}_L \left(\gamma_0 + \sum_{j=1}^{m} \gamma_j x_{i,j} \right), \qquad i = 1, \ldots, n,$$

and the *logit*

$$\overline{F}_L^{-1}(p_i) = -\log \frac{p_i}{1 - p_i} = \gamma_0 + \sum_{j=1}^{m} \gamma_j x_{i,j}, \qquad i = 1, \ldots, n,$$

or log-odds, represents probabilities on a logit scale. An implementation example of logistic regression is presented below using synthetic data.

```
y <- rnorm(15,1)
y<-(y-min(y))/(max(y)-min(y))
x <- rbinom(15,1,prob=y*y)
y<-data.frame(y)
glmreg<-glm(x ~ y,data=y,family="binomial")
xx<-seq(0,1,0.01)
xx<-data.frame(xx)
colnames(xx) <- c('y')
pred<-predict(glmreg,newdata=xx,type="response")
plot(y$y,x,col="blue")
lines(xx$y,pred,col="purple",lw=3)
points(y$y,((y$y-min(y$y))/(max(y$y)-min(y$y)))^2,col="red",lw=2)
```

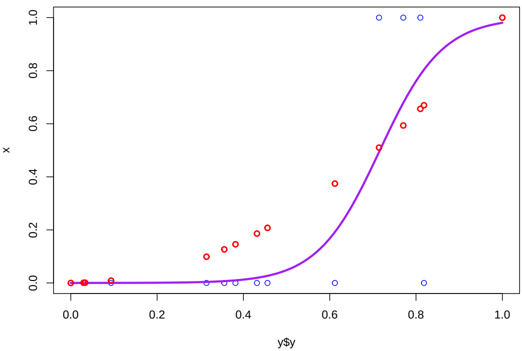


Fig. 8.5: GLM Regression.

```
Train <- createDataPartition(GermanCredit$Class, p=0.6, list=FALSE)
training <- GermanCredit[ Train, ];testing <- GermanCredit[ -Train, ]
```

The data is randomly split into a training set and a testing set using the createDataPartition command in R. The training set is used to fit the data in a generalized linear model using the glm() command. The testing set is then used to estimate the corresponding default probabilities.

```
mod.glm<-glm(Class ~ Age + ForeignWorker + Property.RealEstate + Housing.Own +
      CreditHistory.Critical, data=training, family="binomial")
head(testing$Class); head(predict(mod.glm, newdata=testing, type="response"))
testing <- cbind(rownames(testing), testing); colnames(testing)[1] <- "ID"
testing$Score5<-predict(mod.glm, newdata=testing, type="response")
testsamp <- head(testing,100);colnames(testsamp) <- make.unique(names(testsamp))
ggplot(testsamp, aes(x=ID, y=Score5, fill=Class)) + geom_bar(stat="identity") +
      scale_fill_manual(values=c( "red","darkgreen")) +
      scale_y_continuous(limits=c(0,1),expand = c(0,0)) + theme_bw(base_size = 18) +
      xlab(NULL) + theme(axis.text.x = element_blank(),aspect.ratio=0.5) +
      geom_hline(yintercept = 0.75,lwd=1.6)
```

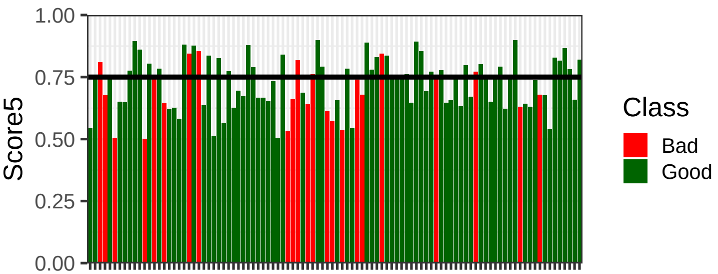


Fig. 8.6: Logistic regression output on 5 criteria.

Based on the above 100 samples, the next R code identifies the True Positive Rate (TPR) = 34/77 = 44.16%, and the False Positive Rate (FPR) = 8/23 = 34.78%, at the threshold $p = 0.75$.

```
cat('TPR5=',length(testsamp$Score5[testsamp$Score5>0.75 & testsamp$Class=="Good"])
      /length(testsamp$Score5[testsamp$Class=="Good"]),'\n')
cat('FPR5=',length(testsamp$Score5[testsamp$Score5>0.75 & testsamp$Class=="Bad"])
      /length(testsamp$Score5[testsamp$Class=="Bad"]),'\n')
pred <- ifelse(testsamp$Score5 > 0.75, "Good", "Bad")
confusionMatrix(factor(noquote(pred)),factor(testsamp$Class))
```

```
mod.glm<-glm(Class ~ Duration + Amount + InstallmentRatePercentage + ResidenceDuration
      + Age + NumberExistingCredits + NumberPeopleMaintenance + Telephone +
      ForeignWorker + CheckingAccountStatus.lt.0 + CheckingAccountStatus.0.to.200 +
      CheckingAccountStatus.gt.200 + CheckingAccountStatus.none +
      CreditHistory.NoCredit.AllPaid + CreditHistory.ThisBank.AllPaid +
      CreditHistory.PaidDuly + CreditHistory.Delay + CreditHistory.Critical +
      Purpose.NewCar + Purpose.UsedCar + Purpose.Furniture.Equipment +
      Purpose.Radio.Television + Purpose.DomesticAppliance + Purpose.Repairs +
      Purpose.Education + Purpose.Vacation + Purpose.Retraining + Purpose.Business +
      Purpose.Other + SavingsAccountBonds.lt.100 + SavingsAccountBonds.100.to.500 +
      SavingsAccountBonds.500.to.1000 + SavingsAccountBonds.gt.1000 +
      SavingsAccountBonds.Unknown + EmploymentDuration.lt.1 +
      EmploymentDuration.1.to.4 + EmploymentDuration.4.to.7 + EmploymentDuration.gt.7
      + EmploymentDuration.Unemployed + Personal.Male.Divorced.Seperated +
      Personal.Female.NotSingle + Personal.Male.Single + Personal.Male.Married.Widowed +
      Personal.Female.Single + OtherDebtorsGuarantors.None +
      OtherDebtorsGuarantors.CoApplicant + OtherDebtorsGuarantors.Guarantor +
      Property.RealEstate + Property.Insurance + Property.CarOther + Property.Unknown +
      OtherInstallmentPlans.Bank + OtherInstallmentPlans.Stores +
      OtherInstallmentPlans.None + Housing.Rent + Housing.Own + Housing.ForFree +
      Job.UnemployedUnskilled + Job.UnskilledResident + Job.SkilledEmployee +
      Job.Management.SelfEmp.HighlyQualified, data=training, family="binomial")
testing$Score61<-predict(mod.glm, newdata=testing, type="response")
testsamp <- head(testing,100);colnames(testsamp) <- make.unique(names(testsamp))
ggplot(testsamp, aes(x=ID, y=Score61, fill=Class)) + geom_bar(stat="identity") +
      scale_fill_manual(values=c( "red","darkgreen")) +
      scale_y_continuous(limits=c(0,1),expand = c(0,0)) + theme_bw(base_size = 18) +
      xlab(NULL) + theme(axis.text.x = element_blank(),aspect.ratio=0.5) +
      geom_hline(yintercept = 0.75,lwd=1.6)
```

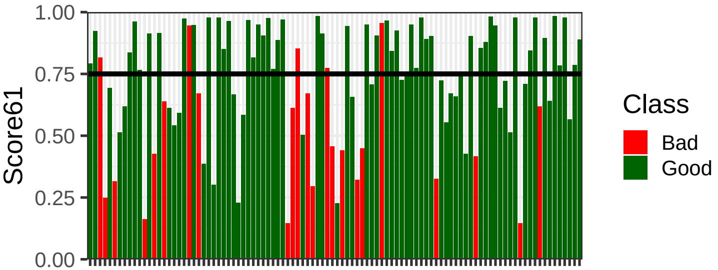


Fig. 8.7: Logistic regression output on 61 criteria.

In comparison with Figure 8.6, the 100 samples in Figure 8.7 above yield a higher True Positive Rate (TPR) $= 48/77 = 62.34\%$ and a lower False Positive Rate (FPR) $= 5/23 = 21.74\%$ at the level $p = 0.75$. In other words, the count of true positive samples has increased from 34 to 48, and the count of false positive samples has decreased from 6 to 5.

```
cat('TPR61=',length(testsamp$Score61[testsamp$Score61>0.75 &
     testsamp$Class=="Good"])/length(testsamp$Score61[testsamp$Class=="Good"]),'\n')
cat('FPR61=',length(testsamp$Score61[testsamp$Score61>0.75 &
     testsamp$Class=="Bad"])/length(testsamp$Score61[testsamp$Class=="Bad"]),'\n')
pred <- ifelse(testsamp$Score61 > 0.75, "Good", "Bad")
confusionMatrix(factor(noquote(pred)),factor(testsamp$Class))
```

8.4 ROC Curve

The ROC curve is a plot of the True Positive Rate values

$$x \longmapsto \overline{F}_G(x)$$

against the False Positive Rate function

$$x \longmapsto \overline{F}_B(x).$$

Definition 8.6. *The Receiver Operating Characteristic (ROC) curve is the function of the threshold* $p \in [0,1]$ *defined as*

$$\begin{array}{rcl} [0,1] & \longrightarrow & [0,1] \\ p & \longmapsto & \mathrm{ROC}(p) := \overline{F}_G\big(\overline{F}_B^{-1}(p)\big), \end{array}$$

where $\overline{F}_B^{-1}$ *denotes the inverse of the tail distribution function* $\overline{F}_B$*.*

The construction of Definition 8.6 proceeds in two steps.

a) Starting from a given FPR level p, compute the associated threshold x^* as $x^* := \overline{F}_B^{-1}(p)$, *i.e.*

$$p = \overline{F}_B(x^*) = \mathbb{P}(X > x^* \mid B).$$

b) From the threshold level, x^* estimate the corresponding TPR as

$$\mathbb{P}(X > x^* \mid G) = \overline{F}_G(x^*) = \overline{F}_G\big(\overline{F}_B^{-1}(p)\big) = \mathrm{ROC}(p).$$

Proposition 8.7. *The ROC function can be rewritten as the integral*

$$\mathrm{ROC}(p) = \int_0^p \lambda\big(\overline{F}_B^{-1}(q)\big)dq, \qquad 0 \leqslant p \leqslant 1,$$

where the likelihood ratio $\lambda(x)$ *is given by*

$$\lambda(x) = \frac{f_X(x \mid G)}{f_X(x \mid B)}, \qquad x \in \mathbb{R}.$$

Proof. The slope of the ROC curve at the point $p \in [0,1]$ is given by

$$\begin{aligned}\frac{d}{dp}\overline{F}_G\big(\overline{F}_B^{-1}(p)\big) &= \overline{F}'_G\big(\overline{F}_B^{-1}(p)\big)\frac{d}{dp}\overline{F}_B^{-1}(p)\\ &= \frac{\overline{F}'_G\big(\overline{F}_B^{-1}(p)\big)}{\overline{F}'_B\big(\overline{F}_B^{-1}(p)\big)}\\ &= \lambda\big(\overline{F}_B^{-1}(p)\big),\end{aligned}$$

hence we have

$$\overline{F}_G\big(\overline{F}_B^{-1}(p)\big) = \int_0^p \lambda\big(\overline{F}_B^{-1}(q)\big)dq, \qquad 0 \leqslant p \leqslant 1.$$

□

When X is Gaussian distributed given $\{G, B\}$ with the conditional densities

$$f_X(x \mid G) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)}$$

and

$$f_X(x \mid B) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)},$$

the likelihood ration is given by

$$\lambda(x) = \frac{f_X(x \mid G)}{f_X(x \mid B)} = \mathrm{e}^{\beta x-(\mu_G^2-\mu_B^2)/(2\sigma^2)}, \qquad x \in \mathbb{R},$$

with $\beta := (\mu_G - \mu_B)/\sigma^2$.

The ROC curve in this Gaussian setting can be computed using the following R code.

```
FG <- function(x, mug, sigma) {1 - pnorm((x - mug) / sigma)}
FBINV <- function(x, mub, sigma) {mub + sigma * qnorm(1 - x)}
ROC <- function(x, mub, mug, sigma) {
  sapply(x, function(x_val) ifelse(x_val == 0, 0, FG(FBINV(x_val, mub, sigma), mug, sigma)))}
```

Figure 8.8 presents three samples of ROC curves in the Gaussian example with successively $(\mu_B, \mu_G) = (1, 4)$, $(\mu_B, \mu_G) = (1, 2)$, and $(\mu_B, \mu_G) = (1, 1)$.

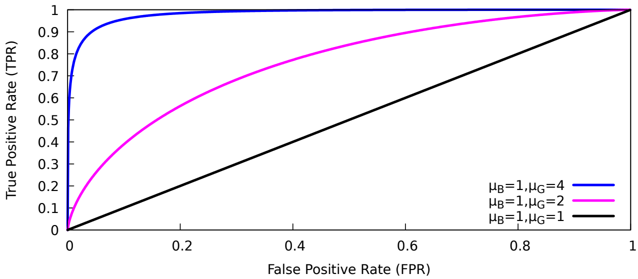


Fig. 8.8: Gaussian ROC curves.

We check that the classification is better when $\mu_B << \mu_G$. The ROC Curve shows the performance of the classification procedure, which is quantified by the Area Under The Curve (AUC). A perfect classification would correspond to a single point with coordinates $(0, 1)$ and AUC equal to 1, which corresponds to FPR=0% false negatives and TPR=100% true positives. The closer the AUC is to 1, the better the classification is performing.

On the other hand, a completely random guess would correspond to a point on the diagonal line. Points above the diagonal represent good classification results (better than random), while points below the line represent poor results (worse than random) (Wikipedia).

```
TPR5<-length(testing$Score5[testing$Score5>0.75 &
      testing$Class=="Good"])/length(testing$Score5[testing$Class=="Good"])
FPR5<-length(testing$Score5[testing$Score5>0.75 &
      testing$Class=="Bad"])/length(testing$Score5[testing$Class=="Bad"])
cat('TPR5=',TPR5,'\n'); cat('FPR5=',FPR5,'\n')
TPR61<-length(testing$Score61[testing$Score61>0.75 &
      testing$Class=="Good"])/length(testing$Score61[testing$Class=="Good"])
FPR61<-length(testing$Score61[testing$Score61>0.75 &
      testing$Class=="Bad"])/length(testing$Score61[testing$Class=="Bad"])
cat('TPR61=',TPR61,'\n'); cat('FPR61=',FPR61,'\n')
```

The above True Positive Rates (TPR) and False Positive Rates (FPR) based on 5 and 61 criteria at the level $p = 0.75$ are recomputed in the above R code on the whole 400 samples, and plotted on the next ROC graphs of Figure 8.9.

```
install.packages("ROCR"); library(ROCR)
pred5<-prediction(as.numeric(testing$Score5),as.numeric(testing$Class))
perf5 <- performance(pred5,"tpr","fpr")
dev.new(width=16,height=7); par(mar = c(4.5,4.5,2,2))
plot(perf5,col="purple",lwd=3, xaxs = "i", yaxs = "i",cex.lab=2,las=1)
segments(FPR5,0,FPR5,TPR5, col="purple", lwd =2)
segments(0,TPR5,FPR5,TPR5, col="purple", lwd =2)
pred61<-prediction(as.numeric(testing$Score61),as.numeric(testing$Class))
perf61 <- performance(pred61,"tpr","fpr"); par(new=TRUE)
plot(perf61,col="blue",lwd=3,main="", ann=FALSE, xaxs="i", yaxs="i")
legend("bottomright", legend=c("61 criteria","5 criteria"),col=c("blue","purple"), lwd=3, cex=3)
segments(0,0,1,1, col="black", lwd =3)
segments(FPR61,0,FPR61,TPR61, col="blue", lwd =2)
segments(0,TPR61,FPR61,TPR61, col="blue", lwd =2)
```

The ROC graphs in the next Figure 8.9 confirm the improvement in classification reached when switching from 5 to 61 criteria.

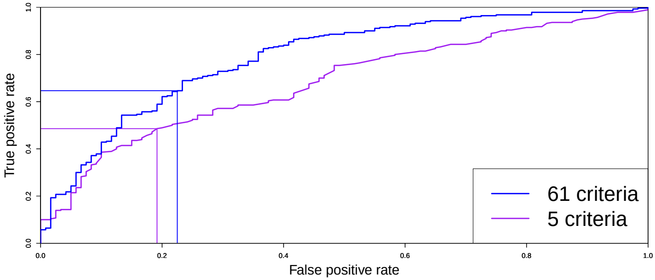


Fig. 8.9: ROC curves based on 5 criteria and 61 criteria.

Using a neural network in R

Note: Restarting R is *needed* before running the following code.

```r
install.packages("neuralnet");
library(neuralnet);library(caret);data(GermanCredit);library(ROCR);
Train <- createDataPartition(GermanCredit$Class, p=0.6, list=FALSE)
training <- GermanCredit[ Train, ];testing <- GermanCredit[ -Train, ]
nn <- neuralnet(Class ~ Age + ForeignWorker + Property.RealEstate + Housing.Own +
       CreditHistory.Critical, data=training, hidden=c(3,1), linear.output=FALSE,
       threshold=0.05)
nn$result.matrix; plot(nn,col.intercept = "blue")
temp_test <- subset(testing, select = c("Age", "ForeignWorker", "Property.RealEstate",
       "Housing.Own", "CreditHistory.Critical"))
head(temp_test); nn.results <- compute(nn, temp_test)
results <- data.frame(actual = testing$Class, prediction = nn.results$net.result)
head(results); pred <- ifelse(results[,3] > 0.75, "Good", "Bad")
confusionMatrix(factor(results$actual),factor(noquote(pred)))
pred2<-prediction(as.numeric(results[,3]),as.numeric(results$actual))
perf <- performance(pred2,"tpr","fpr"); plot(perf,col="purple",lwd=3, xaxs = "i", yaxs = "i")
```

Improved performance may be achieved by rescaling the binary variables to $\{-1, 1\}$.

```r
training$ForeignWorker <- ifelse(training$ForeignWorker == 0, -1, 1)
training$Property.RealEstate <- ifelse(training$Property.RealEstate  == 0, -1, 1)
training$Housing.Own <- ifelse(training$Housing.Own == 0, -1, 1)
training$CreditHistory.Critical <- ifelse(training$CreditHistory.Critical  == 0, -1, 1)
nn2 <- neuralnet(Class ~ Age + ForeignWorker + Property.RealEstate + Housing.Own +
       CreditHistory.Critical, data=training, hidden=c(3,1), linear.output=FALSE,
       threshold=0.05)
nn2.results <- compute(nn2, temp_test)
results2 <- data.frame(actual = testing$Class, prediction = nn2.results$net.result)
pred3 <- ifelse(results2[,3] > 0.75, "Good", "Bad")
confusionMatrix(factor(results2$actual),factor(noquote(pred3)))
pred4<-prediction(as.numeric(results2[,3]),as.numeric(results$actual))
dev.new(width=16,height=7); par(mar = c(4.5,4.5,2,2))
plot(perf,col="purple",lwd=3, xaxs = "i", yaxs = "i",cex.lab=2,las=1)
perf3 <- performance(pred4,"tpr","fpr"); par(new=TRUE)
plot(perf3,col="blue",lwd=3,main="", ann=FALSE, xaxs="i", yaxs="i")
legend("bottomright", legend=c("Rescaled","Non rescaled"),col=c("blue","purple"), lwd=3,
       cex=3)
```

Using a neural network in Python

Download the corresponding **IPython notebook*** that can be run **here** using this **data file**.

Using random forests in Python

Download the corresponding **IPython notebook** that can be run **here** using this **data file**.

* Right-click to save as attachment (may not work on 🍎).

Using XGBoost in R

Download the corresponding R **code**.

Exercises

Exercise 8.1 Consider a set Ω of applicants decomposed into the partition $\Omega = G \cup B$, where each applicant ω is assigned a score $X(\omega)$ which is exponentially distributed given $\{G, B\}$ with the conditional densities

$$f_X(x \mid G) = \lambda_G \mathrm{e}^{-\lambda_G x} \mathbb{1}_{[0,\infty)}(x)$$

and

$$f_X(x \mid B) = \lambda_B \mathrm{e}^{-\lambda_B x} \mathbb{1}_{[0,\infty)}(x),$$

where $\lambda_B > \lambda_G > 0$.

a) Compute the conditional expected values $\mathbb{E}[X \mid G]$ and $\mathbb{E}[X \mid B]$.
b) Compute the *probability default curve*

$$x \longmapsto \mathbb{P}(B \mid X = x) = \frac{\mathbb{P}(B) f_X(x \mid B)}{\mathbb{P}(G) f_X(x \mid G) + \mathbb{P}(B) f_X(x \mid B)}$$

in terms of the likelihood ratio $\lambda(x) := f_G(x)/f_B(x)$, $x \in \mathbb{R}$.
c) Determine the acceptance set

$$\mathscr{A} := \big\{ x \in \mathbb{R} \ : \ D\mathbb{P}(B \mid X = x) \leqslant L\mathbb{P}(G \mid X = x) \big\},$$

where

- L represents the loss incurred by the rejection of an applicant, and
- D represents the loss incurred by the default of an applicant.

Exercise 8.2 Consider a set Ω of applicants decomposed into the partition $\Omega = G \cup B$, where each applicant ω is assigned a uniformly distributed score $X(\omega)$ given $\{G, B\}$, with the conditional densities

$$f_X(x \mid G) = \frac{1}{\lambda_G} \mathbb{1}_{[0,\lambda_G]}(x) dx$$

and

$$f_X(x \mid B) = \frac{1}{\lambda_B} \mathbb{1}_{[0,\lambda_B]}(x) dx,$$

where $0 < \lambda_B < \lambda_G$.

a) Compute the *probability default curve*

$$x \longmapsto \mathbb{P}(B \mid X = x) = \frac{\mathbb{P}(B)f_X(x \mid B)}{\mathbb{P}(G)f_X(x \mid G) + \mathbb{P}(B)f_X(x \mid B)}.$$

b) Compute the conditional expected values $\mathbb{E}[X \mid G]$ and $\mathbb{E}[X \mid B]$.
c) Determine the acceptance set

$$A := \big\{x \in \mathbb{R}_+ \ : \ D\mathbb{P}(B \mid X = x) \leqslant L\mathbb{P}(G \mid X = x)\big\},$$

where

- L represents the missed earnings incurred by the rejection of applicant,
- D represents the loss incurred by the default of an applicant.

Exercise 8.3

a) Compute the ROC function $x \mapsto \overline{F}_G\big(\overline{F}_B^{-1}(x)\big)$ of Exercise 8.1, $x \in [0,1]$, and draw its graph for sample values of λ_G, λ_B.
b) Compute the ROC function $x \mapsto \overline{F}_G\big(\overline{F}_B^{-1}(x)\big)$ of Exercise 8.2, $x \in [0,1]$, and draw its graph for sample values of λ_G, λ_B.

Exercise 8.4 Consider a set Ω of customers decomposed as the partition $\Omega = G \cup B$, where each customer ω is assigned a uniformly distributed score $X(\omega)$ given $\{G, B\}$, with the conditional densities

$$f_X(x \mid G) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu_G)^2/(2\sigma^2)}$$

and

$$f_X(x \mid B) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu_B)^2/(2\sigma^2)},$$

with $\mu_B < \mu_G$.

a) Compute the *probability default curve*

$$x \longmapsto \mathbb{P}(B \mid X = x) := \frac{\mathbb{P}(B)f_X(x \mid B)}{\mathbb{P}(G)f_X(x \mid G) + \mathbb{P}(B)f_X(x \mid B)}.$$

b) Compute the likelihood ratio $\lambda(x)$ defined as $\lambda(x) := f_X(x \mid G)/f_X(x \mid B)$, $x \in \mathbb{R}$.
c) Letting

- $L(x) := \mathrm{e}^{-ax}$ represent the missed earnings incurred by rejecting an applicant with score $x \in \mathbb{R}$, with $a > 0$,
- $D(x) := \mathrm{e}^{bx}$ represent the loss incurred by the default of an applicant with score $x \in \mathbb{R}$, with $b > 0$,

determine the acceptance set

$$\mathcal{A} := \big\{ x \in \mathbb{R}_+ \ : \ D(x)\mathbb{P}(B \mid X = x) \leqslant L(x)\mathbb{P}(G \mid X = x) \big\}.$$