

Cours de Statistique : Théorie de la Régression

Fondamentaux, Non-paramétrique et Régularisation

Lecture 2

1 Introduction et Cadre Probabiliste

L'objectif de la régression est de prédire une variable aléatoire de sortie $Y \in \mathbb{R}$ à partir d'un vecteur de variables d'entrée (prédicteurs) $X \in \mathcal{X} \subset \mathbb{R}^d$.

Soit (X, Y) un couple de variables aléatoires suivant une loi de probabilité jointe inconnue, caractérisée par sa densité $f_{X,Y}(x, y)$. Nous disposons d'un échantillon de N observations indépendantes et identiquement distribuées (i.i.d.) :

$$\mathcal{D}_N = \{(x_n, y_n)\}_{n=1}^N$$

Nous cherchons une fonction de décision $f : \mathcal{X} \rightarrow \mathbb{R}$ telle que $f(X)$ soit une "bonne" approximation de Y .

2 L'approche Naïve et ses Limites

Une approche intuitive consiste à minimiser le risque empirique (erreur quadratique moyenne sur les données observées) :

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N |y_n - f(x_n)|^2 \quad (1)$$

Le problème du sur-apprentissage (Overfitting) Si l'espace de fonctions \mathcal{F} est trop vaste (par exemple, l'ensemble de toutes les fonctions continues), il existe une infinité de solutions annulant parfaitement l'erreur empirique.

- **Polynôme de Lagrange** : On peut construire un polynôme de degré $N - 1$ passant par tous les points (x_n, y_n) .
- **Conséquence** : Bien que l'erreur d'entraînement soit nulle, la capacité de généralisation sur de nouvelles données est médiocre. C'est le phénomène de sur-apprentissage.

3 Caractérisation de la Solution Optimale

Pour définir proprement la "meilleure" fonction, on se place dans le cadre théorique de la minimisation du risque quadratique attendu (L2).

Définition 3.1 (Fonction de régression). *La solution du problème de minimisation théorique :*

$$f^* = \operatorname{argmin}_{f \in L^2(P_X)} \mathbb{E}_{X,Y} [|Y - f(X)|^2]$$

est donnée par l'espérance conditionnelle :

$$m(x) = \mathbb{E}[Y|X = x] \quad (2)$$

Preuve (Approche Bayésienne) : En utilisant la loi des probabilités totales (désintégration de la mesure), on peut décomposer le risque :

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}_X [\mathbb{E}_Y[(Y - f(X))^2 | X = x]]$$

Pour chaque x , le minimum de $\mathbb{E}[(Y - c)^2 | X = x]$ par rapport à la constante c est atteint pour $c = \mathbb{E}[Y | X = x]$.

Modèle de bruit additif : On suppose souvent le modèle suivant :

$$Y = f(X) + \varepsilon, \quad \text{avec } \mathbb{E}[\varepsilon | X] = 0 \text{ et } \text{Var}(\varepsilon | X) = \sigma^2$$

Dans ce cas, la fonction cible est bien $f(x) = \mathbb{E}[Y | X = x]$.

4 Méthodes d'Estimation Non-Paramétriques

Puisque $f_{X,Y}$ est inconnue, nous devons estimer $m(x)$ à partir des données \mathcal{D}_N .

4.1 Approche Heuristique : k -plus proches voisins (k -NN)

L'idée est de moyenner les réponses y_i des observations dont les x_i sont les plus proches de x . Soit σ_x une permutation des indices telle que $\|x - x_{\sigma_x(1)}\| \leq \dots \leq \|x - x_{\sigma_x(N)}\|$.

- Si $k = 1$: $\hat{f}(x) = y_{\sigma_x(1)}$. On interpole les données (Risque de sur-apprentissage).
- Si $k = N$: $\hat{f}(x) = \frac{1}{N} \sum y_n = \bar{Y}$. Modèle constant (Risque de sous-apprentissage).

4.2 Lissage par Noyau : Estimateur de Nadaraya-Watson

On cherche à estimer $m(x) = \int y \frac{f_{X,Y}(x,y)}{f_X(x)} dy$. En remplaçant les densités par leurs estimateurs de noyau (Parzen-Rosenblatt) :

- $\hat{f}_X(x) = \frac{1}{N} \sum_{n=1}^N K_h(x - x_n)$
- $\hat{f}_{X,Y}(x, y) = \frac{1}{N} \sum_{n=1}^N K_h(x - x_n) K_h(y - y_n)$

L'estimateur de **Nadaraya-Watson** devient :

$$\hat{f}(x) = \sum_{n=1}^N w_n(x) y_n, \quad \text{où } w_n(x) = \frac{K_h(x - x_n)}{\sum_{i=1}^N K_h(x - x_i)} \quad (3)$$

Note : Les poids $w_n(x)$ somment à 1 et représentent l'influence relative du point n sur la prédiction en x .

5 Régularisation et Splines de Lissage

Pour éviter le sur-apprentissage tout en restant flexible, on restreint l'espace des solutions en ajoutant une pénalité de régularisation.

5.1 Principe de Projection et Pénalisation

On cherche f dans un sous-espace \mathcal{E} de L^2 ou on minimise :

$$\hat{f} = \underset{f}{\operatorname{argmin}} \sum_{n=1}^N |y_n - f(x_n)|^2 + \lambda \text{Pen}(f)$$

- **Régression Ridge** : $\text{Pen}(f) = \|f\|_{L^2}^2$ (favorise les petites normes).
- **Lasso** : $\text{Pen}(f) = \|f\|_{L^1}$ (favorise la parcimonie).

5.2 Splines de Lissage

On s'intéresse au problème de minimisation sur l'espace des fonctions deux fois dérivables sur $[a, b]$:

$$J(f) = \frac{1}{N} \sum_{n=1}^N (y_n - f(x_n))^2 + \lambda \int_a^b |f''(t)|^2 dt \quad (4)$$

Le terme $\int |f''(t)|^2 dt$ pénalise la courbure de la fonction (sa "rugosité").

Définition 5.1 (Spline Cubique). *Une fonction S est une spline cubique sur une partition $a = t_0 < t_1 < \dots < t_p = b$ si :*

1. *S est un polynôme de degré ≤ 3 sur chaque intervalle $[t_n, t_{n+1}]$.*
2. *S est de classe C^2 sur $[a, b]$.*

Résultat Fondamental : La solution du problème $J(f)$ est unique et est une **spline cubique naturelle** dont les nœuds sont situés aux points d'observation x_1, \dots, x_N . Bien que l'espace C^2 soit de dimension infinie, la solution appartient à un espace de dimension finie N , ce qui rend le calcul possible par des algorithmes d'algèbre linéaire.