Institut de Mathématiques de Toulouse, INSA Toulouse

Supervised Learning- Part II

Formation en machine Learning
EUR NanoX

Béatrice Laurent - Philippe Besse - Olivier Roustant

Methods studied in this course :

Part I

- Linear models for regression
- Linear models for classification

Part II

- Classification And Regression Trees, Bagging, Random Forests
- Neural networks, Introduction to deep learning

Outline

- Linear models for regression
- Linear models for classification

Part I-2 : Classification

- Theory : Optimal Bayes classifier
- Logistic Regression
  - Definitions
  - Estimation of the parameters
  - Application
  - Multiclass classification
- A word on linear discriminant analysis
- Two-class problems : beyond Bayes classifier
  - ROC curve

Part I-2 : Classification

- We now consider supervised classification problems. We have a training data set with n observation points (or objects) $\boldsymbol{X}_i$ and their class (or label) Y_i.
- Suppose that $\boldsymbol{d}^n$ corresponds to the observation of a n-sample $\boldsymbol{D}^n = \{(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)\}$ with joint unknown distribution P on $\mathcal{X} \times \mathcal{Y}$.
- A *classification rule* is a measurable function $f : \mathcal{X} \to \mathcal{Y}$ that associates the output $f(\boldsymbol{x})$ to the input $\boldsymbol{x} \in \mathcal{X}$.
- In order to quantify the quality of the prevision, we introduce a loss function.

Definition

A measurable function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a *loss function* if $\ell(y, y) = 0$ and $\ell(y, y') > 0$ for $y \neq y'$.

- **For classification** : $\mathcal{Y}$ is a finite set. We define $\ell(y, y') = \mathbb{1}_{y \neq y'}$.
- We consider the expectation of this loss, this leads to the definition of the *risk* :

Definition

Given a loss function ℓ, the *risk* - or *generalisation error* - of a prediction rule f is defined by

$$R_P(f) = \mathbb{E}_{(\boldsymbol{X}, Y) \sim P}[\ell(Y, f(\boldsymbol{X}))]$$

- It is important to note that, in the above definition, $(\boldsymbol{X}, Y)$ is independent of the training sample $\boldsymbol{D}^n$ that was used to build the prediction rule f.

- Let $\mathcal{F}$ denote the set of all possible prediction rules. We say that f^* is an optimal rule if $R_P(f^*) = \inf_{f \in \mathcal{F}} R_P(f)$.
- A natural question arises : is it possible to build optimal rules ?
- We define the Bayes rule, which is an optimal rule for classification.

Definition

We call *Bayes rule* any measurable function f^* in $\mathcal{F}$ such that for all $\boldsymbol{x} \in \mathcal{X}$, $\mathbb{P}(Y = f^*(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x}) = \max_{y \in \mathcal{Y}} \mathbb{P}(Y = y | \boldsymbol{X} = \boldsymbol{x})$.

Theorem

— *If f^* is a Bayes rule, then* $R_P(f^*) = \inf_{f \in \mathcal{F}} R_P(f)$.

- The definition of a Bayes rule depends on the knowledge of the distribution P of $(\boldsymbol{X}, Y)$.
- In practice, we have a training sample $\boldsymbol{D}^n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ with joint unknown distribution P, and we construct a classification rule.
- The aim is to find a "good" classification rule, in the sense that its risk is close to the optimal risk of a Bayes rule.

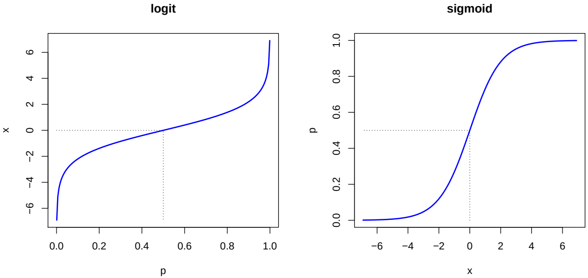
Part I-2

- Theory : Optimal Bayes classifier
- Logistic Regression
  - Definitions
  - Estimation of the parameters
  - Application
  - Multiclass classification
- Two-class problems : beyond Bayes classifier
  - ROC curve

Logistic regression model

The idea for logistic regression is to use a linear model for probabilities, thanks to a one-to-one mapping ("link" function) from $[0,1]$ to $\mathbb{R}$.
The most used is the logit function and its inverse, the sigmoid function :

$$
\begin{array}{lccccl}
 & [0,1] & & \mathbb{R} & & \\
\hline
\textbf{logit}: & \pi & \rightarrow & \ln\left(\frac{\pi}{1-\pi}\right) & & \\
 & \frac{\exp(x)}{1+\exp(x)} & \leftarrow & x & & : \textbf{sigmoid}
\end{array}
$$



Logistic regression model

- We assume that $\mathcal{X} = \mathbb{R}^p$.
- One of the most popular model for binary classification when $\mathcal{Y} = \{0, 1\}$ is the **logistic regression model**, for which it is assumed that for all $x \in \mathcal{X}$ and for some $\beta \in \mathbb{R}^p$,

$$\pi(\boldsymbol{x}) = \mathbb{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\langle \beta, \boldsymbol{x} \rangle)}{1 + \exp(\langle \beta, \boldsymbol{x} \rangle)}$$
$$1 - \pi(\boldsymbol{x}) = \mathbb{P}(Y = 0 | \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp(\langle \beta, \boldsymbol{x} \rangle)},$$

- The quantity $odds(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}$ is called the odds for $\boldsymbol{x}$. For example, if $\pi(\boldsymbol{x}) = 0.8$, then $odd(\boldsymbol{x}) = 4$ which means that the chance of success ($Y = 1$) when $\boldsymbol{X} = \boldsymbol{x}$ is 4 against 1.
- The odds ratio between $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ is $OR(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = odds(\boldsymbol{x})/odds(\tilde{\boldsymbol{x}})$.

Illustration in 1D

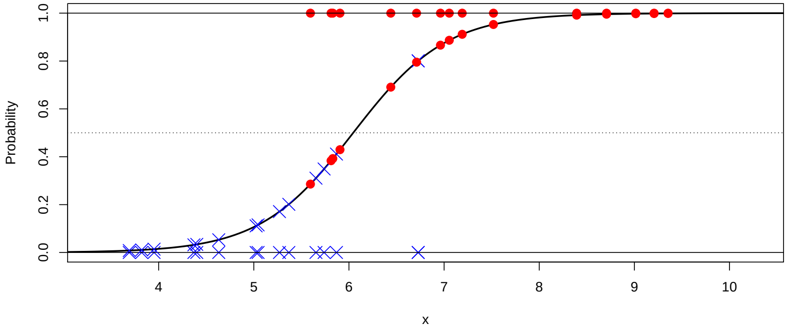


FIGURE – Logistic regression for a dataset composed of 2 groups of size 15, sampled from Normal distributions, centered at 5 and 7, with variance 1.

Parameters estimation

- Given a n-sample $\boldsymbol{D}^n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$, we can estimate the parameter β by maximizing the conditional likelihood of $\underline{Y} = (Y_1, \ldots, Y_n)$ given $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$.
- Since the distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ is a Bernoulli distribution with parameter $\pi_\beta(\boldsymbol{x})$, the conditional likelihood is

$$L(Y_1, \ldots, Y_n, \beta) = \prod_{i=1}^{n} \pi_\beta(\boldsymbol{X_i})^{Y_i}(1 - \pi_\beta(\boldsymbol{X_i}))^{1-Y_i}$$

$$L(\underline{Y}, \beta) = \prod_{i, Y_i=1} \frac{\exp(\langle \beta, \boldsymbol{X_i} \rangle)}{1 + \exp(\langle \beta, \boldsymbol{X_i} \rangle)} \prod_{i, Y_i=0} \frac{1}{1 + \exp(\langle \beta, \boldsymbol{X_i} \rangle)}.$$

Parameters estimation

- Unlike the linear model, there is no explicit expression for the maximum likelihood estimator $\hat{\beta}$.
- It can be shown that computing $\hat{\beta}$ is a convex optimization problem.
- We compute the gradient of the log-likelihood, also called **the score function** $S(\underline{Y}, \beta)$ and use a **Newton-Raphson algorithm** to approximate $\hat{\beta}$ satisfying $S(\underline{Y}, \hat{\beta}) = 0$.
- Variable selection is also possible by maximizing the penalized likelihood (AIC, BIC, LASSO ..).

- We can then predict the probabilities :

$$\hat{\mathbb{P}}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \pi_{\hat{\beta}}(\boldsymbol{x}) = \frac{\exp(\langle \hat{\beta}, \boldsymbol{x} \rangle)}{1 + \exp(\langle \hat{\beta}, \boldsymbol{x} \rangle)}$$

$$\hat{\mathbb{P}}(Y = 0|\boldsymbol{X} = \boldsymbol{x}) = 1 - \pi_{\hat{\beta}}(\boldsymbol{x}) = \frac{1}{1 + \exp(\langle \hat{\beta}, \boldsymbol{x} \rangle)}.$$

- We then compute the logistic regression classifier : we set $\hat{Y}(\boldsymbol{x}) = 1$ if $\hat{\mathbb{P}}(Y = 1|\boldsymbol{X} = \boldsymbol{x})) \geq \hat{\mathbb{P}}(Y = 0|\boldsymbol{X} = \boldsymbol{x})$ which is equivalent to $\langle \hat{\boldsymbol{\beta}}, \boldsymbol{x} \rangle \geq 0$. Hence,

$$\hat{Y}(\boldsymbol{x}) = \mathbb{1}_{\langle \hat{\boldsymbol{\beta}}, \boldsymbol{x} \rangle \geq 0}.$$

Application

- We use the logistic regression model to predict the exceedance of the threshold 150 for the variable O3obs.
- Only with the variable MOCAGE :

```
> logistic <- glm(depseuil ~ MOCAGE, data = ozone,
                  family = binomial(link = "logit"))
> summary(logistic)
```

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.596493	0.389841	-14.36	<2e-16 ***
MOCAGE	0.028659	0.002528	11.34	<2e-16 ***

Application

- We compute the predicted values :

```
> pihat <- logistic$fitted.values
> Yhat <- (pihat > 0.5)
> table(depseuil, Yhat)
```

$Y \setminus \hat{Y}$	0	1
0	830	33
1	152	26

- The misclassification error is 17.7%. There are many false negatives.
- The model tends to underestimate the threshold overflow : only 15% of the overflows have been predicted.
- We try to improve the model by considering more variables.

Application

- We consider the variables JOUR, MOCAGE,TEMPE, RMH2O, NO2, NO

```
> logistic2 <- glm(depseuil ~ MOCAGE + TEMPE + RMH2O + NO2 + NO + JOUR1,
                   data = ozone, family = binomial(link = "logit"))
> summary(logistic2)
```

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14.840457	1.116901	-13.287	< 2e-16 ***
MOCAGE	0.026924	0.004045	6.655	2.82e-11 ***
TEMPE	0.309566	0.029529	10.483	< 2e-16 ***
RMH2O	138.430723	28.548702	4.849	1.24e-06 ***
NO2	-0.210011	0.102607	-2.047	0.0407 *
NO	0.742302	0.552606	1.343	0.1792
JOUR1	0.159047	0.235654	0.675	0.4997

Application

- We compute the predicted values :

```
> pihat <- logistic2$fitted.values
> Yhat <- (pihat > 0.5)
> table(depseuil, Yhat)
```

$Y \setminus \hat{Y}$	0	1
0	829	34
1	88	90

- The misclassification error is 11.7%.
- We have improved the results, but there are still many false negative : only 50% of the overflows have been predicted.

Multinomial or polytomic regression

- Here the response variable Y has M levels $u_1, \ldots, u_M$.
- Define, for all m levels, $\pi_m(\boldsymbol{x}) = \mathbb{P}(Y = u_m | \boldsymbol{X} = \boldsymbol{x})$.
  Observe that :
$$\sum_{m=1}^{M} \pi_m(\boldsymbol{x}) = 1.$$
- We choose a reference for levels, say the first one u_1.
  The multinomial regression model is then defined by
$$\log\left(\frac{\pi_m(\boldsymbol{x})}{\pi_1(\boldsymbol{x})}\right) = \langle \beta^{(\boldsymbol{m})}, \boldsymbol{x} \rangle \qquad \forall m = 2, \ldots, M.$$

- This is equivalent to

$$\pi_m(\boldsymbol{x}) = \frac{\exp(\langle \beta^{(\boldsymbol{m})}, \boldsymbol{x} \rangle)}{1 + \sum_{m'=2}^{M} \exp(\langle \beta^{(\boldsymbol{m'})}, \boldsymbol{x} \rangle)}$$

which generalizes the logistic regression model $(u_1 = 0, u_2 = 1)$.

- The parameters $\beta^{(\boldsymbol{m})}$ are estimted by maximizing the likelihood :

$$L(\underline{Y}, \beta) = \prod_{i=1}^{n} \prod_{m=1}^{M} \pi_m(\boldsymbol{X_i})^{\mathbb{1}_{Y_i = u_m}}.$$

Part I-2 : Classification

- Theory : Optimal Bayes classifier
- Logistic Regression
  - Definitions
  - Estimation of the parameters
  - Application
  - Multiclass classification
- A word on linear discriminant analysis
- Two-class problems : beyond Bayes classifier
  - ROC curve

Linear discriminant analysis

Linear Discriminant Analysis (LDA), probabilistic approach

- Assume that each law $\mathbf{X}|Y = y$ is Normal, with the same variance.
- Applying the Bayes rules gives a form of... logistic regression !

Remarks

- LDA can be viewed as a particular case of logistic regression.
  But parameters are not estimated in the same way (slight differences).
- LDA is equivalent to the geometric method proposed by Fisher : Find a direction s.t. the projections on it maximize the ratio $\frac{\text{inter-class variance}}{\text{intra-class variance}}$.
- Relaxing the constant variance assumption gives a quadratic frontier
  $\rightarrow$ Quadratic Discriminant Analysis (QDA)

Part I-2 : Classification

- Theory : Optimal Bayes classifier

- Logistic Regression
  - Definitions
  - Estimation of the parameters
  - Application
  - Multiclass classification

- A word on linear discriminant analysis

- Two-class problems : beyond Bayes classifier
  - ROC curve

Two-classes problem : ROC curve

Motivation

For two classes $\mathcal{Y} = \{0, 1\}$, the optimal Bayes rule is :

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > \frac{1}{2} \quad \Leftrightarrow \quad \boldsymbol{x} \text{ belongs to class 1}$$

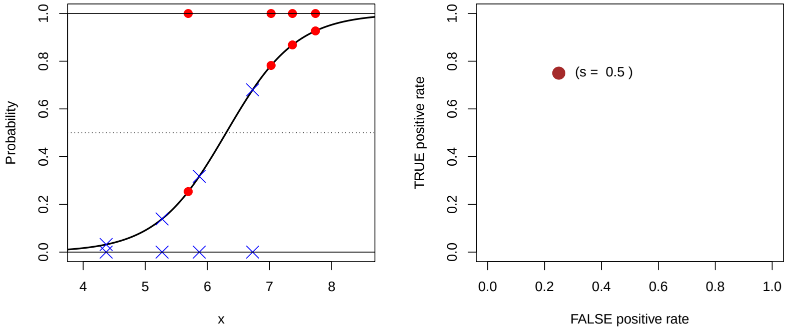
This gives a symmetric role to classes 0 and 1, which is often not desirable (health context, for instance)

The idea is to parameterize the decision by a new threshold parameter s :

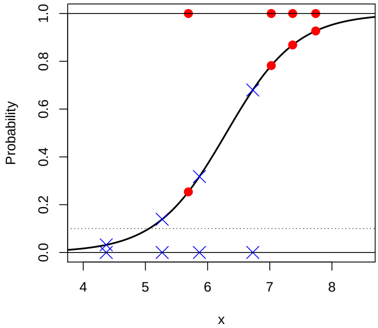
$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > s \quad \Leftrightarrow \quad \boldsymbol{x} \text{ belongs to class 1}$$

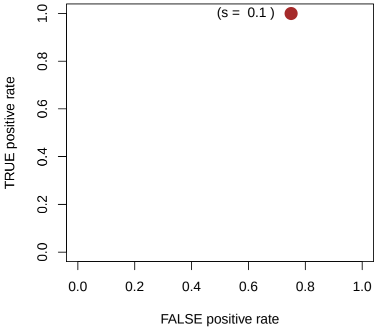
s should be chosen according to policy decision, typically a tradeoff between the rate of true positive and false positive.

ROC curve - Illustration in 1D

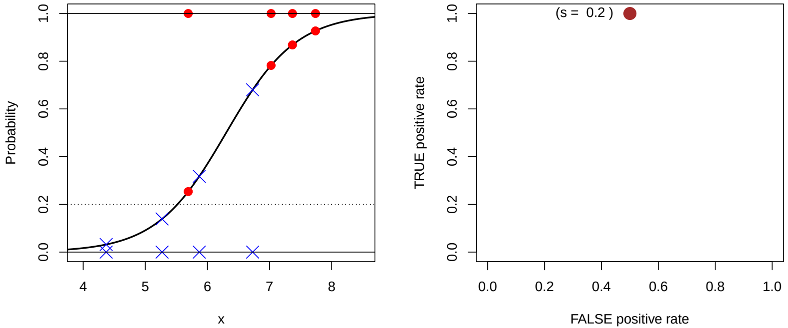


ROC curve - Illustration in 1D

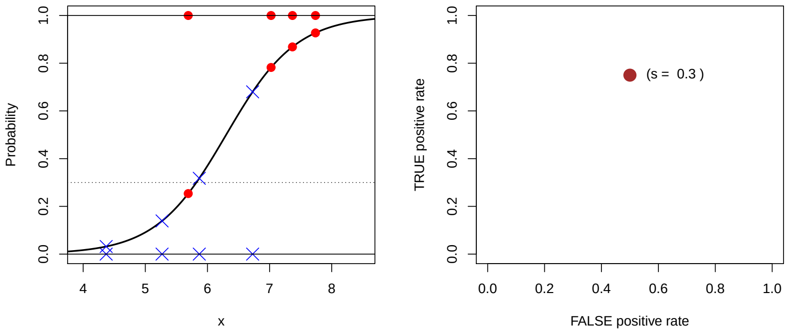




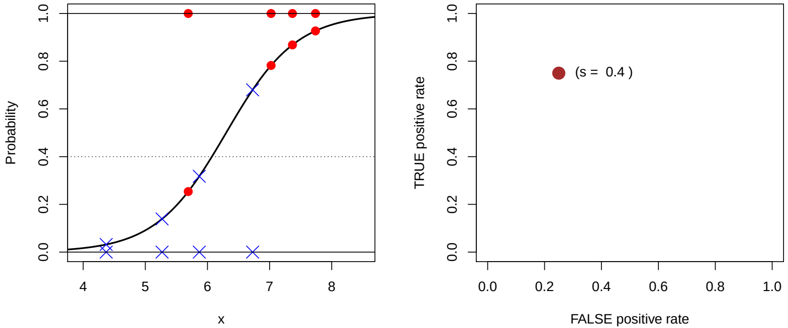
ROC curve - Illustration in 1D



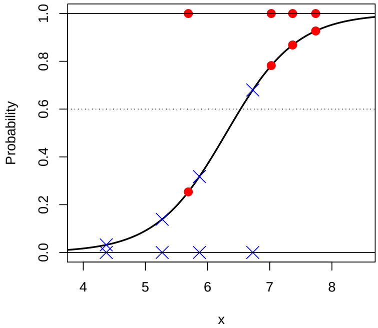
ROC curve - Illustration in 1D

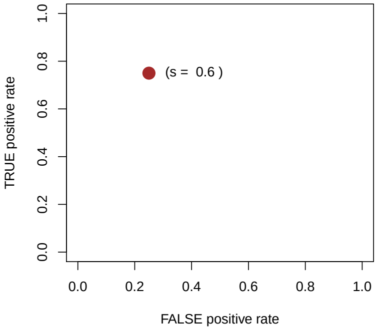


ROC curve - Illustration in 1D

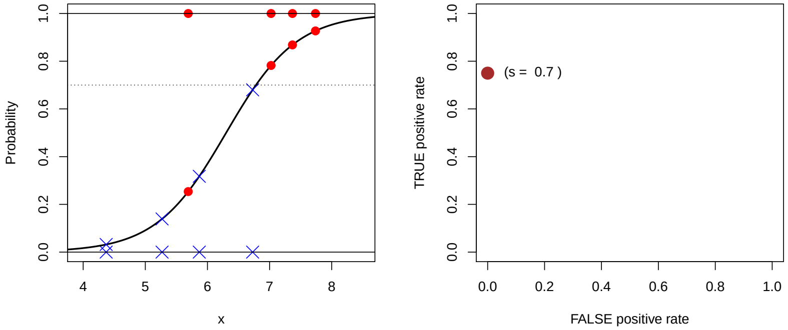


ROC curve - Illustration in 1D

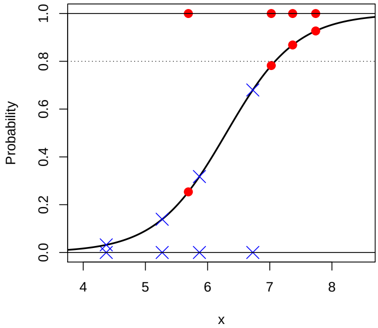


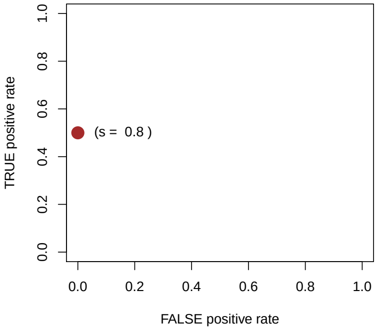


ROC curve - Illustration in 1D

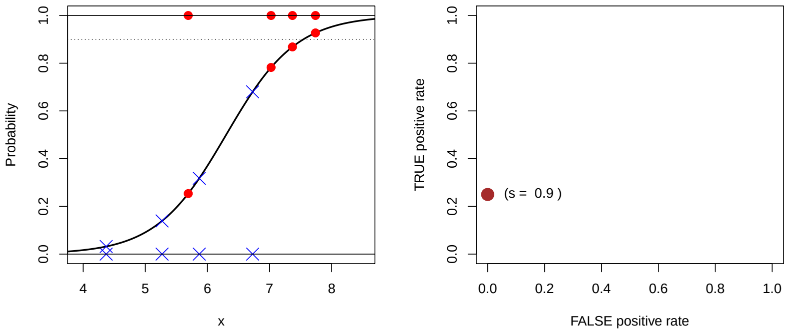


ROC curve - Illustration in 1D

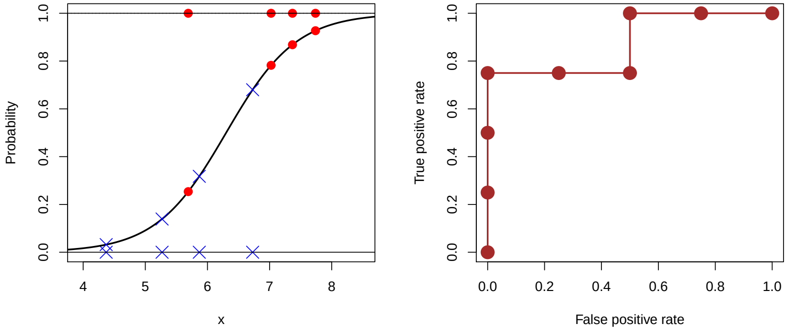




ROC curve - Illustration in 1D



ROC curve - Illustration in 1D



ROC curve - Definition

Definitions from the contingency table

Prediction : if $\widehat{\pi}_i > s,\ \widehat{y}_i = 1$ else $\widehat{y}_i = 0$

Prediction	Observation		Total
	$Y = 1$	$Y = 0$	
$\widehat{y}_i = 1$	$n_{11}(s)$	$n_{10}(s)$	$n_{1+}(s)$
$\widehat{y}_i = 0$	$n_{01}(s)$	$n_{00}(s)$	$n_{0+}(s)$
Total	n_{+1}	n_{+0}	n

- True positive rate : $TPR(s) = \frac{n_{11}(s)}{n_{+1}}$ *(sensitivity, recall)*
- False positive rate : $FPR(s) = \frac{n_{10}(s)}{n_{+0}}$

The ROC curve plots TPR(s) versus FPR(s) for all values of $s \in [0, 1]$.

Usage of ROC curve to select classifiers

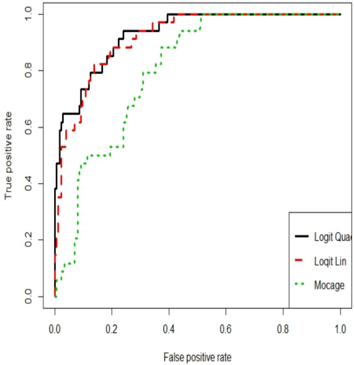


FIGURE – Ozone : ROC curve for three models. Here, logistic regression should be prefered to Mocage.