Institut Paoli-Calmettes

Centre régional de lutte contre le cancer Provence-Alpes-Côte d'Azur

L'overfitting se cache derrière vos prédictions. Trouvez-le et mesurez-le.

Benjamin Esterni, Jean-Marie Boher

Epiclin – Journées des statisticiens des centres

28 mai 2010

Construction d'une règle de prédiction

Jeu d'apprentissage

$$X = (x_{1,} x_{2,} \ldots, x_n)$$

n observations

$$x_i = (t_i, y_i)$$

Règle de prédiction

$$r_X(t)$$

Erreur de prédiction

$$Err(r_X) = E[(y - r_X(t))^2]$$

Estimateurs

$$\overline{Err}, \quad \widehat{Err_{CV}}, \quad \widehat{Err_{BO}}, \quad \widehat{Err_{632}}, \quad \widehat{Err_{632+}}$$

Erreur apparente

$$\overline{Err} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - r_X(t_i)\right)^2$$

$$X = (x_1, x_2, \ldots, x_n)$$

Erreur de la règle $\mathbf{r}_X$, *construite sur* $\mathbf{X}$ *et appliquée sur* $\mathbf{X}$

→ *risque d'optimisme*

Erreur « Cross-Validation »

X est divisé en K sous-ensembles disjoints

$$X = X_1 \cup X_2 \cup \ldots \cup X_K, \cap X_j = \emptyset$$

$$X_j^c = X \setminus X_j$$

$$\widehat{Err}_{CV} = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{n_K} \left(y_i - r_{X_j^c}(t_i) \right)^2$$

Erreur moyenne des règles r_{X_j}, *construites sur* X_j *et appliquées sur* X_j^c

→ *risque de pessimisme*

Erreur « Bootstrap »

Échantillon bootstrap :

tirage aléatoire avec remise de n observations parmi X = (x_1,..,x_n)

$$X_b^* = \{x_{b\cdot 1}^*, x_{b\cdot 2}^*, \ldots, x_{b\cdot n}^*\}, \quad x_{b\cdot j}^* \in X$$

$$X_b^{*c} = \{x_i : x_i \notin X_b^*\}$$

$$\widehat{Err}_{BO} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n_b^{*c}} \left(y_i - r_{X_b^*}(t_i)\right)^2$$

Erreur moyenne des règles $r_{X_b^*}$, *construites sur* X_b^* *et appliquées sur* X_b^{*c}

→ *risque de pessimisme*

A chaque tirage bootstrap:

$$X_b^* = \{x_{b\cdot 1}^*, x_{b\cdot 2}^*, \ldots, x_{b\cdot n}^*\}, \quad x_{b\cdot j}^* \in X$$

$$X_b^0 = \{x_i : x_i \notin X_b^*\}$$

$$P(x \in X_b^*) = 1 - P(x \notin X_b^*) = 1 - \left(1 - \frac{1}{n}\right)^n$$

$$\lim_{n \to +\infty} \left(1 - \left(1 - \frac{1}{n}\right)^n\right) = 1 - e^{-1} = \underline{0.632}$$

Erreur « 632 »

Efron, 1983

$$\widehat{Err}_{632} = \underbrace{0.368}_{1-0.632} \cdot \overline{Err} + 0.632 \cdot \widehat{Err}_{BO}$$

En présence d'un fort surapprentissage
et d'une réponse indépendante des mesures

$$\overline{Err} = 0 \quad \Rightarrow \quad \widehat{Err}_{632} = 0.632\,\widehat{Err}_{BO} < E[(y - r_X)^2]$$

Mesure du surapprentissage

Efron & Tibshirani, 1997

Erreur de non-information

$$\hat{\gamma} = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(y_i - r_X(t_j))^2$$

Taux de surapprentissage

$$\hat{R} = \frac{\widehat{Err}_{BO} - \overline{Err}}{\hat{\gamma} - \overline{Err}}$$

Erreur « 632+ »

Efron & Tibshirani, 1997

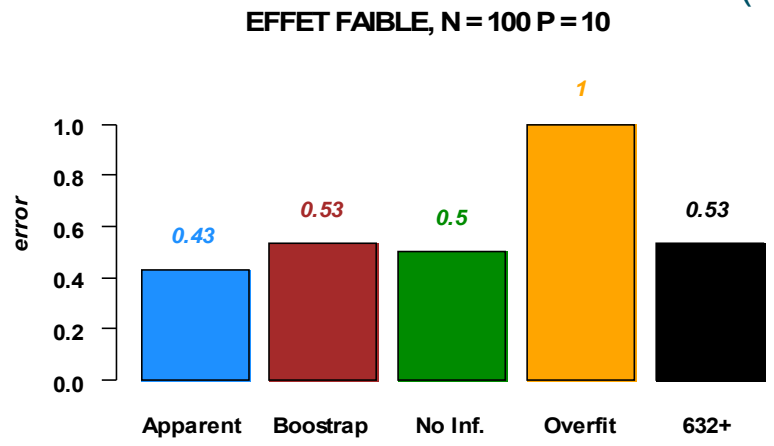
Erreur corrigée

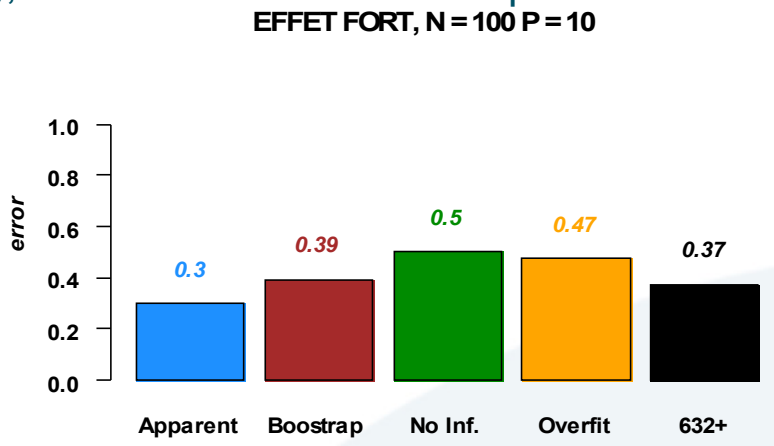
$$\hat{R} = \frac{\widehat{Err}_{BO} - \overline{Err}}{\hat{\gamma} - \overline{Err}} \quad , \quad \hat{\omega} = \frac{0.632}{1 - 0.368\,\hat{R}}$$

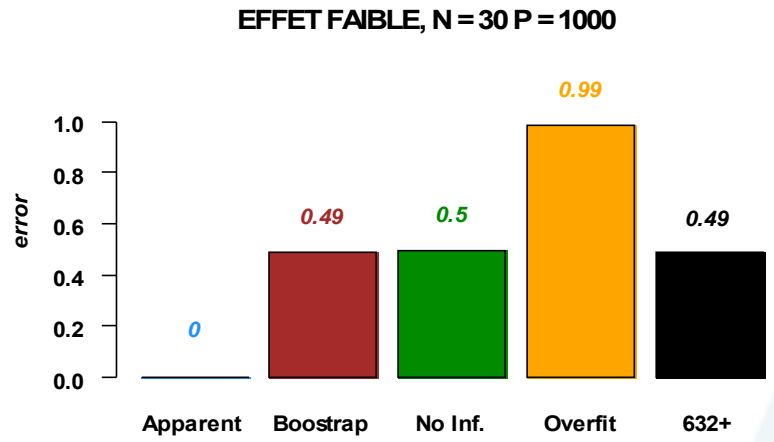
$$\widehat{Err}_{632+} = (1-\hat{\omega}) \cdot \overline{Err} + \hat{\omega} \cdot \widehat{Err}_{BO} = \begin{cases} \widehat{Err}_{632} & si \quad \hat{R} = 0 \\ \widehat{Err}_{BO} & si \quad \hat{R} = 1 \end{cases}$$

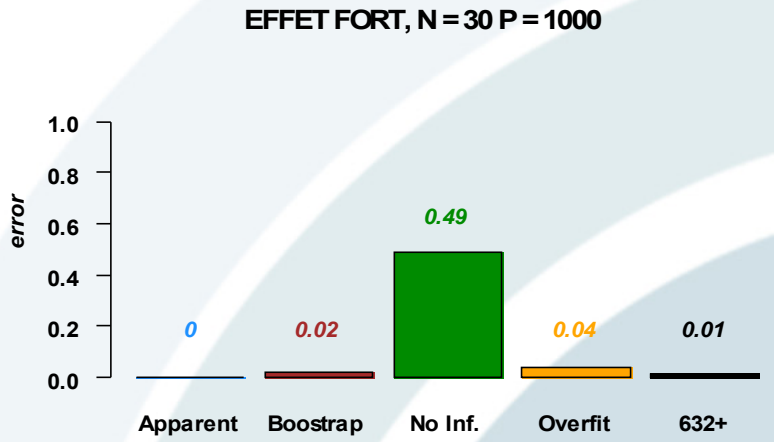
Simulations avec une règle simple

Sélection des 5 meilleures covariables (t-test), distance euclidienne aux profils médians









Extension aux données de survie

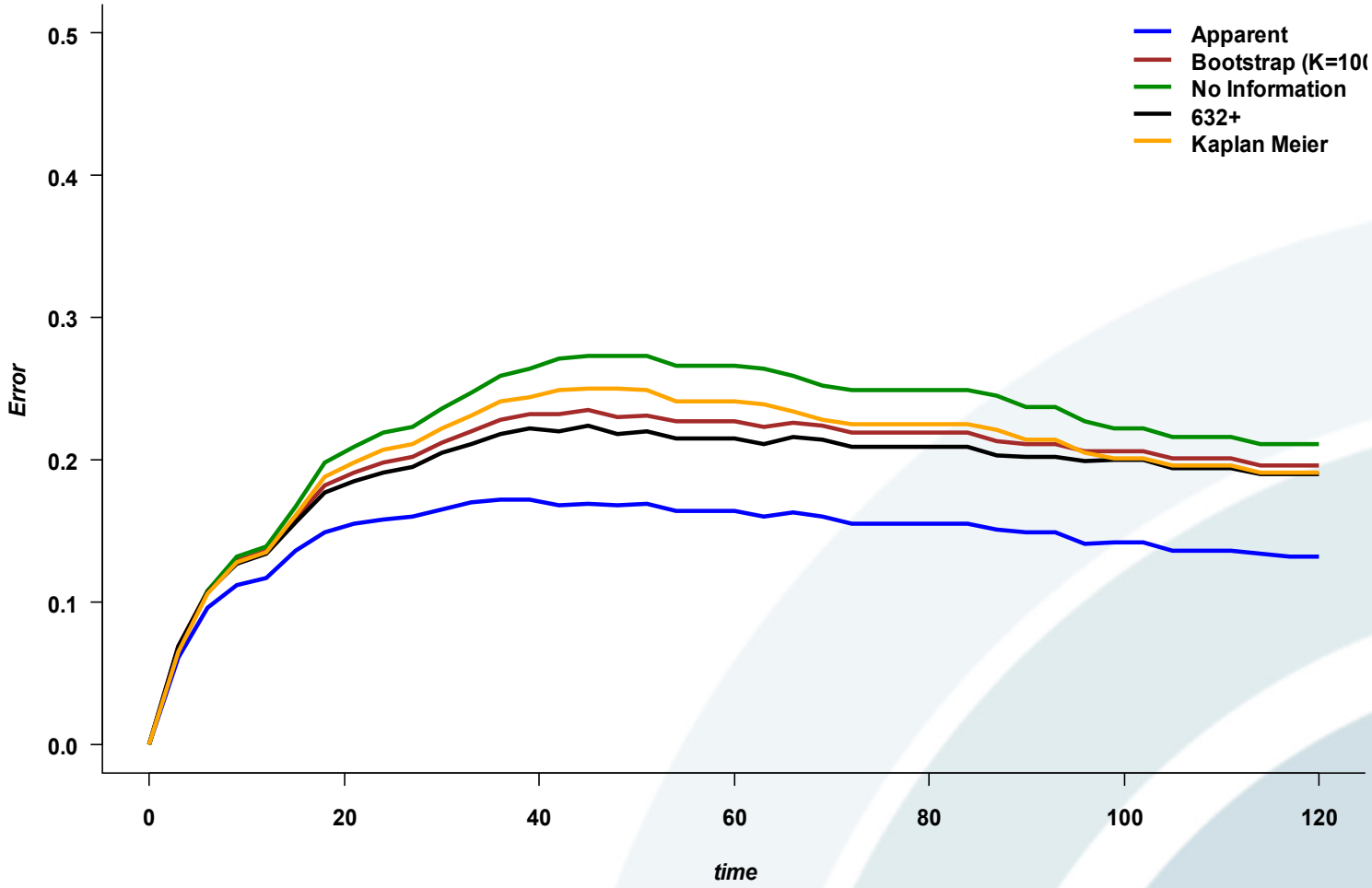
Gerds & Schumacher, 2007

$$\overline{err}(t,r)=\frac{1}{n}\sum_{i=1}^{n}\{Y_i(t)-r_n(t|Z_i)\}^2 W(t,X_i)$$

$$\text{où} \qquad W(t,X_i)=\begin{cases}(P(C>t|Z_i))^{-1} & si \quad \tilde{T}_i>t \\ (\Delta_i\cdot P(C>\tilde{T}_i|Z_i))^{-1} & si \quad \tilde{T}_i\leq t\end{cases}$$

$$\tilde{T}_i=min(T_i,C_i), \quad \Delta_i=I_{T_i\leq C_i}, \quad Z_i=covariates$$

Extension aux données de survie



Références

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation.

Journal of the American Statistical Association 78, 316–331.

Efron, B. Tibshirani, R. (1997). Improvements on crossvalidation: The 0.632+ bootstrap method.

Journal of the American Statistical Association 92, 548–560.

Gerds, T. A. Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times.

Biometrical Journal 48, 1029–1040.

Gerds, T. A. Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis.

Biometrics 63, 1283–1287.