

- Objectifs
 - validation (ajuster les paramètres)
 - test (évaluer la performance, comparer des algorithmes)

- Estimées basées sur des échantillons
 - erreur de test (estimée de holdout, validation simple)
 - validation croisée (deleted estimate, rotation estimate, U-method, jack-knife, leave-one-out)
 - bootstrap
- Estimées basées sur des pénalités
 - AIC, BIC, MDL

- Estimées basées sur des pénalités

- idée: $R(f_n) \simeq \widehat{R}_{em}(f_n) + P(\alpha)$

- la complexité est mesuré par le nombre des paramètres N : $P(\alpha) = P(N)$

- AIC (Akaike information criterion): $P(N) = \Theta(N/n)$
- BIC (Bayesian information criterion): $P(N) = \Theta(N \log n/n)$
- MDS (minimum description length) \equiv BIC, motivé par la théorie de codage optimal

- Validation simple

- échantillon de test: $D_m = \{Z_{n+1}, \dots, Z_{n+m}\}$
- idée: $R(f_n) \simeq \widehat{R}_m(f_n) = \frac{1}{m} \sum_{i=n+1}^{n+m} L(f_n, Z_{n+i})$
- non-biaisé: $E[\widehat{R}_m(f_n) | D_n] = R(f_n)$

- Validation simple avec la classification:

- la perte: $L(Z, f) = L((X, Y), f) = \begin{cases} 0 & \text{si } f(X) = Y, \\ 1 & \text{si } f(X) \neq Y \end{cases}$
- $m\widehat{R}_m(f_n) | D_n \sim \text{BIN}(m, R(f_n))$
- $\text{Var}(\widehat{R}_m(f_n) | D_n) = E[(\widehat{R}_m(f_n) - R(f_n))^2 | D_n] = \frac{R(f_n)(1 - R(f_n))}{m} \leq \frac{1}{4m}$
- inégalité de Hoeffding $\rightarrow \forall \varepsilon > 0 : P\{|\widehat{R}_m(f_n) - R(f_n)| > \varepsilon | D_n\} \leq 2e^{-2m\varepsilon^2}$

- Validation croisée en k blocs

- soit $q = n/k$ (supposé entier)
- j ème ensemble de test: $D_q^{(j)} = \{Z_{(j-1)q+1}, Z_{(j-1)q+2}, \dots, Z_{jq}\}$
- j ème ensemble d'entraînement: $D_{n-q}^{(j)} = D_n \setminus D_q^{(j)}$
- j ème fonction entraînée sur $D_{n-q}^{(j)} : f_{n-q}^{(j)}$
- j ème erreur de test: $\widehat{R}_q^{(j)}(f_{n-q}^{(j)}) = \frac{1}{q} \sum_{i=(j-1)q+1}^{jq} L(f_{n-q}^{(j)}, Z_i)$
- idée: $R(f_n) \simeq \widehat{R}_{cv(k)}(f_n) = \frac{1}{k} \sum_{j=1}^k \widehat{R}_q^{(j)}(f_{n-q}^{(j)})$

- Validation **croisée** en k blocs

- **presque non-biaisé**: $E[\widehat{R}_{cr(k)}(f_n)|D_n] = R(f_{n-q})$

- $k = n$: **jackknife, leave-one-out**

- classification (Rogers et Wagner [1978]):

$$E\{(\widehat{R}_{cr(k)}(f_n) - R(f_n))^2\} \leq \frac{1}{n} + 6P\{f_n(X) \neq f_{n-1}(X)\}$$

- exemple: k -PPV

$$E\{(\widehat{R}_{cr(k)}(f_n) - R(f_n))^2\} \leq \frac{6k+1}{n}$$

- **Tests** statistiques pour **comparer** les algorithmes A et B

- Diettrich[98]: "Approximate Statistical Tests for Comparing ..."
- Hypothèse nulle: $R_n(A) = R_n(B)$
- t -test de **McNemar**
- test de **différence des erreurs**
- t -test des **paires re-échantillonnées**
- t -test des **paires de validation croisée**
- t -test des **paires 5x2cv**

- Sources de la variance

- $\widehat{R} = \widehat{R}_{D_m}(f_n) = \widehat{R}_{D_m}(\text{ALGO}(D_n))$
- $\text{Var}(\widehat{R}|D_n, f_n)$: variance de l'échantillon de **test**
- $\text{Var}(\widehat{R}|D_m, \text{ALGO}(\cdot))$: variance de l'échantillon d'**entraînement**
- $\text{Var}(\widehat{R}|D_m, D_n)$: variance de l'**algorithme**
- (R^*) : erreur de Bayes, **bruit**