# Predictive models in Oncology

Luis Mariano Esteban & Gerardo Sanz

Universidad de Zaragoza (Spain)

CESS 2016

Budapest, October 20-21, 2016

# Problem

`To build and implement binary classification models in oncology with better predictive ability`

Work lines:

- Construction of linear models that maximize the area under ROC curve.
- Selection of thresholds (from density functions) in order to achieve practical rules for using the classifier.

# Binary classification problem

- $X_1, \ldots, X_n$ : predictive variables measured on a set of patients.
- $Y$ denotes the status of each patient.

To predict $Y$ (from $X's$)? (0, healthy, 1 diseased)

## Definition

*A classifier* $\mathbf{Y} := f(X_1, \ldots, X_n)$ *is a function of the predictive variables and* $c$ *is a threshold that separates patients (healthy - diseased).*

*An individual* $i$ *is classified as diseased if* $\mathbf{Y} := f(X_{1i}, \ldots, X_{ni}) > c$, *where* $X_{ki}$ *denotes the value of the* $k$*-th variable measured on the individual* $i$.

*Analogously, the individual is classified as healthy if* $\mathbf{Y} := f(X_{1i}, \ldots, X_{ni}) < c$

# Performance measures: sensibility and specificity

## Definition

*Let $\hat{Y}(\mathbf{X})$ be the status assigned by the classifier to an individual. Let $c$ the cutpoint that separates the two states.*

$$\begin{aligned} \text{Sensibility} &\equiv TPR(c) = P\{(\hat{Y}(\mathbf{X})) \geq c | Y = 1\} \\ \text{Specificity} &\equiv TNR(c) = P\{(\hat{Y}(\mathbf{X})) < c | Y = 0\} \end{aligned}$$

## Definition

$$ROC(c) = \{(1 - \text{Specificity}(c), \text{Sensibility}(c)), c \in R\}$$

If $F_0$ and $F_1$ denote the distribution functions of the populations, healthy (0) and diseased (1), obtained from the results of the classifier,

$$ROC = \{(1 - F_0(c), 1 - F_1(c)), c \in R\}$$

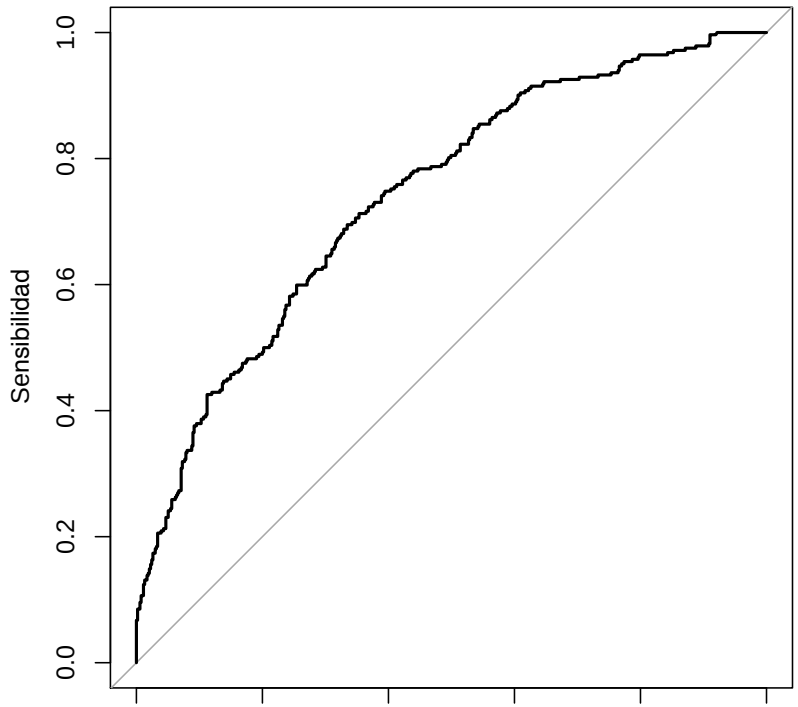
or equivalently: $ROC(t) = \{(t, 1 - F_1(F_0^{-1}(1 - t))), t \in [0, 1]\}$

# ROC curve: Non-parametric approach

$c \in (-\infty, +\infty)$, and $\{p_1, \ldots, p_{n_1}\}$, $\{\overline{p}_1, \ldots, \overline{p}_{n_0}\}$ numerical values provided by the classifier for the classes 1 and 0, respectively.

$$TPR(c) = \sum_{i=1}^{n_1} \frac{I(p_i > c)}{n_1} \quad , \quad FPR(c) = \sum_{i=1}^{n_0} \frac{I(\overline{p}_i > c)}{n_0}$$

$n_1$ and $n_0$ number of individuals of classes 1 and 0 respectively.



# Summary measure of ROC curve: area under the ROC curve (AUC)

## Definition

*Given $F_0$, $F_1$ the distribution functions provided by the classifier for the classes 0 and 1, the AUC is given by*

$$AUC = \int_0^1 (1 - F_1(F_0^{-1}(1-t)))\mathrm{d}t$$

*Under Normality* $Y^0 \sim N(\mu_0, \sigma_0)$, $Y^1 \sim N(\mu_1, \sigma_1)$:

$$\hat{AUC} = \Phi\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right) \quad , \quad \texttt{where} \quad \hat{a} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_1}, \quad \hat{b} = \frac{\hat{\sigma}_0}{\hat{\sigma}_1}$$

## AUC: Non-parametric estimation

Let $Y_0$, $Y_1$ be the values provided by the classifier for classes 0 and 1:

$$\hat{AUC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(Y_j^1 > Y_i^0) \ , \ I(z > x) = \begin{cases} 1 & \text{si } z > x \\ 1/2 & \text{si } z = x \\ 0 & \text{si } z < x \end{cases}$$

# Estimation of linear models under the criterion of maximizing the AUC

**Objective:** Given $(X_1, \ldots, X_n)$, estimate the parameters $(\beta_2, \ldots, \beta_n)$ such that the linear model

$$L(\mathbf{Y}) = Y_1 + \beta_2 \cdot Y_2 + \ldots + \beta_n \cdot Y_n$$

maximizes the AUC.

**Remark.-** The above model and the more complex model:

$$L_g(\mathbf{Y}) = g(\beta_0 + \beta_1 \cdot Y_1 + \ldots \beta_n \cdot Y_n),$$

with $g$ any increasing monotone function, have the same AUC.

(Under multivariate Normality: Linear discriminant function maximizes the AUC.)

$$\text{Normality?} \Longrightarrow \text{Non-parametric approach}$$

# Some proposal

- First of all, a set of possible values for each parameter $\beta_j$ is selected.
- The parameters $\beta_i$ are evaluated only inside the interval $[-1, 1]$, because the estimation of $X_i + \beta X_j$ with $\beta > 1$ or $\beta < -1$ is equivalent to estimate $\frac{1}{\beta} X_i + X_j$ with $\frac{1}{\beta} \in [-1, 1]$.
- If $k$ values are considered for each coefficient $\beta_i$ and the predictive variables are $n$, the number of AUC's to analize is of order $nk^{n-1}$ !!!!!

**It is not a viable procedure even if the number of variables is small.**

# Our proposal: Step by step algorithm

**Sketch of the algorithm**

- First of all, the best combination of two variables is selected (AUC max).
- Then, we estimate the coefficient $\beta_3$ such that $(X_i + \beta_2 X_j) + \beta_3 X_k$ con $k \neq i, j = 1, \ldots, n$ has maximum AUC.
- The process is repeated till all variables are included.
- This algorithm has a computational complexity obviously smaller: $k(n-1)(\frac{3}{2}n - 1)$.

# Step by step algorithm

Technical questions

- Normalizations, multiple optima
- Exclusion of models with poor predictive ability:

$$H_0 : AUC_{\text{Model n variables}} \leq AUC_{\text{Model n}-\text{k variables}}$$
$$H_1 : AUC_{\text{Model n variables}} > AUC_{\text{Model n}-\text{k variables}}$$

- Correlations between variables.
- Use of additive models. Number of terms in additive models that maximize the AUC.

# Discussion

## Discussion

- Under normality (independent or not) the algorithm converges to the maximum theoretical value.
- In absence of normality, the algorithm performs better that the logistic regression.
- For a real database of prostate cancer the algorithm and the logistic model shown a similar behaviour.
- The algorithm has been used successfully in prediction of breast cancer ( Nicolosi et al, JBSE 2013).

# Choosing a threshold point in binary classification problems

## Choosing a threshold

- The practical application of the classification models requires the selection of a threshold point $c$ that defines a satisfactory classification rule.
- To determine whether a threshold point provides a good classification, we can analyse the ROC curve.
- The best..... high sensibility and specificity.

# Classical criteria for choosing a threshold

## Definition

*Let $Y$ be a classifier: $Y = f(X_1, \ldots, X_n)$. The Youden index is the point $c \in R$ that maximizes*

$$TPR(c) - FPR(c) \qquad (Sensitivity + Specificity - 1)$$

## Definition

*Let $Y$ be a classifier: $Y = f(X_1, \ldots, X_n)$. The optimum point $c$ is the point that minimizes the quantity*

$$\sqrt{(1 - TPR(c))^2 + FPR(c)^2}$$

*( (TPR(c), FPR(c)) is the closest point to (0,1) )*

# Our proposal

- Estimation of density functions $f_0$ and $f_1$ (Smoothing through kernel functions).
- Selection of a threshold point to separate the two populations.

**Remark.-** Smoothing of a function $f$ from a data set $x_1, \ldots, x_n$:

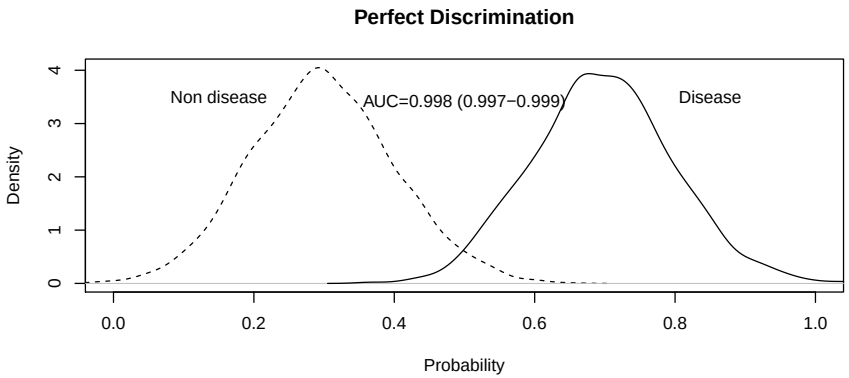
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

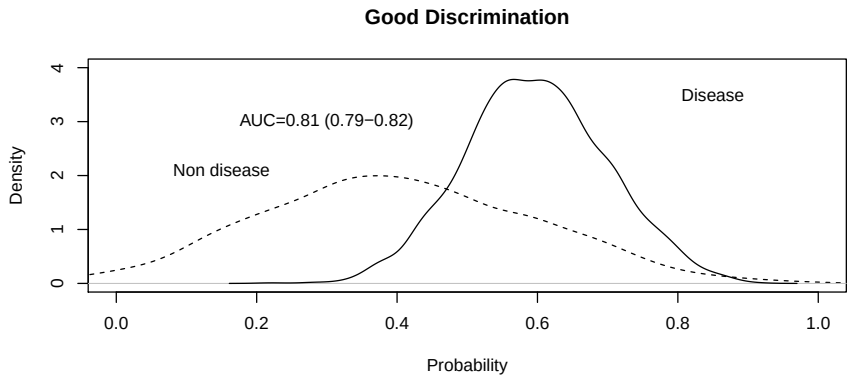
where $h$ is the bandwidth and $k$ the kernel function.
Suggestions:

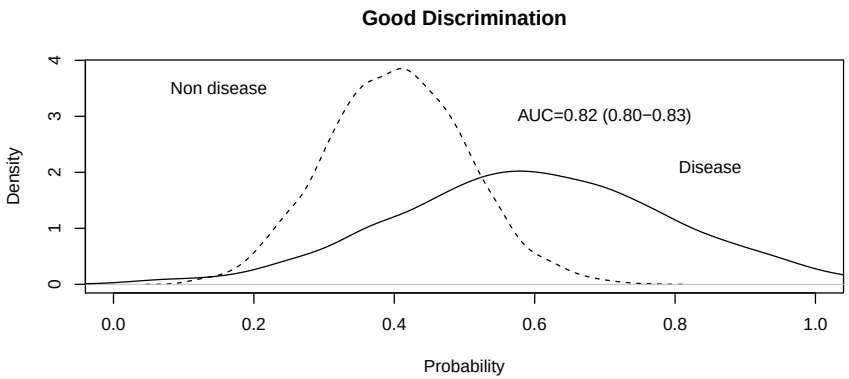
$$h_n = 0.9 \min\{SD, IQR/1.35\} n^{-1/5}$$

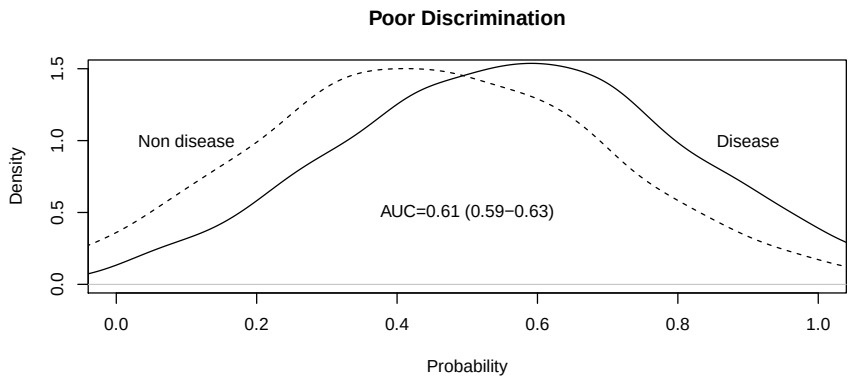
$$k(t) = \frac{15}{16}(1 - t^2)^2, \; with \, t \in (-1, 1)$$

# Different situations



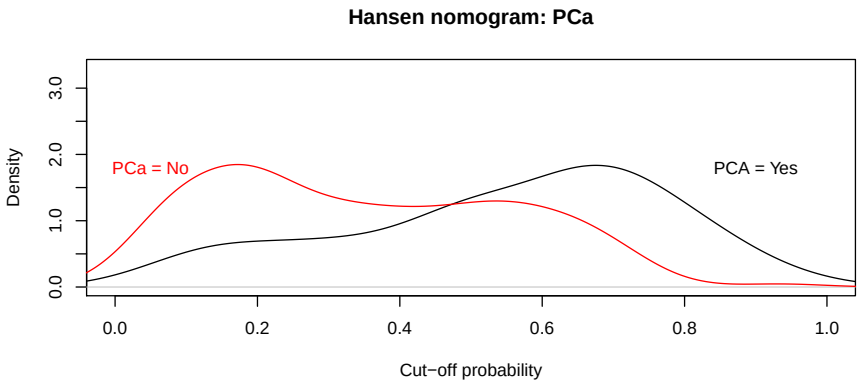



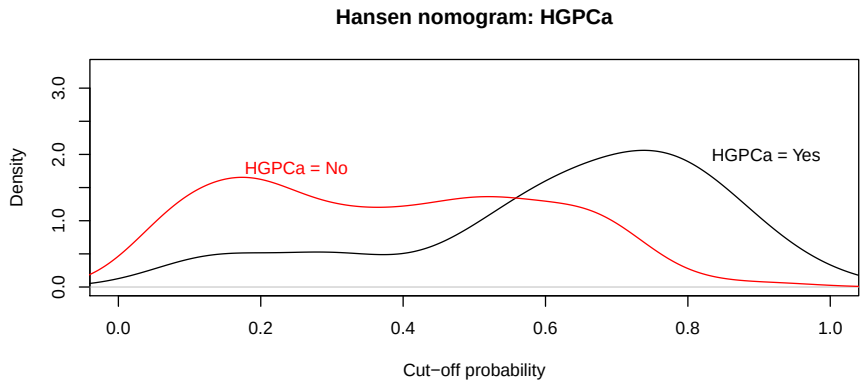



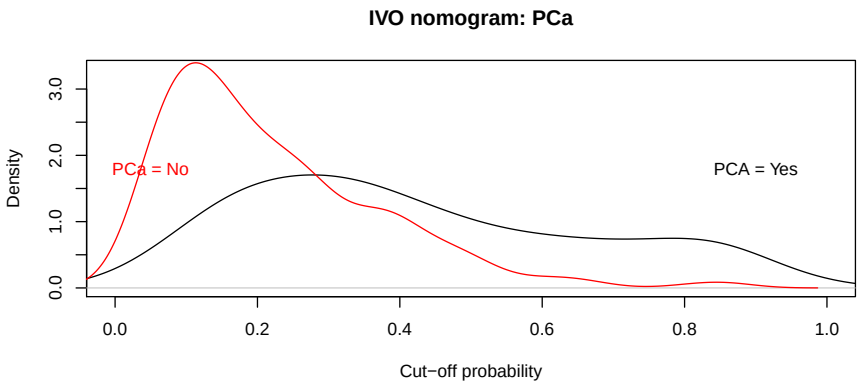



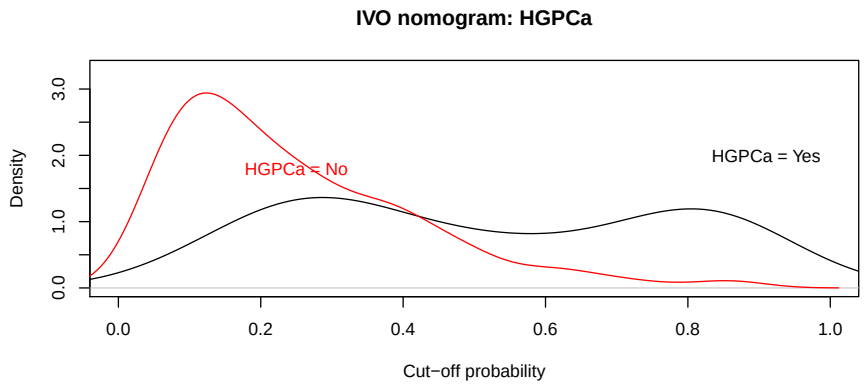

Universidad Zaragoza
1542

# Examples




Universidad Zaragoza

# Application in Prostate cancer. Real data

- Prediction of Organ-confined disease.
- Validation and comparison of the ability predictive and clinical utility of our models with some of the "golden rules" to predict organ-confined disease: Very good results.
- Proposal of the most reliable threshold point to use the estimated model (nomogram).
- Prediction of biochemical recurrence after radical prostatectomy.

# Summary

- Implementation in R code of a step by step algorithm for estimating linear classifiers with better discriminative ability.
- Modelling strategies in practical cases to increase the capacity of discrimination of classifiers. The global behaviour of the model is analysed (AUC for discrimination ability, concordance probabilities).
- Use of graphical methods, based on smoothing of density functions, for selection of cut-offs that increase the clinical utility of classifiers.
  We can analyse easily the changes in the utility and accuracy of the model changing the threshold.
- Applications in prostate cancer.

# References


Esteban, L. M., Sanz, G., Borque, A. (2011). A step-by-step algorithm for combining diagnostic tests. Journal of Applied Statistics, 38(5), 899-911.

Borque, Á., Esteban, L. M., Sanz, G. et al. (2013). Genetic predisposition to early recurrence in clinically localized prostate cancer. BJU international, 111(4), 549-558.

Borque, Á., Esteban, L.M., Sanz, G. et al. (2014). Implementing the use of nomograms by choosing threshold points in predictive models: 2012 updated Partin Tables vs a European predictive nomogram for organ confined disease in prostate cancer. BJU international, 113(6), 878-886.

López-Torrecilla, J., Esteban, L. M., Sanz, G. et al. (2015). Three linked nomograms for predicting biochemical failure in prostate cancer treated with radiotherapy plus androgen deprivation therapy. Strahlentherapie und Onkologie, 191(10), 792-800.

Rubio-Briones, J., Borque, A., Esteban, L.M, Sanz, G.et al. (2015). Optimizing the clinical utility of PCA3 to diagnose prostate cancer in initial prostate biopsy. BMC cancer, 15(1), 633.