Statistical Modeling of complex Data

Anne Gégout-Petit
(Université de Lorraine, Faculté des Sciences, IECL, Inria Nancy)
with the Benoît Liquet help !

4 Lectures on Model Selection and high-dimensional statistical analysis

1. Multiple testing issues
   - FWER: Family Wise Error Rate
   - False Discovery Rate
2. Model selection and assessment
   - Problematic, error
   - Criteria for linear model : AIC, BIC, Cp
   - Cross Validation and bootstrap method
   - Variable selection: subset
3. Regularization Methods
   - Ridge Regression
   - Lasso method
   - Elastic-net
4. Reduction method
   - Principal Component Analysis
   - Partial Least Square regression
   - Sparse Methods
   - Discriminant Analysis version

Plan

Introduction

Formalisation, error

Criteria by penalization

Cross-Validation methods

About the assessment of a model when Y is a binary variable

What is model selection ?

Examples

Model Selection: examples 1/5

Variables selection: subset selection
We have a model (multiple linear model here)

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

The purpose is to identify a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model using this subset.

- 2^p possible subset $\rightarrow$ choose between 2^p possible models
- case $n \geq p$ and $n < p$

Model Selection: examples 1/5

Variables selection: subset selection
We have a model (multiple linear model here)

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

The purpose is to identify a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model using this subset.

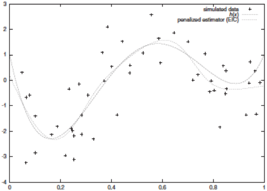
- 2^p possible subset → choose between 2^p possible models
- case $n \geq p$ and $n < p$

Our problem in this case is to select the appropriate variables. In this framework model selection means **variables selection** or subset selection.

Model Selection: examples. 2/5

Model:

$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$



Model Selection: examples. 2/5

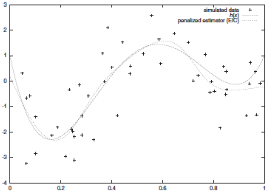
Model:

$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$

- Parametric approach: Polynomial estimator

$$h(X) = \beta_0 + \beta_1 X^1 + \ldots + \beta_d X^d$$

  - choice of the degree ?



Model Selection: examples. 2/5

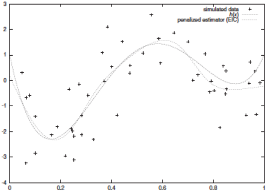
Model:

$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$

- Parametric approach: Polynomial estimator

$$h(X) = \beta_0 + \beta_1 X^1 + \ldots + \beta_d X^d$$

  - choice of the degree ?



In this framework model selection means **choice of the degree ie choice of the complexity of the model**.

Model Selection: examples. 3/5

Model:

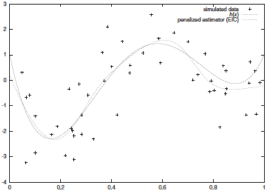
$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$

- non-parametric approach: h modelled by splines:

$$\hat{h}(x) = \sum_{j, s_j \in B} \beta_j s_j(x)$$

  - Which basis for the splines?

$$pen\mathcal{L}_\lambda(\mathcal{X}, \mathcal{Y}) = \ln(\mathcal{L}(\mathcal{X}, \mathcal{Y})) - \lambda \int h''^2(u) du$$



Model Selection: examples. 3/5

Model:

$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$

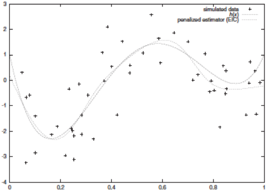
- non-parametric approach: h modelled by splines:

$$\hat{h}(x) = \sum_{j, s_j \in B} \beta_j s_j(x)$$

  - Which basis for the splines?

$$pen\mathcal{L}_\lambda(\mathcal{X}, \mathcal{Y}) = \ln(\mathcal{L}(\mathcal{X}, \mathcal{Y})) - \lambda \int h''^2(u)du$$

In this framework model selection means **choice of the base of splines** B



Model Selection: examples. 4/5

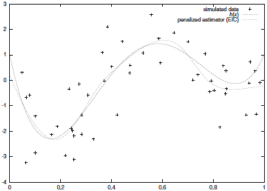
Model:

$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$

- non-parametric approach: kernel modellig

$$\hat{h}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{\lambda}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{\lambda})}$$

  - choice of the window parameter λ



Model Selection: examples. 4/5

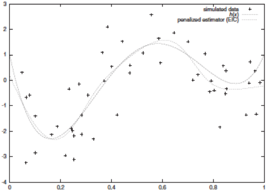
Model:

$$Y = h(X) + \varepsilon, \qquad h \text{ unknown}$$

- non-parametric approach: kernel modellig

$$\hat{h}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{\lambda}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{\lambda})}$$

  - choice of the window parameter λ



In this framework model selection means **choice of an hyperparameter of the model**

Model Selection: examples

Survival models (about T time to event of interest)

$$\lambda(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0^+} \mathbb{P}[t < T < t + \Delta t | T > t]$$

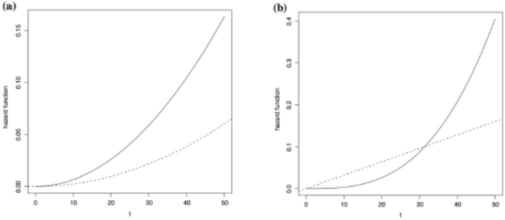
Adjusted Model (e.g. X = Gender)

- Proportional hazard model:

$$\lambda(t|X_i) = \lambda^0(t) \exp(\beta X_i) \qquad i = 1, \ldots, n$$

- Stratified model:

$$\lambda(t|X_i) = \begin{cases} \lambda^0(t) & \text{if } X_i = 0 \\ \lambda^1(t) & \text{if } X_i = 1 \end{cases}$$



Model Selection: examples

Survival models (about T time to event of interest)

$$\lambda(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0^+} \mathbb{P}[t < T < t + \Delta t | T > t]$$

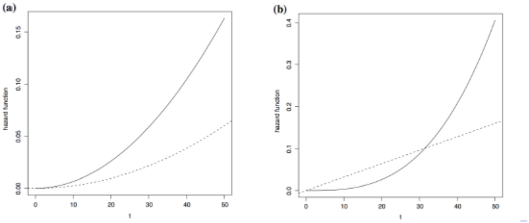
Adjusted Model (e.g. $X =$ Gender)

- Proportional hazard model:

$$\lambda(t|X_i) = \lambda^0(t) \exp(\beta X_i) \qquad i = 1, \ldots, n$$

- Stratified model:

$$\lambda(t|X_i) = \begin{cases} \lambda^0(t) & \text{if } X_i = 0 \\ \lambda^1(t) & \text{if } X_i = 1 \end{cases}$$



In this framework model selection means **choice between two different modelling**

Model Selection

- Model selection could have different meaning
- It could be the choice of the type of model (linear, polynomial, non-parametric, ...)
- For a given model, it could mean the choice of the variables to include in the model
- For a given model, it could mean the choice of the hyperparameters of the model
- And all together !

Plan

Introduction

Formalisation, error

Criteria by penalization

Cross-Validation methods

About the assessment of a model when Y is a binary variable

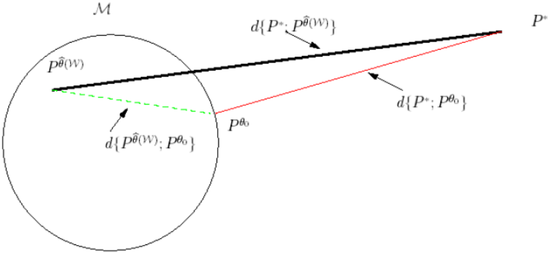
Model selection, formalisation

Model selection, formalisation

- Let $(\mathcal{X}, \mathcal{Y}) = \mathcal{W}$ be a random element $\simeq \mathbb{P}^*$ (density f^*)
- Principle: $\mathbb{P}^*$ estimated via a model $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$
- Observation about $\mathcal{W}$ (for instance a sample $(X_i, Y_i)_{1 \leq i \leq n}$) leads to $\mathbb{P}_{\hat{\theta}(\mathcal{W})}$

Model selection, formalisation

- Let $(\mathcal{X}, \mathcal{Y}) = \mathcal{W}$ be a random element $\simeq \mathbb{P}^*$ (density f^*)
- Principle: $\mathbb{P}^*$ estimated via a model $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$
- Observation about $\mathcal{W}$ (for instance a sample $(X_i, Y_i)_{1 \leq i \leq n}$) leads to $\mathbb{P}_{\hat{\theta}(\mathcal{W})}$


$\mathcal{M}$
$d\{P^*; P^{\hat{\theta}(\mathcal{W})}\}$
P^*
$P^{\hat{\theta}(\mathcal{W})}$
$d\{P^*; P^{\theta_0}\}$
P^{θ_0}
$d\{P^{\hat{\theta}(\mathcal{W})}; P^{\theta_0}\}$

- $d(\mathbb{P}^*, \mathbb{P}_{\theta_0})$: mis-specification risk
- $d(\mathbb{P}_{\theta(\hat{\mathcal{W}})}, \mathbb{P}_{\theta_0})$: statistical risk

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Prediction at point $X = x_0$: $\hat{h}^{(\mathcal{W})}(x_0)$
- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

Possible loss function

In regression

$$\begin{aligned} Loss(y, \hat{h}^{(\mathcal{W})}(x_0)) &= (y - \hat{h}^{(\mathcal{W})}(x_0))^2 \\ Loss(y, \hat{h}^{(\mathcal{W})}(x_0)) &= |y - \hat{h}^{(\mathcal{W})}(x_0)| \end{aligned}$$

In classification

$$\begin{aligned} Loss(y, \hat{g}^{(\mathcal{W})}(x_0)) &= \mathbf{1}_{y \neq \hat{f}^{(\mathcal{W})}(x_0)} \\ Loss(y, \hat{p}(Y = y|X)) &= -2\ln(\hat{p}(Y = y|X)) \end{aligned}$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

$$\mathbb{E}err(x_0) \quad = \quad \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

$$\begin{aligned}
\mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \left(\mathbb{E}[(\hat{h}^{(\mathcal{W})}(x_0)] - h(x_0)\right)^2 + \mathbb{E}[\left(\hat{h}^{(\mathcal{W})}(x_0) - \mathbb{E}[\hat{h}^{(\mathcal{W})}(x_0)]\right)^2] \\
&= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{h}^{(\mathcal{W})}(x_0)) + Var(\hat{h}^{(\mathcal{W})}(x_0)) \\
&= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

Example: k-nearest neighbor regression

$$\hat{h}(x_0) = \frac{1}{k} \sum_{l=1}^{k} Y_{(l)} \quad \text{where } x_{(1)} \ldots x_{(k)} \text{ are the } k \text{ nearest of } x_0$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

$$\begin{aligned}
\mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \Big(\mathbb{E}[(\hat{h}^{(\mathcal{W})}(x_0)] - h(x_0)\Big)^2 + \mathbb{E}[\Big(\hat{h}^{(\mathcal{W})}(x_0) - \mathbb{E}[\hat{h}^{(\mathcal{W})}(x_0)]\Big)^2] \\
&= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{h}^{(\mathcal{W})}(x_0)) + Var(\hat{h}^{(\mathcal{W})}(x_0)) \\
&= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \left[h(x_0) - \frac{1}{k}\sum_{l=1}^{k} h(x_{(l)})\right]^2 + \frac{\sigma_\varepsilon^2}{k}
\end{aligned}$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

$$\begin{aligned}
\mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \Big(\mathbb{E}[(\hat{h}^{(\mathcal{W})}(x_0)] - h(x_0)\Big)^2 + \mathbb{E}[\Big(\hat{h}^{(\mathcal{W})}(x_0) - \mathbb{E}[\hat{h}^{(\mathcal{W})}(x_0)]\Big)^2] \\
&= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{h}^{(\mathcal{W})}(x_0)) + Var(\hat{h}^{(\mathcal{W})}(x_0)) \\
&= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

Example: Linear Regression

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

$$\begin{aligned}
\mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \Big(\mathbb{E}[(\hat{h}^{(\mathcal{W})}(x_0)] - h(x_0)\Big)^2 + \mathbb{E}[\Big(\hat{h}^{(\mathcal{W})}(x_0) - \mathbb{E}[\hat{h}^{(\mathcal{W})}(x_0)]\Big)^2] \\
&= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{h}^{(\mathcal{W})}(x_0)) + Var(\hat{h}^{(\mathcal{W})}(x_0)) \\
&= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

Example: Linear Regression

$$\begin{aligned}
\mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + \Big[h(x_0) - \mathbb{E}[x_0^T \hat{\beta}]\Big]^2 + \sigma_\varepsilon^2 x_0^T (X'X)^{-1} x_0
\end{aligned}$$

Ponctuel prediction

$$Y = h(X) + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon] = 0, \quad var(\varepsilon) = \sigma_\varepsilon^2$$

- Expected prediction error at x_0:

$$\mathbb{E}err(x_0) = \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0]$$

$$f(y, \hat{y}) = (y - \hat{y})^2 \quad \text{is called the loss function}$$

$$\begin{aligned} \mathbb{E}err(x_0) &= \mathbb{E}[(Y - \hat{h}^{(\mathcal{W})}(x_0))^2 | X = x_0] \\ &= \sigma_\varepsilon^2 + \left(\mathbb{E}[(\hat{h}^{(\mathcal{W})}(x_0)] - h(x_0)\right)^2 + \mathbb{E}[\left(\hat{h}^{(\mathcal{W})}(x_0) - \mathbb{E}[\hat{h}^{(\mathcal{W})}(x_0)]\right)^2] \\ &= \sigma_\varepsilon^2 + \text{Bias}^2(\hat{h}^{(\mathcal{W})}(x_0)) + Var(\hat{h}^{(\mathcal{W})}(x_0)) \\ &= \text{Irreducible error} + \text{Bias}^2 + \text{Variance} \end{aligned}$$

Example: Linear Regression
The last term depends on x_0 but if we average on the sample $(x_i)_{1 \leq i \leq n}$:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}err(x_i) = \sigma_\varepsilon^2 + (\frac{1}{n}\sum_{i=1}^{n}\left[h(x_i) - \mathbb{E}[x_i^T\hat{\beta}]\right])^2 + \frac{p}{n}\sigma_\varepsilon^2$$

→ variance directly linked with the number of covariates p.

Examples

Framework: 80 observations and 20 predictors, uniformly distributed in the hypercube $[0,1]^{20}$.
Two cases:

- Y is 0 if $X_1 \leq 1/2$ and 1 if $X_1 > 1/2$, and we apply k-nearest neighbors,
- Y is 1 if $\sum_{j=1}^{p} X_j \geq 5$, 0 otherwise and we use best subset linear regression of size p to predict $P(Y=1|X)$
- Both problems are tackle by regression (prediction of probability to be one, MSE) or by classification (0-1 loss)

Examples: error k-NN

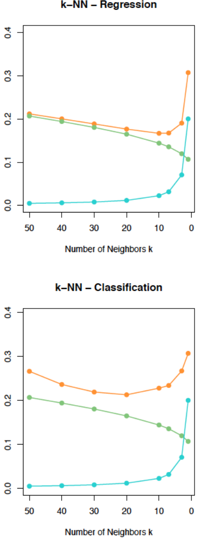


FIGURE 7.3. *Expected prediction error (orange), squared bias (green) and variance (blue) for a simulated example. The top row is regression with squared error loss; the bottom row is classification with 0–1 loss. The models are k-nearest neighbors (left) and best subset regression of size p (right). The variance and bias curves are the same in regression and classification, but the prediction error curve is different.*

Examples: error lin reg

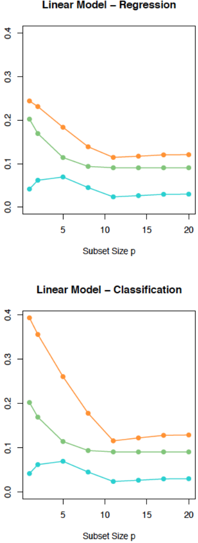


FIGURE 7.3. *Expected prediction error (orange), squared bias (green) and variance (blue) for a simulated example. The top row is regression with squared error loss; the bottom row is classification with 0–1 loss. The models are k-nearest neighbors (left) and best subset regression of size p (right). The variance and bias curves are the same in regression and classification, but the prediction error curve is different.*

Model Selection

How to measure the error of the model ?

Some theoretical definitions

Definition

Conditional error. Given a training set $\mathcal{T} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the generalization error of a model $\hat{f}$ is defined by

$$\mathbb{E}err(\mathcal{T}) = \mathbb{E}_{X^0, Y^0}\left[Loss(Y^0, \hat{f}(X^0)|\mathcal{T}\right]$$

- $\mathcal{T}$, training set used to estimate f by $\hat{f}$, $\mathcal{T}$ is fixed
- X^0, Y^0 means that it is a new test data point.

If we average on the training set, we obtain the **Expected Error**

$$\mathbb{E}err = \mathbb{E}_{\mathcal{T}}\left[\mathbb{E}_{X^0, Y^0}\left[Loss(Y^0, \hat{f}(X^0)|\mathcal{T}\right]\right]$$

Estimations

Definition

Training error.

$$\bar{\text{err}} = \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, \hat{f}(x_i))$$

for the particular loss $Loss(y_i, \hat{f}(x_i)) = (y_i - \hat{f}(x_i))^2$, $\bar{\text{err}}$ is the Residual Sum of Squares (RSS) divided by n,

$$\bar{\text{err}} = RSS/n = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

and its square root is called RMSE (Root Mean Square Errors)

$$RMSE = \sqrt{\bar{\text{err}}} = \sqrt{RSS/n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}$$

Estimations

Definition

Training error.

$$\bar{\mathrm{err}} = \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, \hat{f}(x_i))$$

Remark: Underestimated the $\mathbb{E}err(\mathcal{T})$ because we use the points of the training set to estimate $\hat{f}$.

Two solutions:

- Add a penalty to estimate $\mathbb{E}err(\mathcal{T})$ without bias
- Estimate it on another set or by bootstrap and cross validation

Plan

Introduction

Formalisation, error

Criteria by penalization

Cross-Validation methods

About the assessment of a model when Y is a binary variable

Estimation by adding a penalization : C_p; AIC, ...

Definition

In-Sample Prediction Error and optimism

$$\begin{aligned}
\mathbb{E}err_{\text{in}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y^0}\left[Loss(Y^0, \hat{f}(x_i))|\mathcal{T}\right] \\
\text{op} &= \mathbb{E}err_{\text{in}} - \bar{\text{err}} \qquad \text{(op for optimism)} \\
\omega &= \mathbb{E}_Y[\text{op}]
\end{aligned}$$

Proposition

For squared-error, $0-1$*, and other loss function, one can prove that*

$$op = \frac{2}{n}\sum_{i=1}^{n} cov(Y_i, \hat{Y}_i)$$

op depends on how y_i affects its own prediction. The harder we fit the data, the greater $cov(y_i, \hat{y}_i)$ will be.

Estimations

$$\mathbb{E}_Y[\mathbb{E}err_{\text{in}}] = \mathbb{E}_Y[\bar{\text{err}}] + \frac{2}{n}\sum_{i=1}^{n} cov(y_i, \hat{y}_i)$$

Estimations

$$\mathbb{E}_Y[\mathbb{E}err_{\text{in}}] = \mathbb{E}_Y[\bar{\text{err}}] + \frac{2}{n}\sum_{i=1}^{n} cov(y_i, \hat{y}_i)$$

Particular case of linear regression with p inputs or basis functions:

$$\sum_{i=1}^{n} cov(y_i, \hat{y}_i) = p\sigma_{\varepsilon}^2.$$

$$\mathbb{E}_Y[\mathbb{E}rr_{\text{in}}] = \mathbb{E}_Y[\bar{\text{err}}] + \frac{2p}{n}\sigma_{\varepsilon}^2 \tag{1}$$

In-sample prediction error

If we estimate In-sample prediction error on different models, we choose the model corresponding to the smallest one. But

$$\mathbb{E}err_{\text{in}} = \text{op} + \bar{\text{err}}$$

An obvious way to estimate in-sample prediction error is to estimate op and add $\bar{\text{err}}$. We have two families of methods

- methods based on (1) leading to criteria AIC, C_p and so on
- Estimation by bootstrap and cross validation

Estimate In-sample prediction error, C_p

$$\mathbb{E}err_{\text{in}} = \text{op} + \bar{\text{err}}$$

An obvious way to estimate in-sample prediction error is to estimate *op* and add $\bar{\text{err}}$.

$$\widehat{\mathbb{E}err_{\text{in}}} = \bar{\text{err}} + \hat{\text{op}}$$

If we use $\text{op} = \frac{2p}{n}\sigma_\varepsilon^2$ available for a class of additive models, we obtain the so-called C_p statistic:

Definition

C_p criterion

$$C_p = \bar{\text{err}} + \frac{2p}{n}\hat{\sigma}_\varepsilon^2$$

Estimate In-sample prediction error, AIC

The Akaike Information Criterion (AIC) is a similar but more generally applicable estimate of $\mathbb{E}err_{\text{in}}$ when a log-likelihood loss function is used. It is based on an relationship similar to (1) that holds asymptotically with n:

$$-2\mathbb{E}\left[\log f_{\hat{\theta}}(Y)\right] \sim -\frac{2}{n}\mathbb{E}\left[\text{loglik}\right] + \frac{2p}{n}\sigma_{\varepsilon}^{2}$$

Estimate In-sample prediction error, AIC

$$-2\mathbb{E}\left[\log f_{\hat{\theta}}(Y)\right] \sim -\frac{2}{n}\mathbb{E}\left[\text{loglik}\right] + \frac{2p}{n}\sigma_{\varepsilon}^{2}$$

where

- f_θ is a family of densities for Y including the true one,
- $\hat{\theta}$ the maximum likelihood estimator
- loglik is the maximized log-likelihood

$$L = \text{loglik} = \sum_{i=1}^{n} log(f_{\hat{\theta}}(y_i))$$

Estimate In-sample prediction error, AIC

$$-2\mathbb{E}\left[\log f_{\hat{\theta}}(Y)\right] \sim -\frac{2}{n}\mathbb{E}\left[\text{loglik}\right] + \frac{2p}{n}\sigma_{\varepsilon}^{2}$$

Thus, AIC is defined by

$$AIC/n = -\frac{2}{n}L + \frac{2p}{n}$$

or by equivalence

Definition

Akaike information criterion (AIC)

$$AIC = -2L + 2p$$

Selection model, AIC

As AIC/n estimates the model's error,

$$-2\mathbb{E}\left[\log f_{\hat{\theta}}(Y)\right] \sim -\frac{2}{n}\mathbb{E}\left[\text{loglik}\right] + \frac{2p}{n}\sigma_{\varepsilon}^{2}$$

→ **we choose the model with smallest AIC**.

Selection model, AIC

As AIC/n estimates the model's error,

$$-2\mathbb{E}\left[\log f_{\hat{\theta}}(Y)\right] \sim -\frac{2}{n}\mathbb{E}\left[\text{loglik}\right] + \frac{2p}{n}\sigma_{\varepsilon}^{2}$$

→ **we choose the model with smallest AIC**.
In gaussian linear model C_p and AIC are equivalent

But what if p is not well defined ?

Selection model, AIC

As AIC/n estimates the model's error,

$$-2\mathbb{E}\left[\log f_{\hat{\theta}}(Y)\right] \sim -\frac{2}{n}\mathbb{E}\left[\text{loglik}\right] + \frac{2p}{n}\sigma_{\varepsilon}^{2}$$

→ **we choose the model with smallest AIC**.

But what if p is not well defined ?

Need to define the "effective number of parameters"

AIC, Effective number of parameters

Definition

Effective number of parameter *df*

When the predictor is linear:

$$\hat{y} = Sy$$

with S an $n \times n$ matrix depending only on the x_i, then the effective number of parameters is

$$df(S) = \text{trace}(S)$$

Note that in the linear regression model,

$$\hat{y} = \underbrace{(X'X)^{-1}X'}_{=S} y \qquad \text{and } \text{trace}(S) = p$$

AIC, Effective number of parameters

Definition

Effective number of parameter *df*

When the predictor is linear:

$$\hat{y} = Sy$$

with S an $n \times n$ matrix depending only on the x_i, then the effective number of parameters is

$$df(S) = \mathrm{trace}(S)$$

Note that in the linear regression model,

$$\hat{y} = \underbrace{(X'X)^{-1}X'}_{=S} y \qquad \text{and trace}(S) = p$$

If the model is additive: $Y = f(X) + \varepsilon$,

$$df(\hat{y}) = \frac{\sum_{i=1}^{n} cov(y_i, \hat{y}_i)}{\sigma_\varepsilon^2}$$

Selection model, BIC

Definition

BIC for Bayeisan Information Criteria

$$BIC = -2L + p\ln(n)$$

One can show that in the gaussian model, BIC, is proportional with AIC and C_p with factor 2 replaced by $\ln(n)$.
BIC penalizes complex models more heavily, preferring simpler models.

Selection model, BIC

Definition

BIC for Bayeisan Information Criteria

$$BIC = -2L + p\ln(n)$$

One can show that in the gaussian model, BIC, is proportional with AIC and C_p with factor 2 replaced by $\ln(n)$.
BIC penalizes complex models more heavily, preferring simpler models.

BIC is motivated by Bayesian choice of model, BIC will choose the model that optimises the posterior probability when the prior is uniform.

Plan

Introduction

Formalisation, error

Criteria by penalization

Cross-Validation methods

About the assessment of a model when Y is a binary variable

Cross validation methods

Mainly the most widely used method

Princip:

- Use a "training" set to estimate the model
- Use a independent "test" or validation set, from the joint distribution (Y, X) to estimate error $\mathbb{E}rr$,
- several possibilities

Cross validation methods

Split the original dataset in two sets: one for the training called the "train set", the second for the validation of the model fitted by train set, called the "test set"

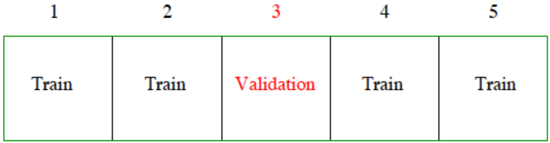
Princip: from the original dataset

- Choose Randomly a subset for the training (according to the size of the sample, it could be from 65% to 80 % of the data)
- fit the model on the "train set"
- evaluate the model on the "test set" (that is the complementary of the train set)
- Eventually, repeat the operations B times and see the variability of the evaluation's parameter

K-Fold Cross

Princip:

- Split your data in K subsets → leading to a K-partition of the data



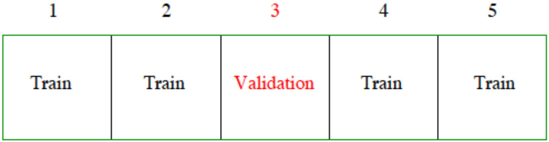
1	2	3	4	5
Train	Train	Validation	Train	Train

- For $k = 1$ to K, choose the k-th subset as validation test and the $K - 1$ union of the others as training set

K-Fold Cross

Princip:

- Split your data in K subsets → leading to a K-partition of the data



1	2	3	4	5
Train	Train	Validation	Train	Train

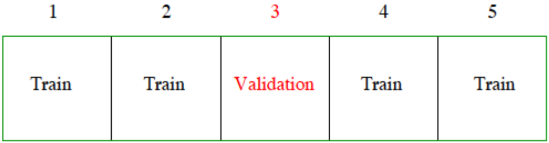
- For $k = 1$ to K, choose the k-th subset as validation test and the $K - 1$ union of the others as training set
- For $k = 1$ to K, predict $\hat{h}^{-k}$ and calculate the prediction error when predicting the k-th subset

$$CV\mathbb{E}rr = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{h}^{-k(i)}(x_i))^2$$

where $\hat{h}^{-k(i)}$ is the prediction of the subject i based in a model fitted with the $k(i)$-th part of the data removed

B - *K*-Fold Cross validation

Choose B K-partitions randomly



1	2	3	4	5
Train	Train	Validation	Train	Train

$$BCV\mathbb{E}rr = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{h}^{-k_b(i)}(x_i))^2\right)$$

B - *K*-Fold Cross validation

Choose B K-partitions randomly

1	2	3	4	5
Train	Train	Validation	Train	Train

In the context of model selection, we have to do it for

$$BCV\mathbb{E}rr(\alpha) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{h}^{-k_b(i)}(x_i, \alpha))^2\right)$$

Choose the model corresponding to α that minimizes $BCV\mathbb{E}rr(\alpha)$
very time consuming !

Cross validation

- If $K = n$, n-Fold CV is called leave-one-out CV
- How to choose K for K-Fold cross validation method ?

Cross validation

- If $K = n$, n-Fold CV is called leave-one-out CV
- How to choose K for K-Fold cross validation method ?
- with lower K, CV $\mathbb{E}rr$ has a low variance to estimate the error but bias could be a problem
- with $K = n$, it is almost unbiased but variance is high (the training set are almost the same !)

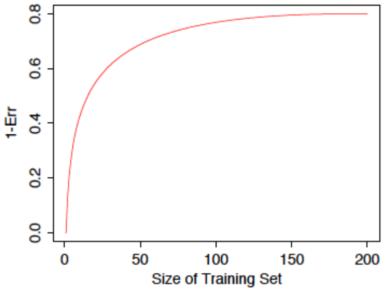


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of* $1 - \mathrm{Err}$ *versus the size of the training set* N*. With a dataset of* 200 *observations,* 5*-fold cross-validation would use training sets of size* 160*, which would behave much like the full set. However, with a dataset of* 50 *observations fivefold cross-validation would use training sets of size* 40*, and this would result in a considerable overestimate of prediction error.*

Cross validation

- If $K = n$, n-Fold CV is called leave-one-out CV
- How to choose K for K-Fold cross validation method ?
- with lower K, CV $\mathbb{E}rr$ has a low variance to estimate the error but bias could be a problem
- with $K = n$, it is almost unbiased but variance is high (the training set are almost the same !)
- The most common are 5-Fold and 10-Fold, but if the data set is small ... leave one out is better !

Bootstrap methods

General princip

- Denote the dataset by $Z = (z_1 \ldots, z_n)$ where $z_i = (x_i, y_i)$,
- randomly draw a training dataset with replacement from original data
- This is done B times (e.g. $B = 1000$)
- Refit the model to each of the bootstrap datasets and examine the behavior over the B replications
- From the bootstrap sample, we can estimate any aspect of the distribution of $S(Z)$, where $S(z)$ can be any quantity computed from the data

Bootstrap methods

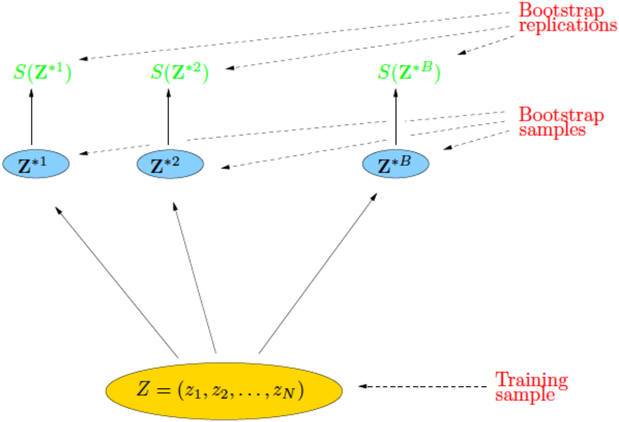


FIGURE 7.12. *Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity $S(\mathbf{Z})$ computed from our dataset. B training sets $\mathbf{Z}^{*b}$, $b = 1, \dots, B$ each of size N are drawn with replacement from the original dataset. The quantity of interest $S(\mathbf{Z})$ is computed from each bootstrap training set, and the values $S(\mathbf{Z}^{*1}), \dots, S(\mathbf{Z}^{*B})$ are used to assess the statistical accuracy of $S(\mathbf{Z})$.*

Bootstrap in selection model

Estimating the prediction error

Bootstrap in selection model

Estimating the prediction error

- First idea:

$$\hat{\mathbb{E}rr}_{boot} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\left(\sum_{i=1}^{n}(y_i - \hat{h}^{*b}(x_i))^2\right)$$

- $\hat{\mathbb{E}rr}_{boot}$ does not provide a good estimate
  - Bootstrap dataset is acting as both training and testing (common observations)
  - The overfit predictions will look unrealistically good
- By mimicking CV, better bootstrap estimates
- Only keep track of predictions from bootstrap samples not containing the observations

Bootstrap in selection model

The leave-one-out bootstrap estimate of prediction error

$$\hat{\mathbb{E}rr}^{(1)}_{boot} = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} (y_i - \hat{h}^{*b}(x_i))^2\right)$$

- C^{-i} is the set of indices of the bootstrap sample b that do not contain observation i.
- We either have to choose B large enough to ensure that all of $|C^{-i}|$ is greater than zero, or just leave-out the terms that correspond to $|C^{-i}|$'s that are zero.

The .632 estimator !

The leave-one-out bootstrap estimate of prediction error Why this name ?

Probability for observation i to be in the bootstrap sample b ?

The .632 estimator !

The leave-one-out bootstrap estimate of prediction error Why this name ?

Probability for observation i to be in the bootstrap sample b ? $\hat{\mathbb{E}rr}^{(1)}_{boot}$ suffers from bias due to training-set-size bias. Some learners have proposed the ".632 estimator " defined by

$$\hat{\mathbb{E}rr}^{(.632)}_{boot} = 0.368\bar{err} + 0.632\hat{\mathbb{E}rr}^{(1)}_{boot}$$

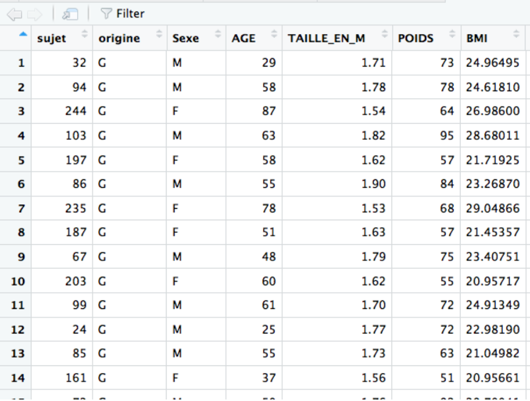
Best subset selection

Best subset selection

Best subset selection

A practical example

- Medical context: explain physiological variable TLCO
- Potential Covariates



	sujet	origine	Sexe	AGE	TAILLE_EN_M	POIDS	BMI
1	32	G	M	29	1.71	73	24.96495
2	94	G	M	58	1.78	78	24.61810
3	244	G	F	87	1.54	64	26.98600
4	103	G	M	63	1.82	95	28.68011
5	197	G	F	58	1.62	57	21.71925
6	86	G	M	55	1.90	84	23.26870
7	235	G	F	78	1.53	68	29.04866
8	187	G	F	51	1.63	57	21.45357
9	67	G	M	48	1.79	75	23.40751
10	203	G	F	60	1.62	55	20.95717
11	99	G	M	61	1.70	72	24.91349
12	24	G	M	25	1.77	72	22.98190
13	85	G	M	55	1.73	63	21.04982
14	161	G	F	37	1.56	51	20.95661

Best subset selection

Best Subset Selection: we fit a separate least square regression for each possible combination of the p predictors !

- Fit all p models that contains exactly one predictor
- Fit all $\binom{2}{p}$ models that contain exactly two predictors
- and so on ...
- To choose the best model we use a two steps algorithm

Best subset selection

Algorithm:

1. Let $\mathcal{M}_0$ denote the null model without predictor.
2. For $k = 1, \ldots, p$
    1. Fit $\binom{k}{p}$ models that contain exactly k predictors,
    2. Pick the best among these k models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, Cp, AIC, BIC, or adjusted R^2.
4. very long when p is large !
   $p = 10 \rightarrow 1000$ possible models
   $p = 20 \rightarrow$ one million of models

Forward stepwise selection

Algorithm:

1. Let $\mathcal{M}_0$ denote the null model without predictor.
2. For $k = 0, \ldots, p - 1$
    2.1 Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
    2.2 Choose the best among these models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, Cp, AIC, BIC, or adjusted R^2.

Forward stepwise selection

Algorithm:

1. Let $\mathcal{M}_0$ denote the null model without predictor.
2. For $k = 0, \ldots, p-1$
   - 2.1 Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   - 2.2 Choose the best among these models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, Cp, AIC, BIC, or adjusted R^2.

$$\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2 \qquad \text{visited models}$$

$p = 20 \rightarrow 211$ visited models.

Backward stepwise selection

Algorithm:

1. Let $\mathcal{M}_p$ denote the full model with p predictors.
2. For $k = p, \ldots, 1$
   2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$ for a total of $k - 1$ predictors.
   2.2 Choose the best among these models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p, AIC, BIC, or adjusted R^2.

$$\sum_{k=0}^{p-1} (p - k) = 1 + p(p + 1)/2 \qquad \text{visited models}$$

$p = 20 \rightarrow 211$ visited models.

Backward stepwise selection

Algorithm:

1. Let $\mathcal{M}_p$ denote the full model with p predictors.
2. For $k = p, \ldots, 1$
   - 2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$ for a total of $k-1$ predictors.
   - 2.2 Choose the best among these models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p, AIC, BIC, or adjusted R^2.

$$\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2 \qquad \text{visited models}$$

$p = 20 \rightarrow 211$ visited models.

- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit).
- Forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Exercice

In the linear regression model, write the log likelihood and compute an explicit value of the AIC, BIC and C_p

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

Plan

Introduction

Formalisation, error

Criteria by penalization

Cross-Validation methods

About the assessment of a model when Y is a binary variable

About the assessment of a model when Y is a binary variable

Be careful with the loss function $Loss(y, \hat{g}^{(\mathcal{W})}(x_0)) = \mathbf{1}_{y \neq \hat{f}^{(\mathcal{W})}(x_0)}$ when evaluating

$$\mathbb{E}err(x_0) = \mathbb{E}[Loss(y, \hat{f}(x_i))|X = x_0]$$

by

$$\bar{\text{err}} = \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, \hat{f}(x_i)).$$

Indeed, if in the sample $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is very near from 0 (or 1) then without a model it is very easy to have a good $\bar{\text{err}}$ by predicting everybody by $\hat{f}(x_i) = 0$. In this case, $\bar{\text{err}} = \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, \hat{f}(x_i)) = 1 - \hat{p}$!

About the assessment of a model when Y is a binary variable

Some elements to assess the model :

- Don't forget the AIC and BIC based on the log-lokelihood !
- Elements from the confusion matrix
- ROC curve
- Recall-precision curve

Confusion matrix and derived parameters

Confusion matrix

Prevision $\hat{Y}$ Truth Y	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	n

Withe $n = a + b + c + d$.

- Error rate: e$\bar{\mathrm{r}}$r $= \frac{b+c}{n}$
- Sensibility (or recall) $S_e = TPR = \frac{a}{a+b}$
- Precision $\frac{a}{a+c}$ (a sort of 1 -FDR !)
- Specificity $S_p = TNR = \frac{d}{c+d} = 1 - FPR = 1 - \frac{c}{c+d}$

Confusion matrix and derived parameters

Confusion matrix

Prevision $\hat{Y}$ Truth Y	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	n

Withe $n = a + b + c + d$.

- Error rate: $\bar{err} = \frac{b+c}{n}$
- Sensibility (or recall) $S_e = TPR = \frac{a}{a+b}$
- Precision $\frac{a}{a+c}$ (a sort of 1 -FDR !)
- Specificity $S_p = TNR = \frac{d}{c+d} = 1 - FPR = 1 - \frac{c}{c+d}$
- A good model presents high values of S_e, S_p, precision and low values of $\bar{err}$, and $FPR = \frac{c}{a+c}$
- $\bar{err}$ is symetric and gives the same importance to FP c and FN b
- Sensibility and Precision give a more importance to positive values
- When the sensibility increases, the specificity decreases ! A model that is better for this two values has to be chosen.

Confusion matrix and derived parameters

Confusion matrix

Prevision $\hat{Y}$ Truth Y	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	n

Withe $n = a + b + c + d$.

- Error rate: $\bar{err} = \frac{b+c}{n}$
- Sensibility (or recall) $S_e = TPR = \frac{a}{a+b}$
- Precision $\frac{a}{a+c}$ (a sort of 1 -FDR !)
- Specificity $S_p = TNR = \frac{d}{c+d} = 1 - FPR = 1 - \frac{c}{c+d}$
- F-measure that is an harmonic mean of the recall and precision

$$F_\beta = \frac{(1 + beta^2) S_e \times Precision}{\beta^2 precision + S_e}$$

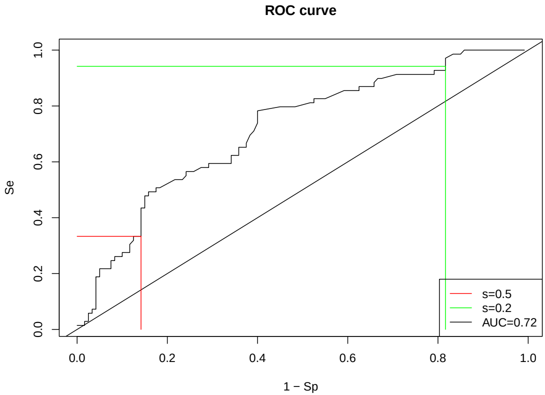
  used very frequently with $\beta = 1$

  - If $\beta < 1$, more importance is given to recall
  - If $\beta > 1$, more importance is given to precision

ROC curve

If model estimates $\hat{p}_i = P(Y_i = 1|x_i)$, the rule is to predict Y_i by $\hat{Y}_i = 1$ if $\hat{p}_i > 0.5$ and 0 otherwise.
But we can choose another threshold: $\hat{Y}_i = \mathbf{1}_{\hat{p}_i > s}$

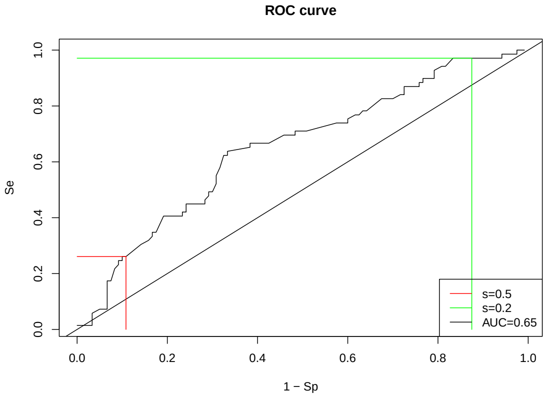
The idea of the ROC curve is to draw the False Positive Rate (FPR or S_e) according to the True Positive Rate (TPR or 1-S_p) by computing these values for a continuum of threshold s.



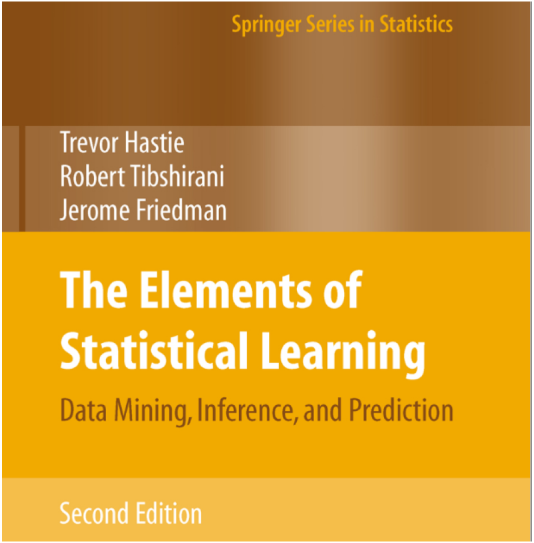
ROC curve

If model estimates $\hat{p}_i = P(Y_i = 1|x_i)$, the rule is to predict Y_i by $\hat{Y}_i = 1$ if $\hat{p}_i > 0.5$ and 0 otherwise.
But we can choose another threshold: $\hat{Y}_i = \mathbf{1}_{\hat{p}_i > s}$

The idea of the ROC curve is to draw the False Positive Rate (FPR or S_e) according to the True Positive Rate (TPR or 1-S_p) by computing these values for a continuum of threshold s.



References



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

pdf available on internet