Université Paul Sabatier Toulouse III
Master 2 Professionnel IMAT
Méthodes d'apprentissage et data-mining
Année 2011-2012

T.P. 1 — Régression polynomiale

Le but de ce T.P. est d'illustrer la question du compromis biais/variance et le choix d'un bon modèle parmi une collection d'estimateurs. Il s'agit d'un problème que l'on retrouve dans toutes les approches nonparamétrique. Pour cela on va considérer un problème de régression polynomiale et s'intéresser à la question de l'estimation d'un degré optimal.

1 Principe de la régression polynomiale

On suppose que l'on dispose de l'observation de n variables aléatoires réelles $Y_i, i = 1, \ldots, n$ indépendantes, et l'on se place dans le cadre du modèle de régression standard :

$$\begin{aligned} Y_i &= f(x_i) + \epsilon_i,\ i = 1, \ldots, n, \\ \epsilon_1, \ldots, \epsilon_n &\text{ i.i.d. avec } \mathbb{E}(\epsilon_i) = 0,\ \mathrm{Var}(\epsilon_i) = \sigma^2, \end{aligned} \tag{1}$$

où les prédicteurs x_i sont **déterministes** et à valeurs dans $\mathbb{R}$ et les ϵ_i sont des variables Gaussiennes. On se place dans le cas de la régression polynomiale, et on suppose que f est de fa forme

$$f(x) = \theta_0 + \theta_1 x + \ldots + \theta_{m^*} x^{m^*},$$

pour tout $x \in \mathbb{R}$, où m^* est un entier **inconnu**, et

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{m^*}) \in \mathbb{R}^{m^*+1}$$

un ensemble de paramètres à estimer. Le problème est alors d'estimer f par un polynôme en choisissant le "bon" degré m parmi un ensemble d'entiers possible $\{1, \ldots, M\}$ où M est le degré maximal qu'on s'autorise.

Dans ce qui suit, on désignera par $\boldsymbol{Y}$ le vecteur colonne de $\mathbb{R}^n$ qui contient les observations i.e. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$. De plus, pour un degré $m \in \{1, \ldots, M\}$, on notera par $\boldsymbol{X}_m$ la matrice de taille $(n \times m + 1)$ qui s'écrit sous la forme :

$$\boldsymbol{X}_m = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^m \\ 1 & x_2^1 & \ldots & x_2^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \ldots & x_n^m \end{pmatrix}$$

Le modèle de régression polynomiale (1) pour le degré m^* peut alors s'écrire sous la forme d'un modèle linéaire :

$$\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\epsilon}^* = \boldsymbol{X}_{m^*} \cdot \boldsymbol{\theta}_{m^*} + \boldsymbol{\epsilon}^*,\ \text{avec } \boldsymbol{\epsilon}^* \sim N(0, \sigma^2 I_n),$$

où $\boldsymbol{f}^* = \boldsymbol{X}_{m^*} \cdot \boldsymbol{\theta}_{m^*}$ et $\boldsymbol{\theta}_{m^*}$ est un vecteur dont toutes les coordonnées sont non nulles. Donc la vraie fonction f est exactement polynôme de degré m^* ! En pratique on ne connait

pas le degré m^* donc on va estimer f par différents modèles polynomiaux de degré $m \in \{1, \dots, M\}$.

Soit $m \in \{1, \dots, M\}$ un degré donné et P_m la matrice de projection orthogonale sur le sous-espace vectoriel V_m de $\mathbb{R}^n$ engendré par les colonnes de $\boldsymbol{X}_m$. On suppose que les points x_i sont choisis de sorte que V_m est de dimension $m+1$. Et donc dans ce cas :

$$P_m = \boldsymbol{X}_m(\boldsymbol{X}_m^t \boldsymbol{X}_m)^{-1}\boldsymbol{X}_m^t$$

On notera :

- $\hat{\boldsymbol{f}}_m = P_m \boldsymbol{Y}$ l'estimation par moindres carrés de $\boldsymbol{Y}$ dans le modèle m,
- $\boldsymbol{f}_m^* = P_m \boldsymbol{f}^*$

On définit également :
- le risque de l'estimateur $\hat{\boldsymbol{f}}_m$ par $R(m) = \mathbb{E}\|\hat{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2$
- l'erreur empirique des moindres carrées par $R_n(m) = \|\boldsymbol{Y} - \hat{\boldsymbol{f}}_m\|^2$

Idéalement, on voudrait choisir le degré m qui donne la plus petite erreur $R(m)$ i.e.

$$m_{ideal} = \arg \min_{1 \le m \le M} R(m),$$

mais ce calcul n'est pas possible en pratique car on ne connait pas $\boldsymbol{f}^*$ et donc on ne peut pas calculer $R(m)$. L'idée la plus simple qui vient ensuite est de chercher le degré $\hat{m}$ qui minimise $R_n(m)$ i.e.

$$\hat{m}_{naif} = \text{argmin}_{1 \le m \le M} R_n(m)$$

mais nous allons voir que c'est une **très mauvaise méthode** car ceci conduit toujours au choix du degré le plus élevé, car $R_n(m)$ n'est pas une bonne approximation de $R(m)$.

On définit également la variance $\text{Var}(\hat{\boldsymbol{f}}_m)$ et le biais (au carré) $\text{Biais}^2(\hat{\boldsymbol{f}}_m)$ de l'estimateur $\hat{\boldsymbol{f}}_m$ par

$$\text{Var}(\hat{\boldsymbol{f}}_m) = \mathbb{E}\|\hat{\boldsymbol{f}}_m - \mathbb{E}\hat{\boldsymbol{f}}_m\|^2 \text{ et } \text{Biais}^2(\hat{\boldsymbol{f}}_m) = \|\mathbb{E}\hat{\boldsymbol{f}}_m - \boldsymbol{f}^*\|^2$$

et on peut montrer que

$$R(m) = \text{Biais}^2(\hat{\boldsymbol{f}}_m) + \text{Var}(\hat{\boldsymbol{f}}_m).$$

De plus dans le cas d'un modèle Gaussien, on a les formules explicites suivantes :

$$\text{Var}(\hat{\boldsymbol{f}}_m) = \sigma^2(m+1) \text{ et } \text{Biais}^2(\hat{\boldsymbol{f}}_m) = \|P_m \boldsymbol{f}^* - \boldsymbol{f}^*\|^2$$

Une méthode possible pour estimer un bon modèle est d'utiliser le critère pénalisé suivant (appelé C_p de Mallows)

$$\hat{m} = \text{argmin}_{1 \le m \le M} C_n(m)$$

où

$$C_n(m) = R_n(m) + 2\sigma^2(m+1).$$

2 Simulation des données

Q. 1 *Utiliser le code ci-dessous pour simuler des données qui correpondent à* $n = 30$ *observations bruitées d'un polynôme de degré* $m^* = 5$ *pour des points* $x_i = \frac{i}{n},\ i = 1, \ldots, n$. *Visualiser graphiquement la vraie fonction de régression et les observations.*

```
# Vraie fonction
theta <- c(1,-1,2,-0.8,0.6,-1)
T <- 100
t <- (1:T)/T
XT <- cbind(rep(1,T),t,t^2,t^3,t^4,t^5)
f <- XT%*%theta

# Visualisation de la fonction de regression
x11()
plot(t,f,type = "l", col="red", lwd=2)


# Choix des points du design : n valeurs regulierement espacees sur [0,1]
n <- 30
x <- (1:n)/n
X <- cbind(rep(1,n),x,x^2,x^3,x^4,x^5)

# Data 1
sigma <- 0.05 # Niveau de bruit
Y <- as.vector(X%*%theta + sigma*rnorm(n))

# Visualisation des donnees
x11()
plot(x,Y,type="p",pch=19)

# Visualisation des donnees avec la fonction a reconstruire
x11()
plot(x,Y,type="p",pch=19)
lines(t,f, type = "l", col="red", lty="dotdash", lwd=2)
```

3 Estimation

3.1 Estimation des paramètres et de la fonction de régression

La fonction `lm` permet d'ajuster un modèle de regression et la fonction `predict` d'obtenir les valeurs prédites par le modèle soit sur les points du design soit sur d'autres points. Par exemple pour ajuster un polynôme de degré 4 aux données, on peut utiliser les lignes de commandes suivantes :

```
# Estimation par moindres carrés
# Degre 4
```

```
donnees <- data.frame(y=Y,x1=x,x2=x^2,x3=x^3,x4=x^4)
mod4 <- lm(y~x1+x2+x3+x4,data=donnees)
pred4 <- data.frame(x1=t,x2=t^2,x3=t^3,x4=t^4)

# Visualisation des donnees
x11()
plot(x,Y,type="p",pch=19)
lines(t,predict(mod4,pred4), lwd=2, col='black')

# Superposition de la vraie fonction de regression
lines(t,f, type = "l", col="red", lty="dotdash", lwd=2)
```

Q. 2 *Utiliser le code ci-dessus pour ajuster aux données des polynômes de degré 1, 4 et 15. Comparer graphiquement les résultats obtenus. Par rapport aux données quel est selon vous le meilleur estimateur ? Par rapport à la vraie fonction de régression quel est selon vous le meilleur estimateur ?*

3.2 Choix du meilleur degré

Q. 3 *Pour les degrés* $m = 1, 2, 3, \ldots, 10$ *calculer* $R_n(m)$ *puis déterminer* $\hat{m}_{naif}$. *Comparer les valeurs de* $\hat{m}_{naif}$ *et* m^*.

Q. 4 *Pour les degrés* $m = 1, 2, 3, \ldots, 10$ *calculer* $C_n(m)$ *puis déterminer* $\hat{m}$. *Comparer les valeurs de* $\hat{m}$ *et* m^*.

Q. 5 *Reprendre les questions précédentes en simulant un nouveau jeu de données avec les mêmes points de design. Comment varient les valeurs de* $\hat{m}_{naif}$ *et* m^* *?*

3.3 Evolution du biais et de la variance en fonction du degré

Q. 6 *Créer* $S = 100$ *jeux de données différents* Y_i^s, $i = 1, \ldots, n$, $s = 1, \ldots, S$ *à partir du polynôme de degré 5 donné dans la question 1 en conservant les mêmes points de design pour chaque jeu de données.*

Q. 7 *Pour chaque jeu de données* Y^s, $s = 1, \ldots, S$, *calculer une estimation* $\hat{\boldsymbol{f}}_m^s$ *de la fonction de régression par un polynôme de degré* m *pour* m *variant de 1 à 10. Calculer alors les approximations numériques suivantes respectivement pour le biais au carré et la variance :*

$$\tilde{Var}(\hat{\boldsymbol{f}}_m) = \frac{1}{S}\sum_{s=1}^{S} \|\hat{\boldsymbol{f}}_m^s - \bar{\boldsymbol{f}}_{m,S}\|^2 \text{ et } \tilde{Biais}^2(\hat{\boldsymbol{f}}_m) = \|\bar{\boldsymbol{f}}_{m,S} - \boldsymbol{f}^*\|^2$$

où $\bar{\boldsymbol{f}}_{m,S} = \frac{1}{S}\sum_{s=1}^{S} \hat{\boldsymbol{f}}_m^s$. *Puis caculer une approximation numérique de* $R(m)$ *par*

$$\tilde{R}(m) = \tilde{Biais}^2(\hat{\boldsymbol{f}}_m) + \tilde{Var}(\hat{\boldsymbol{f}}_m).$$

Q. 8 *Tracer les courbes* $m \mapsto \tilde{Var}(\hat{\boldsymbol{f}}_m)$, $m \mapsto \tilde{Biais}^2(\hat{\boldsymbol{f}}_m)$ *et* $m \mapsto \tilde{R}(m)$. *Que constatez-vous quand* m *augmente ? Quelle est la valeur de* m *(entre 1 et 10) qui minimise* $\tilde{R}(m)$ *?*

Q. 9 *Pour chaque jeu de données, Y^s, $s = 1, \ldots, S$ quelle est la valeur de $\hat{m}^s_{naif}$ (entre 1 et 10) qui minimise l'erreur empirique des moindres carrées*

$$R^s_n(m) = \|\boldsymbol{Y}^s - \hat{\boldsymbol{f}}^s_m\|^2$$

et la valeur de $\hat{m}^s$ (entre 1 et 10) qui minimise le critère pénalisé

$$C^s_n(m) = \|\boldsymbol{Y}^s - \hat{\boldsymbol{f}}^s_m\|^2 + 2\sigma^2(m+1)$$

pour $s = 1, \ldots, S$? Comment varient $\hat{m}^s_{naif}$ et $\hat{m}^s$ pour $s = 1, \ldots, S$.