Apprentissage automatique / Statistique

Qualité de Prévision et Risque

PHILIPPE BESSE

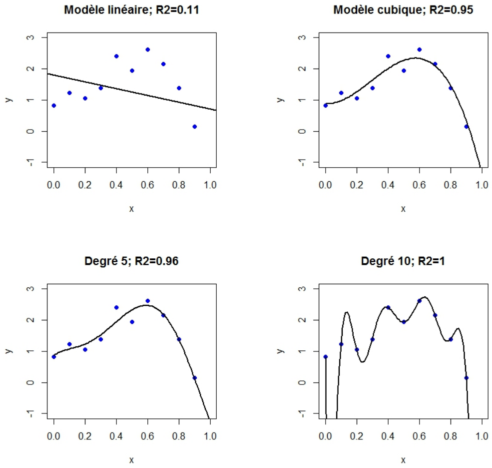
INSA de Toulouse
Institut de Mathématiques

Objectif

- Mesurer la performance d'un modèle, sa capacité de prévision ou de généralisation
  - Optimiser la sélection au sein d'une famille de modèles
  - choix de la méthode en comparant chacun des modèles
  - estimer la confiance accordée à une prévision
  - Enjeu : estimation sans biais de l'erreur de prévision
  - Capacité de généralisation du modèle ou algorithme

Sans modèle probabiliste, trois stratégies :

1. Pénalisation de l'erreur d'ajustement ou risque empirique
2. Partition de l'échantillon : apprentissage, (validation), test
3. Simulation : validation croisée, bootstrap



Régression polynomiale de degré 1, 2, 5 et 10

Notations

- $\boldsymbol{D}_n$ observations d'un n-échantillon $\boldsymbol{D}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ de loi conjointe inconnue P sur $\mathcal{X} \times Y$
- $\boldsymbol{x}$ observation de la variable $\boldsymbol{X}$ multidimensionnelle
- $\boldsymbol{D}_n$ est appelé échantillon d'apprentissage
- $\boldsymbol{D}_n$ est supposé indépendant de $(\boldsymbol{X}, Y)$
- Une règle de prévision (ou prédicteur) est une fonction (mesurable) $f : \mathcal{X} \to \mathcal{Y}$, $\boldsymbol{x} \to f(\boldsymbol{x})$
- Une fonction $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ est une fonction de perte si $l(y, y) = 0$ et $l(y, y') > 0$ pour $y \neq y'$
- Si f est une règle de prévision, $l(y, f(\boldsymbol{x}))$ mesure la perte de f en $\boldsymbol{x}$

Définitions

- Régression réelle : pertes $\mathbb{L}^p$ $(p \geq 1)$: $l(y, y') = |y - y'|^p$
  perte absolue si $p = 1$, perte quadratique si $p = 2$
- Discrimination binaire : $l(y, y') = \mathbb{1}_{y \neq y'} = \frac{|y-y'|}{2} = \frac{(y-y')^2}{4}$
- Risque d'une règle f : $R_P(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim P}[l(Y, f(\boldsymbol{X}))]$
- Risque empirique associé à $\boldsymbol{D}_n$:
  $\widehat{R_n}(f, \boldsymbol{D}_n) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(\boldsymbol{X}_i))$
- Minimisation du risque empirique sur un sous-ensemble F (un modèle) de $\mathcal{F}$: $\hat{f}_F(\boldsymbol{D}_n) \in \text{argmin}_{f\in F}\widehat{R_n}(f, \boldsymbol{D}_n)$
- Problème : choix de F !
- La règle oracle est telle que : $R_P(f^*) = \inf_f R_P(f)$

Décomposition du risque empirique

$$R_P(\hat{f}_F(\mathbf{D}_n)) - R_P(f^*) =$$

$$\underbrace{\left\{ R_P(\hat{f}_F(\mathbf{D}_n)) - \inf_{f \in F} R_P(f) \right\}} + \underbrace{\left\{ \inf_{f \in F} R_P(f) - R_P(f^*) \right\}}$$

Erreur d'estimation et d'approximation

(Variance) (Biais)

$\nearrow$ $\searrow$ (taille de F)

- Plus le modèle F est complexe ou flexible,
- plus le biais est réduit mais
- plus la partie variance risque d'augmenter
- Enjeu : meilleur compromis biais / variance

Risque empirique ou qualité d'ajustement

$$\widehat{R_n}(\widehat{f}(\boldsymbol{D}_n), \boldsymbol{D}_n) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \widehat{f}(\boldsymbol{D}_n)(\boldsymbol{x}_i))$$

- Minimum des moindres carrés dans le cas quantitatif
- Taux de mal classés dans le cas qualitatif
- Estimation biaisée, par optimisme

Estimation sans biais sur un échantillon indépendant

- Partition : $\boldsymbol{D}_n = \boldsymbol{D}_{n_1}^{\text{Appr}} \cup \boldsymbol{D}_{n_2}^{\text{Valid}} \cup \boldsymbol{D}_{n_3}^{\text{Test}}$
- $\widehat{R_n}(\widehat{f}(\boldsymbol{D}_{n_1}^{\text{Appr}}), \boldsymbol{D}_{n_1}^{\text{Appr}})$ pour estimer un modèle choisi $\widehat{f}(\boldsymbol{D}_{n_1}^{\text{Appr}})$
- $\widehat{R_n}(\widehat{f}(\boldsymbol{D}_{n_1}^{\text{Appr}}), \boldsymbol{D}_{n_2}^{\text{Valid}})$ pour optimiser un modèle
- $\widehat{R_n}(\widehat{f}, \boldsymbol{D}_{n_3}^{\text{Test}})$ pour comparer les meilleurs modèles

C_p de Mallows

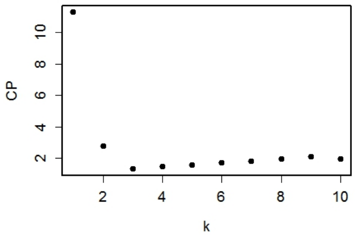
- Décomposition de l'erreur de prévision ou risque quadratique :

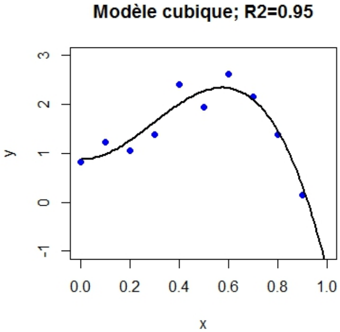
$$\widehat{R_P}(\widehat{f}(\boldsymbol{d}^n)) = \widehat{R_n}(\widehat{f}(\boldsymbol{d}^n), \boldsymbol{d}^n) + \text{Optim}$$

- Estimation normalisée :

$$C_p = \widehat{R_n}(\widehat{f}(\boldsymbol{d}^n), \boldsymbol{d}^n) + 2\frac{d}{n}\widehat{\sigma}^2$$

  - d : nombre de paramètres du modèle
  - n : nombre d'observations
  - s^2 : estimation de la variance de l'erreur par modèle de faible biais





Régression polynomiale : C_p de Mallows fonction du degré et modèle sélectionné.

Critère d'Akaïke

- Basé sur la dissemblance de Kullback
- compare la loi de Y et celle de $\widehat{Y}$
  - Suppose que la famille de lois du modèle contient la "vraie" loi de Y
  - Pour tout modèle estimé par minimisation d'une log-vraisemblance $\mathcal{L}$

$$\text{AIC} = -2\mathcal{L} + 2\frac{d}{n}$$

  - Cas gaussien et variance connue : AIC et C_p équivalents
- AIC_c adapté aux petits échantillons gaussiens

$$\text{AIC}_c = -2\mathcal{L} + \frac{n+d}{n-d-2}$$

Critère BIC de Schwarz

- BIC (Bayesian information criterion)
  - modèle de plus grande probabilité *a posteriori*

$$\text{BIC} = -2\mathcal{L} + \log(n)\frac{d}{n}$$

  - Cas gaussien et variance connue : BIC proportionnel à AIC
  - $n > e^2 \approx 7,4$, BIC pénalise plus les modèles complexes
  - Asymptotiquement, la probabilité pour BIC de choisir le bon modèle tend vers 1
  - différent d'AIC qui tend à choisir des modèles trop complexes
  - À Taille fini, BIC risque de se limiter à des modèles trop simples

Algorithme de V-*fold cross validation*

- *Découper* aléatoirement l'échantillon en V segments (V-fold) de tailles similaires selon une loi uniforme ;
- **for** k=1 à V
  - Mettre de côté le segment k,
  - Estimer le modèle sur les $V-1$ segments restants,
  - Calculer la moyenne des erreurs sur le segment k
- **end for**
- Moyenner toutes les erreurs

Utilisation

- Soit $\tau : \{1, \dots, n\} \mapsto \{1, \dots, V\}$ la fonction d'indexation
- $\widehat{f}^{(-v)}$ estimation de f sans le $v$ième segment de l'échantillon
- Estimation par validation croisée de l'erreur de prévision :

$$\widehat{R_{\text{CV}}} = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \widehat{f}^{(-\tau(i))}(\boldsymbol{x}_i))$$

- Choix de V : n (variance), petit (biais), 10 par défaut
- Utilisation fréquente en choix de modèle :

$$\widehat{\theta} = \arg\min_{\theta} \widehat{R_{\text{CV}}}(\theta)$$

Introduction au Bootstrap

- Simulation (Monte Carlo) de la distribution d'un estimateur
- Principe : substituer P_n, à la distribution inconnue P
- Tirage avec remise d'un échantillon bootstrap de même taille
- Itération et convergence

Estimateur bootstrap naïf

- Échantillon bootstrap : $\mathbf{z}^*$
- Estimateur plug-in (remplacer F par $\widehat{F}$ de $R_P(\hat{f}(\boldsymbol{d}_n))$:
  $\widehat{R}_n(\hat{f}_{z^*}, \boldsymbol{d}_n) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \hat{f}_{z^*}(\boldsymbol{x}_i))$
- $\hat{f}_{z^*}$ désigne l'estimation de f à partir de $\mathbf{z}^*$
- Estimation bootstrap de l'erreur moyenne de prévision $\mathbb{E}_{\boldsymbol{D}_n \sim P^{\otimes n}}[R_P(\hat{f}(\boldsymbol{D}_n)]$:
- $R_{\text{Boot}} = E_{\boldsymbol{Z}^* \sim \widehat{F}}[\widehat{R}_n(\hat{f}_{\boldsymbol{Z}^*}, \boldsymbol{D}_n)] = E_{\boldsymbol{Z}^* \sim \widehat{F}}\left[\frac{1}{n}\sum_{i=1}^{n} l(y_i, f_{\boldsymbol{Z}^*}(\boldsymbol{x}_i))\right]$

Estimation bootstrap par simulation

$$\widehat{R_{\text{Boot}}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} l(y_i, f_{z^{*b}}(\boldsymbol{x}_i))$$

Estimation biaisée par optimisme

Estimateur bootstrap out-of-bag

Distinguer les observations de l'échantillon bootstrap et les autres

$$\widehat{R_{\text{oob}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B_i}\sum_{b\in K_i} l(y_i, f_{z^{*b}}(\boldsymbol{x}_i))$$

- K_i est l'ensemble des indices b des échantillons bootstrap ne contenant pas la $i$ème observation à l'issue des B simulations
- $B_i = |K_i|$ est le nombre de ces échantillons
- $\widehat{R_{\text{oob}}}$ résout le problème d'un biais optimiste de $\widehat{R_{\text{Boot}}}$ mais biais pessimiste comme en validation croisée $(\widehat{R_{\text{CV}}})$

Estimateur .632-bootstrap

- Correctif basé sur la
- probabilité qu'une observation soit tirée dans un échantillon bootstrap :

$$P[\mathbf{x}_i \in \mathbf{x}^{*b}] = 1 - (1 - \frac{1}{n})^n \approx 1 - \frac{1}{e} \approx 0,632$$

- Sur-évaluation de l'erreur analogue à celle de la validation croisée avec $K = 2$
- Compensation :

$$\widehat{R}_{.632} = 0,368 \times \widehat{R}_n(\hat{f}(\boldsymbol{D}_n), \boldsymbol{D}_n) + 0,632 \times \widehat{R}_{\text{oob}}$$

Remarques

- Estimations de l'erreur asymptotiquement équivalentes
- Pas de choix *a priori*
- Bootstrap plus compliqué et encore peu utilisé mais
- Central dans les algorithmes de combinaison de modèles
- Problèmes du .632-bootstrap en sur-ajustement
- Rectificatif complémentaire : le *.632+bootstrap*
- Utiliser le même estimateur pour comparer deux méthodes

Matrice de confusion - Notations

Prévision : Si $\widehat{\pi}_i > s, \ \widehat{y}_i = 1$ sinon $\widehat{y}_i = 0$

Prévision	Observation		Total
	$Y = 1$	$Y = 0$	
$\widehat{y}_i = 1$	$n_{11}(s)$	$n_{10}(s)$	$n_{1+}(s)$
$\widehat{y}_i = 0$	$n_{01}(s)$	$n_{00}(s)$	$n_{0+}(s)$
Total	$n_{+}1$	$n_{+}0$	n

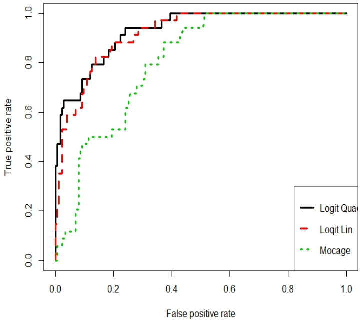
- Nombre de conditions positives $P = n_{+1}$
- Nombre de conditions négatives $N = n_{0+}$
- Vrais positifs $TP = n_{11}(s)$ bien classés ($\widehat{y}_i = 1$ et $Y = 1$),
- Vrais négatifs $TN = n_{00}(s)$ bien classés ($\widehat{y}_i = 0$ et $Y = 0$),
- Faux négatifs $FN = n_{01}(s)$ mal classés ($\widehat{y}_i = 0$ et $Y = 1$),
- Faux positifs $FP = n_{10}(s)$ mal classés ($\widehat{y}_i = 1$ et $Y = 0$),

Notations

- *Accuracy* et Taux d'erreur : $ACC = \frac{TN+TP}{N+P} = 1 - \frac{FN+FP}{N+P}$
- Taux de vrais positifs, *sensitivity, recall* $TPR = \frac{TP}{P}$
- Taux de vrais négatifs *spécificity, selectivity* $TNR = \frac{TN}{N}$
- Précision ou *positive predictive value*
  $PPV = \frac{TP}{TP+FP} = 1 - FDR$
- Taux de faux positifs $FPR = \frac{FP}{N} = 1 - TNR$
- Taux de faux négatifs $FNR = \frac{FN}{P} = 1 - TPR$
- Taux de fausses découvertes $FDR = \frac{FP}{FN+TN}$
- F_1 score ou moyenne harmonique de la précision et de la sensibilité, $F_1 = 2 \times \frac{PPV \times TPR}{PPV+TPR} = \frac{2 \times TP}{2 \times TP+FP+FN}$,
- $F_\beta(\beta \in \mathbb{R}^+)$ score, $F_\beta = (1+\beta^2)\frac{PPV \times TPR}{\beta^2 PPV+TPR}$.

- Taux de vrais positifs : TPR ou *sensibilité*
- Taux de faux positifs : $= FPR = 1 -$ Spécificité
- AUC : aire sous la courbe
- Score de Pierce : $H - F = TPR - FPN$
- Log loss

$$\text{Ll} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{L} y_{ik}\log(\widehat{\pi_{ik}})$$



Banque : Courbes ROC et aire sous la courbe