Data Science for Actuaries (ACT6100)

Arthur Charpentier

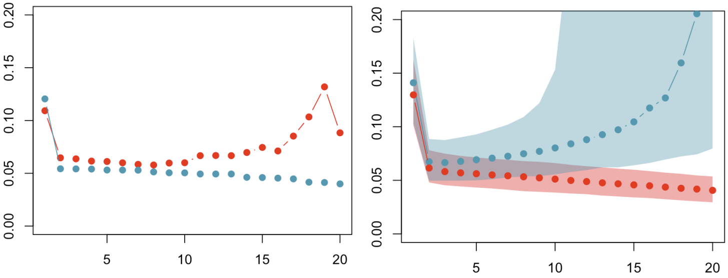
Supervisé # 1.2 (Concepts Fondamentaux - 2)

automne 2020

https://github.com/freakonometrics/ACT6100/

Base d'entrainement et base de validation

- Tel que mentionné précédemment, il est possible de diviser la base de données initiale en une base d'entrainement ($\sim 70\%$) et une base de validation ($\sim 30\%$).



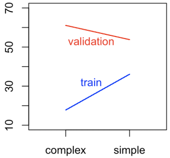
Base d'entrainement et base de validation

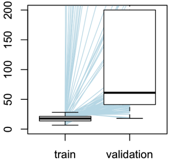
Classical Approach : split the sample $\mathcal{D}_n$ in two parts

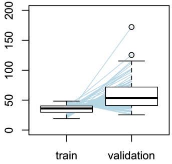
Hold-Out Cross Validation

1. Split $\{1, 2, \cdots, n\}$ in T (training) and V (validation)
2. Estimate $\widehat{m}$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in T$: $\widehat{m}_T$
3. Compute $\displaystyle\frac{1}{|V|}\sum_{i\in V}\ell\big(y_i, \widehat{m}_T(x_{1,i}, \cdots, X_{p,i})\big)$

```
chicago = read.table("http://freakonometrics.free.fr/chicago.txt",header=TRUE,sep=";")
idx = sample(1:nrow(chicago), nrow(chicago)*.7)
train = chicago[idx,]
valid = chicago[-idx,]
```



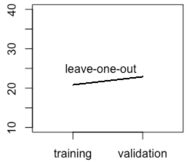




Validation croisée, Leave-One-Out

Leave-one-Out Cross Validation

1. Estimate n models : $\widehat{m}_{-j}$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in \{1, \cdots, n\} \backslash \{j\}$
2. Compute $\displaystyle \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \widehat{m}_{-i}(\boldsymbol{x}_i))$



Can be computationally intensive...

In the case of a linear regression, there is a simple formula to compute $\widehat{\boldsymbol{\beta}}_{-j}$ when observation j is removed. Let $\boldsymbol{H} = \mathcal{P}_{\mathcal{V}(\boldsymbol{X})} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$

$$(\boldsymbol{X}_{(j)}^\top \boldsymbol{X}_{(j)})^{-1} \boldsymbol{X}_{(j)}^\top = \frac{1}{1 - H_{j,j}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_{(j)}^\top$$

Validation croisée, k-fold

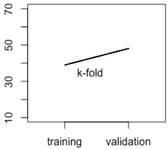
Instead of fitting n models, fit only K

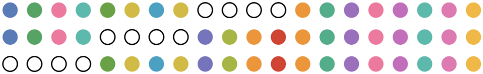
K-Fold Cross Validation

1. Split $\{1, 2, \cdots, n\}$ in K groups $V_1, \cdots, V_K$
2. Estimate K models : $\widehat{m}_k$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in \{1, \cdots, n\} \backslash V_k$
3. Compute $\displaystyle\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|V_k|}\sum_{i\in V_k}\ell(y_i, \widehat{m}_k(x_{1,i}, \cdots, x_{p,i}))$

If $K = 10$ we fit on 90% of the observation, and validate on the remaining 10%

Here the K groups should be created randomly...





k-validation croisée

- En pratique, on pose généralement $k = 10$ (*ten-fold cross validation*) ou $k = n$ (*leave-one-out cross validation*).
- Lorsque le modèle optimal est sélectionné, on cherche à estimer l'erreur quadratique de prédiction, c'est-à-dire
$$\mathbb{E}\Big(Y^* - \widehat{f}(\mathbf{X}^*)\Big)^2,$$
où $(Y^*; \mathbf{X}^*)$ est une nouvelle observation.
- Il faut alors faire attention car le tMSE généralement sous-estime cette valeur.
- Pour contourner ce problème, si la base de données est de taille suffisante, on peut garder une portion des données pour consituer une base de test.
- On estime alors l'erreur quadratique de prédiction comme étant
$$\widehat{\mathbb{E}\Big(Y^* - \widehat{f}(\mathbf{X}^*)\Big)^2} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(y_i^* - \widehat{f}(\mathbf{x}^*{}_i)\Big)^2.$$

Validation croisée

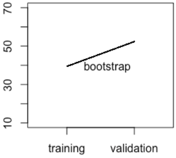
Bootstrap Cross Validation

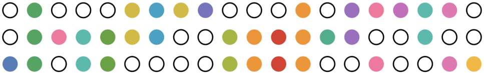
1. Generate B boostrap samples from $\{1, 2, \cdots, n\}$, $I_1, \cdots, I_B$
2. Estimate B models : $\widehat{m}_b$ on sample $(y_i, \boldsymbol{x}_i)$, $i \in I_b$
3. Compute $\displaystyle\frac{1}{B}\sum_{b=1}^{B}\frac{1}{n-|I_b|}\sum_{i\notin I_b}\ell\big(y_i, \widehat{m}_b(x_{1,i}, \cdots, X_{p,i})\big)$

The probability that $i \notin I_b$ is

$$\left(1-\frac{1}{n}\right)^n \sim e^{-1} (= 36.78\%)$$

At stage b, we validate on $\sim 36.78\%$ of the dataset.





Cas des séries chronologiques

Time Series

A time series is a sequence of observations (y_t) ordered in time.

Write $y_t = s_t + u_t$, with systematic part s_t (signal / trend) and 'residual' term u_t

(u_t) is supposed to be a strictly stationary time series

(s_t) might be a 'linear' trend, plus a seasonal cycle

Buys-Ballot (1847, Les changements périodiques de température, dépendants de la nature du soleil et de la lune, mis en rapport avec le pronostic du temps, déduits d'observations néerlandaises de 1729 à 1846) - original probably in Dutch.

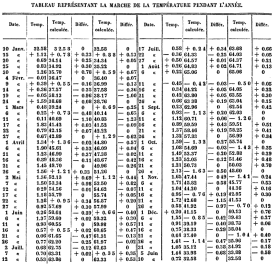


TABLEAU REPRÉSENTANT LA MARCHE DE LA TEMPÉRATURE PENDANT L'ANNÉE.

Date.	Temp.	Temp. calculée.	Differ.	Temp. calculée.	Differ.	Date.	Temp.	Temp. calculée.	Differ.	Temp. calculée.	Differ.
10 Janv.	32.58	32.58	0	32.58	0	17 Juill.	0.58	+ 0.24	+ 0.34	63.68	+ 0.66
15 «	+ 1.12	+ 0.78	+ 0.23	+ 0.88	+ 0.13	22 «	— 0.36	64.33	— 0.25	64.03	+ 0.05
20 «	0.69	34.14	+ 0.25	34.34	+ 0.05	27 «	+ 0.50	64.57	+ 0.01	64.37	+ 0.21
25 «	0.83	34.92	+ 0.30	35.22	0	1 Août	+ 0.26	64.83	+ 0.02	64.71	+ 0.13
30 «	1.26	35.70	+ 0.78	+ 0.59	+ 0.67	6 «	+ 0.22	65.06	0	65.06	0
4 Févr.	— 0.01	36.47	0	36.40	+ 0.07						
9 «	+ 0.39	+ 0.55	— 0.16	36.99	— 0.13	11 «	— 0.45	— 0.42	— 0.03	— 0.50	+ 0.05
14 «	+ 0.36	37.57	— 0.35	37.58	— 0.36	16 «	0.34	64.22	+ 0.05	64.05	+ 0.22
19 «	— 0.05	38.13	— 0.96	38.17	— 1.00	21 «	0.42	63.80	+ 0.05	63.55	+ 0.30
24 «	+ 1.59	38.68	+ 0.08	38.76	0	26 «	0.66	63.38	+ 0.19	63.04	+ 0.15
1 Mars	0.48	39.24	0	+ 0.69	— 0.25	1 Sept.	0.23	62.96	0	62.54	+ 0.42
6 «	0.25	+ 0.73	— 0.48	40.14	— 0.65	6 «	0.93	— 1.13	+ 0.20	62.03	0
11 «	0.11	40.69	— 1.10	40.83	— 1.23	11 «	1.12	60.71	+ 0.06	— 1.26	0
17 «	1.83	41.42	+ 0.03	41.52	— 0.09	16 «	0.89	59.59	+ 0.43	59.51	+ 0.51
22 «	0.79	42.15	+ 0.07	42.22	0	21 «	1.37	58.46	+ 0.19	58.25	+ 0.41
27 «	0.67	42.89	0	+ 1.29	— 0.62	26 «	1.32	57.33	0	56.99	+ 0.34
1 Avril	1.34	+ 1.36	— 0.02	44.80	— 0.57	1 Oct.	1.59	— 1.32	0.27	55.74	0
6 «	1.90	45.61	+ 0.52	46.09	+ 0.04	6 «	1.08	54.69	— 0.03	— 1.43	+ 0.35
11 «	1.23	46.99	+ 0.37	47.38	— 0.02	11 «	1.49	53.37	— 0.20	52.88	+ 0.29
16 «	0.89	48.36	— 0.11	48.67	— 0.42	16 «	1.23	52.05	— 0.12	51.46	+ 0.48
21 «	1.45	49.70	0	49.96	— 0.26	21 «	1.21	50.73	0	50.03	+ 0.70
26 «	1.56	+ 1.21	+ 0.33	51.26	0	26 «	2.13	— 1.63	— 0.50	48.60	0
2 Mai	1.56	52.13	+ 0.69	+ 1.12	+ 0.44	1 Nov.	1.65	47.44	— 0.49	— 1.41	— 0.24
7 «	1.50	53.34	+ 0.98	53.50	+ 0.82	6 «	1.71	45.82	— 0.58	45.77	— 0.53
12 «	0.29	54.56	— 0.01	54.62	— 0.07	11 «	1.04	44.20	0	44.36	+ 0.16
17 «	1.18	55.79	0	55.74	+ 0.05	16 «	0.95	— 0.76	+ 0.19	42.95	+ 0.30
22 «	1.28	+ 0.95	+ 0.34	56.87	+ 0.20	21 «	1.72	42.68	— 1.15	41.53	0
27 «	0.92	57.69	+ 0.30	57.99	0	26 «	0.58	41.92	— 0.97	— 0.70	+ 0.12
1 Juin	0.26	58.64	— 0.39	+ 0.66	— 0.40	1 Déc.	+ 0.20	41.15	0	40.13	+ 0.76
6 «	1.37	59.60	+ 0.02	59.32	+ 0.30	6 «	1.55	— 0.85	— 0.42	39.43	+ 0.27
11 «	0.93	60.55	0	59.98	+ 0.57	11 «	0.91	39.19	— 0.40	38.73	+ 0.06
16 «	0.57	+ 0.55	+ 0.02	60.65	+ 0.47	16 «	0.75	38.33	— 0.29	38.04	0
21 «	0.06	61.65	— 0.47	61.31	— 0.13	21 «	0.64	37.40	0	— 1.04	+ 0.40
26 «	0.77	62.20	— 0.25	61.97	— 0.02	26 «	1.61	— 1.14	— 0.47	35.96	— 0.17
2 Juill.	0.68	62.75	— 0.12	62.63	0	31 «	1.05	35.12	— 0.38	34.92	— 0.18
7 «	0.70	63.31	+ 0.02	+ 0.35	+ 0.35	5 Juin	1.44	33.98	— 0.68	33.88	— 0.58
12 «	0.53	63.86	0	63.33	+ 0.53	10 «	0.72	32.58	0	32.58	0

Cas des séries chronologiques: Lissage Exponentiel

Exponential smoothing - Simple

From time series (y_t) define a smooth version

$$s_t = \alpha \cdot y_t + (1-\alpha) \cdot s_{t-1} = s_{t-1} + \alpha \cdot (y_t - s_{t-1})$$

for some $\alpha \in (0,1)$ and starting point $s_0 = y_1$ Forecast is ${}_t\widehat{y}_{t+h} = s_t$

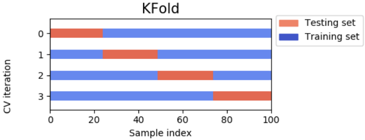
It is called exponential smoothing since

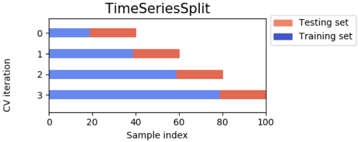
$$\begin{aligned} s_t &= \alpha y_t + (1-\alpha) s_{t-1} \\ &= \alpha y_t + \alpha(1-\alpha) y_{t-1} + (1-\alpha)^2 s_{t-2} \\ &= \alpha \left[y_t + (1-\alpha) y_{t-1} + (1-\alpha)^2 y_{t-2} + \cdots + (1-\alpha)^{t-1} y_1 \right] + (1-\alpha) \end{aligned}$$

corresponding to exponentially weighted moving average

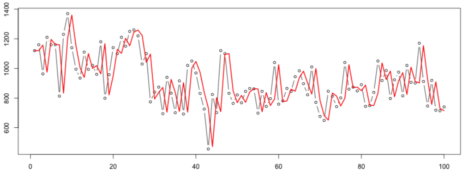
Need to adapt cross-validation techniques,

Cas des séries chronologiques: Lissage Exponentiel

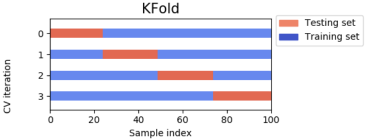


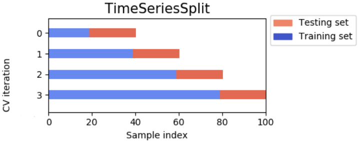


Optimal α ? $\alpha^\star \in \text{argmin}\left\{\sum_{t=2}^{T}\ell_2(y_t - {}_{t-1}\widehat{y}_t)\right\}$ (leave-one-out strategy)

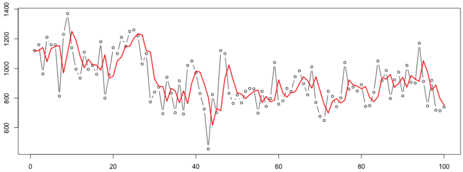


Cas des séries chronologiques: Lissage Exponentiel

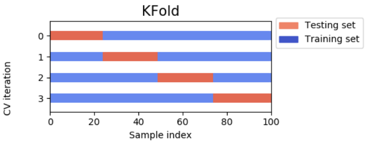


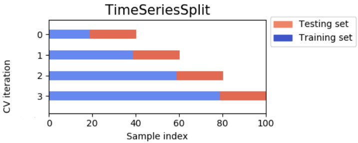


Optimal α ? $\alpha^\star \in \text{argmin}\left\{\sum_{t=2}^{T} \ell_2(y_t - {}_{t-1}\widehat{y}_t)\right\}$ (leave-one-out strategy)

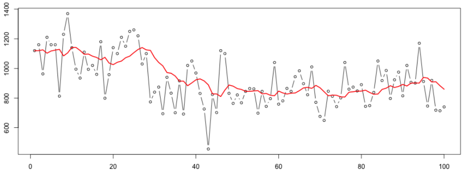


Cas des séries chronologiques: Lissage Exponentiel





Optimal α ? $\alpha^\star \in \text{argmin}\left\{\sum_{t=2}^{T}\ell_2(y_t - {}_{t-1}\widehat{y}_t)\right\}$ (leave-one-out strategy)



Cas des séries chronologiques: Lissage Exponentiel

See Hyndman *et al.* (2008, Forecasting with Exponential Smoothing)

Exponential smoothing - Double

From time series (y_t) define a smooth version

$$\begin{cases} s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1})\, s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1}) \\ b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1}\, b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1} \end{cases}$$

for some $\alpha \in (0,1)$, some trend $\beta \in (0,1)$ and starting points s_0 and b_0, $s_0 = y_0$ and $b_0 = y_1 - y_0$. Forecast is ${}_t\widehat{y}_{t+h} = s_t + h \cdot b_t$.

Cas des séries chronologiques: Lissage Exponentiel

See Hyndman *et al.* (2008, Forecasting with Exponential Smoothing)

Exponential smoothing - Double

From time series (y_t) define a smooth version

$$\begin{cases} s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1})\, s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1}) \\ b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1}\, b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1} \end{cases}$$

for some $\alpha \in (0,1)$, some trend $\beta \in (0,1)$ and starting points s_0 and b_0, $s_0 = y_0$ and $b_0 = y_1 - y_0$. Forecast is ${}_t\widehat{y}_{t+h} = s_t + h \cdot b_t$.

Cas des séries chronologiques: Lissage Exponentiel

See Hyndman *et al.* (2008, Forecasting with Exponential Smoothing)

Exponential smoothing - Double

From time series (y_t) define a smooth version

$$\begin{cases} s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1})\, s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1}) \\ b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1}\, b_t = \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1} \end{cases}$$

for some $\alpha \in (0,1)$, some trend $\beta \in (0,1)$ and starting points s_0 and b_0, $s_0 = y_0$ and $b_0 = y_1 - y_0$. Forecast is ${}_t\widehat{y}_{t+h} = s_t + h \cdot b_t$.

Cas des séries chronologiques: Lissage Exponentiel

Exponential smoothing - Seasonal with lag L (Holt-Winters)

From time series (y_t) define a smooth version

$$\begin{cases} s_t = \alpha \dfrac{x_t}{c_{t-L}} + (1-\alpha)(s_{t-1}+b_{t-1}) \\ b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1} \\ c_t = \gamma \dfrac{y_t}{s_t} + (1-\gamma)c_{t-L} \end{cases}$$

for some $\alpha \in (0,1)$, some trend $\beta \in (0,1)$, some seasonal change smoothing factor, $\gamma \in (0,1)$ and starting points $s_0 = y_0$. Forecast is ${}_t\widehat{y}_{t+h} = (s_t + hb_t)c_{t-L+1+(h-1) \mod L}$.

```r
stats::HoltWinters()
```