Introduction au Machine Learning

Méthodes Linéaire

Théo Lopès-Quintas

BPCE Payment Services,
Université Paris Dauphine

2022 - 2025

Régression Linéaire

Régression Logistique

Comment prédire un nombre ?

Formalisation d'un problème de régression

On considère le problème de régression associé au dataset décrit par (1) : on cherche à prédire un nombre.

$$\mathcal{D} = \left\{ (x^{(i)}, y_i) \mid \forall i \leqslant \underbrace{n}_{\text{Nombre d'observations}},\ x^{(i)} \in \mathbb{R}^{\overbrace{d}^{\text{Nombre d'informations}}}, y_i \in \mathbb{R} \right\} \tag{1}$$

On suppose qu'il existe un lien entre la cible et les informations contenue dans $\mathcal{D}$. Ce lien est une fonction notée $f : \mathbb{R}^d \to \mathbb{R}$ de paramètre $\theta \in \mathbb{R}^{d'}$. On cherche donc à résoudre le problème d'optimisation :

$$\theta^* = \underset{\theta \in \mathbb{R}^{d'}}{\arg\min} \sum_{i=1}^{n} \left(\underbrace{y_i}_{\text{Vraie valeur}} - \overbrace{f_\theta\left(x^{(i)}\right)}^{\text{Valeur prédite}} \right)^2 \tag{2}$$

Comment prédire un nombre ?

Formalisation d'un problème de régression linéaire

Dans le cadre d'une régression **linéaire** on suppose que le lien entre la cible et les informations disponibles est **linéaire** :

$$\hat{y} = \theta_0 + \sum_{j=1}^{d} \theta_j x_j$$

On peut réécrire notre problème (2) comme :

$$\theta^* = \underset{\theta \in \mathbb{R}^{d+1}}{\arg\min} \sum_{i=1}^{n} \left[y_i - \left(\theta_0 + \sum_{j=1}^{d} \theta_j x_j^{(i)} \right) \right]^2$$

Comment prédire un nombre ?

Formalisation d'un problème de régression linéaire : visualisation

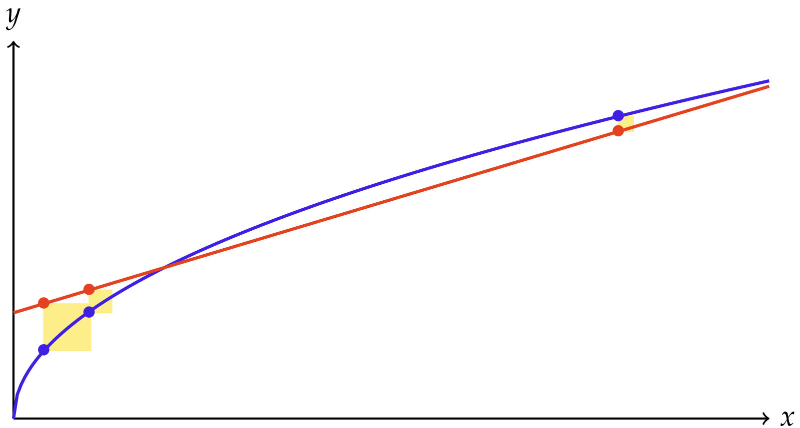


Figure – Visualisation de l'erreur entre la régression linéaire et la vraie courbe

Exercice 1 (Régression linéaire avec une seule information)

On suppose que l'on dispose d'un dataset $\mathcal{D} = \left\{(x^{(i)}, y_i) \mid \forall i \leqslant n : x^{(i)} \in \mathbb{R}, y_i \in \mathbb{R}\right\}$. *On a donc une seule information pour prédire la valeur* y.

1. *Écrire le problème (2) dans le cadre de l'exercice.*
2. *Donner le meilleur vecteur de paramètre* θ.

   On note $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$. *On rappelle avec cette convention que pour* $u, v \in \mathbb{R}^n$ *:*

$$\begin{aligned} \mathrm{Cov}(u, v) &= \overline{uv} - \overline{u} \times \overline{v} \\ \mathbb{V}\left[u\right] &= \overline{u^2} - \overline{u}^2 \end{aligned}$$

3. *Montrer que* θ_0^* *et* θ_1^* *les deux paramètres optimaux peuvent s'écrire :*

$$\begin{aligned} \theta_0^* &= \overline{y} + \theta_1^* \times \overline{x} \\ \theta_1^* &= \frac{\mathrm{Cov}(x, y)}{\mathbb{V}\left[x\right]} \end{aligned}$$

Comment prédire un nombre ?

Trouver les paramètres optimaux : en général

Puisqu'on suppose un lien linéaire, on peut exploiter l'écriture matricielle :

$$\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} & \cdots & x_d^{(n)} \end{pmatrix} \times \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\
&\iff \\
Y &= X\theta + \varepsilon \text{ , avec } \varepsilon \text{ un vecteur de bruit.}
\end{aligned}$$

On peut réécrire notre problème (2) comme $\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min} \|Y - X\theta\|^2$

Proposition 1

Si la matrice X est de rang plein, alors $\theta^* = ({}^tXX)^{-1}\,{}^tXY$

Comment mesurer la performance d'une régression ?

Erreur quadratique moyenne

Une première manière de mesurer la performance est de considérer l'erreur quadratique moyenne (MSE) :

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Cette métrique peut être décomposée en deux quantitées :

- $\text{Bias}\left[\hat{f}(x)\right] = \mathbb{E}\left[\hat{f}(x)\right] - f(x)$: l'écart moyen entre la valeur prédite et la vraie valeur
- $\mathbb{V}\left[\hat{f}(x)\right] = \mathbb{E}\left[\left(\mathbb{E}\left[\hat{f}(x)\right] - \hat{f}(x)\right)^2\right]$: la dispersion moyenne des valeurs prédites autour de la moyenne

$$\text{MSE}(y, \hat{f}(x)) = \left(\text{Bias}\left[\hat{f}(x)\right]\right)^2 + \mathbb{V}\left[\hat{f}(x)\right] + \underbrace{\boxed{\sigma^2}}_{\text{Erreur incompressible}} \tag{3}$$

COMMENT MESURER LA PERFORMANCE D'UNE RÉGRESSION ?

RMSE

Une deuxième manière de mesurer la performance est de calculer la racine carrée de la MSE :

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (y_i - \hat{y}_i)^2}$$

Exercice 2 (Ordre de grandeur)

Montrer que :

$$\mathrm{RMSE}(y, \overline{y}) = \overline{y^2} - \overline{y}^2$$

En déduire une interprétation de la RMSE et un critère de performance d'une régression.

Comment mesurer la performance d'une régression ?

Coefficient de détermination R^2

Une troisième manière de mesurer la performance est d'étudier le coefficient de détermination R^2 :

Prédiction du modèle

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \boxed{\hat{y}_i})^2}{\sum_{i=1}^{n} (y_i - \boxed{\overline{y}})^2}$$

Moyenne de la cible

Exercice 3 (Interprétation de R^2)

On suppose que l'on dispose des vecteurs y et $\hat{y}$.

1. *Comment interpréter la valeur 1 pour le R^2 ? Et la valeur 0 ?*
2. *Le R^2 peut-il être négatif ?*

Régressions pénalisées

Régression Ridge

On dit qu'on **pénalise** un modèle quand on modifie le problème d'optimisation :

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min} \overbrace{\boxed{\sum_{i=1}^{n} \left(y_i - f_\theta \left(x^{(i)} \right) \right)^2}}^{\text{Apprentissage}} + \underbrace{\boxed{\mathcal{P}_\lambda(\theta)}}_{\text{Pénalisation}} \qquad \text{(Pénalisation)}$$

On définit la régression Ridge comme une régression pénalisée :

$$\theta^*_{\text{Ridge}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \|Y - X\theta\|^2 + \lambda\|\theta\|^2 \qquad \text{(Ridge)}$$

Dans ce cas, on obtient l'estimateur optimal à l'aide de la formule :

$$\theta^*_{\text{Ridge}} = \left({}^t XX + n\lambda \mathbb{I}_d\right)^{-1} {}^t XY$$

On dit qu'on **pénalise** un modèle quand on modifie le problème d'optimisation :

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min} \overbrace{\boxed{\sum_{i=1}^{n} \left(y_i - f_\theta\left(x^{(i)}\right)\right)^2}}^{\textbf{Apprentissage}} + \underbrace{\boxed{\mathcal{P}_\lambda(\theta)}}_{\textbf{Pénalisation}} \quad \text{(Pénalisation)}$$

On définit la régression LASSO comme une régression pénalisée :

$$\theta^*_{\text{LASSO}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \|Y - X\theta\|^2 + \lambda \|\theta\|_1 \quad \text{(LASSO)}$$

Exercice 4 (Biais/Variance pour Ridge et LASSO)

Pour la régression Ridge, puis la régression LASSO, comment évolue le biais quand λ augmente ? Même question pour la variance.

Comment prédire une classe ?

Formalisation d'un problème de classification

On considère le problème de classification associé au dataset décrit par (4) : on cherche à prédire un nombre.

$$\mathcal{D} = \left\{ (x^{(i)}, y_i) \mid \forall i \leqslant \underbrace{n}_{\text{Nombre d'observations}},\ x^{(i)} \in \mathbb{R}^{\overbrace{d}^{\text{Nombre d'informations}}}, y_i \in \mathcal{Y} \subset \mathbb{N} \right\} \tag{4}$$

On suppose qu'il existe un lien entre la cible et les informations contenue dans $\mathcal{D}$. Ce lien est une fonction notée $f : \mathbb{R}^d \to \mathcal{Y}$ de paramètre $\theta \in \mathbb{R}^{d'}$. On cherche à résoudre le problème d'optimisation :

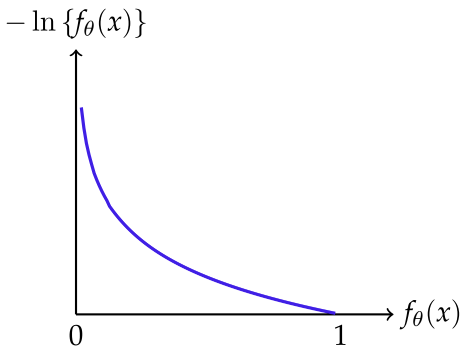
$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d'}} - \left[\overbrace{y \ln \{f_\theta(x)\}}^{\text{Observation positive}} + \underbrace{(1-y) \ln \{1 - f_\theta(x)\}}_{\text{Observation négative}} \right]$$

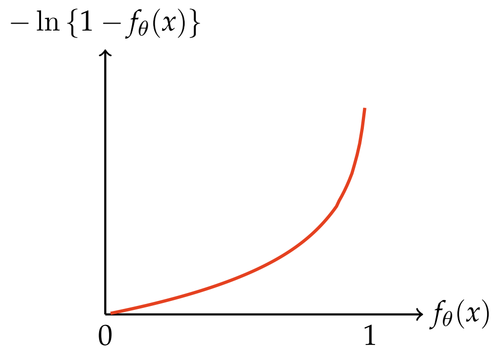
Comment prédire une classe ?

Formalisation d'un problème de classification : visualisation

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{d'}} - \left[\overbrace{y \ln\{f_\theta(x)\}}^{\text{Observation positive}} + \underbrace{(1-y)\ln\{1 - f_\theta(x)\}}_{\text{Observation négative}} \right]$$

Pour le problème d'optimisation considéré, on peut visualiser les variations de chaque partie comme :





Comment prédire une classe ?

Descente de gradient

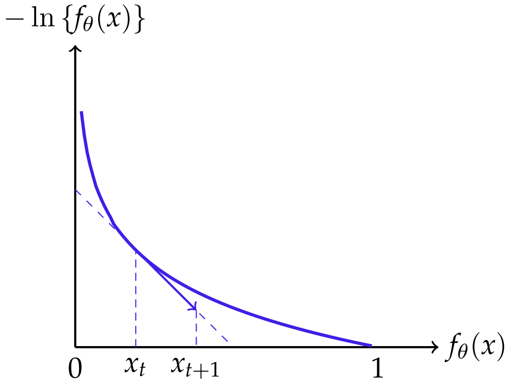
Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction différentiable. Pour le problème d'optimisation :

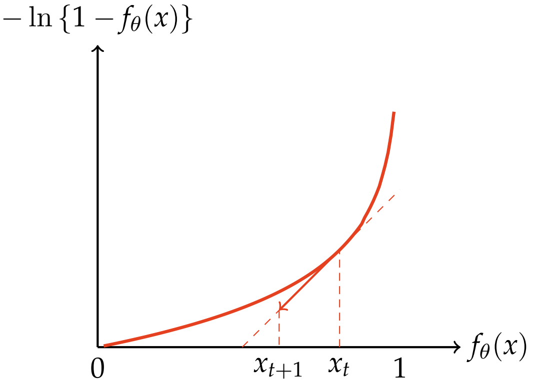
$$x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$$

La descente de gradient correspond au schéma :

$$x_{t+1} = x_t - \eta \nabla f(x_t) \quad \text{avec } \eta > 0$$

Visuellement pour notre problème cela se traduit par :





COMMENT PRÉDIRE UNE CLASSE ?

MODÉLISATION D'UN PROBLÈME DE CLASSIFICATION PAR LA RÉGRESSION LOGISTIQUE

La **régression logistique** suppose un lien *linéaire* entre les informations et la côte que l'observation soit de la classe d'intérêt. On modélise cela par la fonction f :

$$f_\theta(x) = \frac{1}{1+e^{-(x_1\theta_1 + \ldots + x_d\theta_d)}} = \frac{1}{1+e^{-\langle x, \boxed{\theta} \rangle}} \qquad \text{(Régression logistique)}$$

$\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$

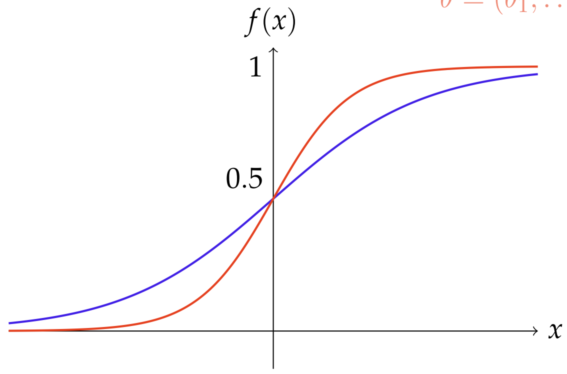


Figure – $f(x) = \frac{1}{1+e^{-x}}$ et $f(x) = \frac{1}{1+e^{-2x}}$

Exercice 5 (Équation de la descente de gradient)

On rappelle que $f_\theta(x) = \dfrac{1}{1+e^{-<\theta,x>}}$ *et on note* $\mathcal{L}(\theta; x, y) = -\left[y \ln\{f_\theta(x)\} + (1-y)\ln\{1 - f_\theta(x)\}\right]$.
Montrer que :

1. $f_\theta(x) = \dfrac{e^{<\theta,x>}}{1+e^{<\theta,x>}}$
2. $f_\theta(-x) = 1 - f_\theta(x)$
3. $\dfrac{\partial \ln}{\partial \theta_j}\left(f_\theta(x)\right) = x_j\left(1 - f_\theta(x)\right)$
4. $\dfrac{\partial \ln}{\partial \theta_j}\left(1 - f_\theta(x)\right) = -x_j f_\theta(x)$
5. $\dfrac{\partial \mathcal{L}}{\partial \theta_j}\left(\theta; x^{(i)}, y_i\right) = x_j^{(i)}\left(f_\theta\left(x^{(i)}\right) - y_i\right)$
6. *Conclure que la descente de gradient pour notre problème est :*

$$\theta_j^{t+1} = \theta_j^t - \eta \sum_{i=1}^{n} x_j^{(i)}\left(f_\theta\left(x^{(i)}\right) - y_i\right)$$

À partir de la régression logistique, nous pouvons obtenir la **matrice de confusion** :

		Prédit	
		Classe 0 (baisse)	Classe 1 (hausse)
Réel	Classe 0	TN	FP
	Classe 1	FN	TP

On définit l'**accuracy** comme la proportion de bonne prédiction :

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Exercice 6

On souhaite prédire une hausse exceptionnelle, et dans le dataset que l'on a à disposition, il y a 1% de classe 1 (hausse exceptionnelle). Construire un algorithme qui permet d'atteindre 99% d'accuracy.

Comment mesurer la performance d'une classification ?

Précision, Recall et F1-Score

		Prédit	
		Classe 0 (baisse)	Classe 1 (hausse)
Réel	Classe 0	TN	FP
	Classe 1	FN	TP

À partir de la matrice de confusion, on peut extraire d'autre métriques qui apportent des éclairages plus précis :

$$\text{Précision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2}{\frac{1}{\text{Précision}} + \frac{1}{\text{Recall}}}$$

Exercice 7

Vous demandez à votre data scientist de concevoir un algorithme qui priorise les mails en essayant de prédire ceux qui sont les plus importants. Un dataset d'entraînement et un dataset de test lui sont fournit. Dans le dataset de test, il y a 1000 mails dont 200 sont importants. Il vous présente un premier modèle qui, pour un certain seuil (A), présente la matrice de confusion suivante :

		Prédit	
		Classe 0	Classe 1
Réel	Classe 0	700	100
	Classe 1	50	150

Pour un autre seuil (B) il présente cette matrice de confusion :

		Prédit	
		Classe 0	Classe 1
Réel	Classe 0	760	40
	Classe 1	80	120

1. *Calculer l'accuracy, la précision, le recall et le F1-score de chacun des seuils.*
2. *Conclure sur le seuil que vous souhaitez conserver.*