TP/TD 1 : RÉGRESSIONS LINÉAIRES ET POLYNOMIALES

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, AUTOMNE 2014

Rémi Lajugie
remi.lajugie@ens.fr

Résumé. Dans ce TP, on considère le problème de la régression de $\mathbb{R}$ dans $\mathbb{R}$. Il s'agit, étant donné des observations $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^2$ et un sous ensemble fonctionnel $\mathcal{F} \subset L^2(([0,1])$ de rechercher le minimiseur du risque quadratique.

Ce que vous serez amené à faire dans ce TP servira à d'autres occasions durant le cours. Conserver les codes produits est donc une bonne idée.

Tout au long de ce TP on considère le couple de variables aléatoires (X, Y), tel que X suive une loi uniforme sur $[0, 1]$ et $Y = f^*(X) + \epsilon$ où $f^*(x) = \exp(x)$ et ϵ est un bruit Gaussien indépendant de X de variance unité et de moyenne nulle.

On rappelle qu'un bruit gaussien dans $\mathbb{R}$ est une variable aléatoire distribué suivant une loi de probabilité admettant comme densité par rapport à la mesure de Lebesgue $p(x) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2\sigma^2})$, avec $\sigma \in \mathbb{R}$

1. Prologue : Rapides rappels de probabilités

On rappelle dans cette partie les quelques résultats de probabilité nécessaires pour le cours. En apprentissage, il n'est pas nécessaire de connaître in extenso toute la théorie des probabilités. Un bon apercu de ce qu'il faut savoir peut etre trouvé dans des livres d'apprentissage comme [1] (chapitre 1) ou [3] (chapitre 2). Ceux qui souhaiteront approfondir les notions de proba que l'on effleurera durant le cours et les TD pourront consulter un cours de probabilité et de théorie de la mesure [2] ou un ouvrage d'introduction aux probabilités comme [4].

1.1. Fondements mathématiques.

Tribus. Soit Ω un espace d'événements. On suppose qu'il existe sur Ω une famille de parties disposant des propriétés suivantes :
- $\Omega \in \mathcal{A}$.
- $\mathcal{A}$ est stable par passage au complémentaire.
- $\mathcal{A}$ est stable par union dénombrable.

Une telle famille est une tribu sur Ω.

Probabilité. Une *probabilité* sur $(\Omega, \mathcal{A})$ est une application $\mathbb{P}$ de $\mathcal{A}$ dans $\mathbb{R}^+$ qui vérifie :
- $\mathbb{P}(\Omega) = 1$.
- Pour une suite $(A_n)_{n\in\mathbb{N}}$ d'éléments de $\mathcal{A}$ disjoints deux à deux, on a $\mathbb{P}(\cup_n A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$.

Variable aléatoire. Dans la suite on considèrera deux types de fonctions :

- Celles à valeurs dans $\mathbb{R}^n$. Dans ce cas on considère que $\mathbb{R}^n$ est munie de sa tribu borélienne $\mathcal{B}$, la plus petite tribu rendant les pavés mesurables.
- Celles à valeurs dans un espace discret $\mathbb{D}$ (identifié à un sous ensemble de $\mathbb{N}$). Dans ce cas, on considère la tribu des sous ensembles de $\mathbb{D}$.

Dans tous les cas, si on appelle $\mathcal{A}$ la mesure associée à l'espace d'arrivée de la fonction X, on écrit

$$\forall A \in \mathcal{A}, \mathbb{P}(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)).$$

Une telle fonction X est appelée une *variable aléatoire*.

1.2. Variables aléatoires à densité.

Variables aléatoires discrètes. Une variable aléatoire discrète admet une densité par rapport à la mesure de comptage sur $\mathbb{D}$ si il existe une fonction $p : \mathbb{D} \to [0, 1]$ telle que

$$\mathbb{P}(X \in A) = \sum_{x \in A} p(x).$$

On dit que $p(x)$ est la probabilité que X vaille x.

Très souvent, on appelle cette densité la *masse* de probabilité.

Variables aléatoires continues. Une variable alatoire continue admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ si il existe une fonction $p : \mathbb{R}^n \to [0, 1]$ telle que

$$\mathbb{P}(X \in A) = \int_A p(x)dx.$$

De manière abusive, on dit souvent que $p(x)$ est la probabilité que X vaille x.

A partir de maintenant on supposera que toutes les variables aléatoires que l'ont rencontre admettent une densité, par rapport à la mesure de Lebesgue si elles sont continues, par rapport à la mesure de comptage si elles sont discrètes.

1.3. Espérance, conditionnement.

Espérance. On appelle *espérance* de la variable aléatoire continue X, la quantité, notée $\mathbb{E}[X] = \int_{\mathbb{R}^n} xp(x)dx$. Dans le cas discret, $\mathbb{E}[X] = \sum_{x \in \mathbb{N}} xp(x)$.

Conditonnement. Soit X et Y deux variables aléatoires.

On devrait noter leurs densités avec deux notations différentes : par exemple $p(x)$ et $q(y)$. Cependant, en apprentissage, on utilise plutôt l'abus de notation suivant : la densité de X est $p(x)$ et celle de Y, $p(y)$.

On appelle la *densité jointe* $p(x, y)$ de X et de Y, la densité de la variable aléatoire (X, Y). On dit que X et Y sont *indépendantes* si $p(x, y) = p(x)p(y)$. Continuant cet abus de notation, on appelle *probabilité conditionnelle* de l'événement $X \in A \in \mathcal{A}$ sachant l'événement $Y \in B \in \mathcal{A}$ la quantité

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(X \in A \cap Y \in B)}{\mathbb{P}(Y \in B)}.$$

En supposant que $p(y) > 0$, on appelle *densité conditionnelle* la densité de probabilité définie par $p(x|y) = \frac{p(x,y)}{p(y)}$.

Espérance conditionnelle. L'*espérance conditionnelle* est alors la **variable aléatoire** notée $\mathbb{E}[X|Y]$ dont la densité est donnée par $\int p(x|y)dx$.

L'espérance conditionnelle possède les propriétés suivantes :
- Si X et Y sont indépendantes, alors $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
- $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ (loi de l'espérance itérée)
- Pour une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbb{E}[f(X)|X] = f(X)$.

Conditionnement "mixte". Il est fréquent qu'en apprentissage on s'intéresse au comportement de deux variables jointes vivant dans des espaces différents. Par exemple pour de la classification binaire, la sortie Y vit dans $\{0, 1\}$ alors que les descripteurs à partir desquels on veut faire la prédiction sont dans $\mathbb{R}^n$. Dans ce cas, la densité conditionnelle du couple (X, Y) s'écrit toujours $p(x, y)$.

Règle de Bayes. On appelle *Règle de Bayes* la formule d'échange entre information a priori et a posteriori

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

La règle de Bayes reste vraie en remplacant les probabilités par les densités conditionnelles.

2. Partie II : Premières simulations

1) a) Rappelez la définition du risque pour le cas de la perte quadratique.

b) Rappeler l'expression de la fonction cible associée à une perte. Si l'on considère l'ensemble des fonctions $C^\infty([0, 1])$, quelle est la fonction cible de notre régression ?

2) a) Générer 20 points du plan (x_i, y_i), réalisations i.i.d. des variables aléatoires X et Y. Visualisez les.
En Matlab, la fonction `randn(1,n)` génère un n échantillon de réalisations indépendantes d'une loi normale centrée réduite (moyenne nulle, variance unité). La fonction `rand(1,n)` les génère uniformément sur l'intervalle $[0, 1]$.

b) Séparer ces points en deux ensembles de taille égale. Par la suite, on appellera la première moitié des données "ensemble d'entraînement", et l'autre "ensemble de test".

3) a) Rappeler la définition du risque quadratique pour la régression de Y sur une fonction $f(X)$.

b) Comment, à partir des seules données d'apprentissage peut-on donner une estimation de ce risque ?

4) Commencer par faire une régression linéaire simple. On considère la classe de fonctions $\mathcal{G} = \{g, \exists \theta_1, \theta_2 \in \mathbb{R}, \forall x \in \mathbb{R}, g(x) = \theta_1 x + \theta_2\}$. On cherche dans cette question à estimer les paramètres θ_1^* et θ_2^* qui minimisent $R(\theta_1, \theta_2) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_2)^2$.

a) En écrivant la condition d'annulation du gradient pour R, vérifier que l'expression de θ_1^* et θ_2^* est compatible avec celle vue en cours. (On pourra considérer que la matrice X que l'on veut régresser est la concaténation des vecteurs "augmentés" $(x_i, 1), \ldots$)

b) Estimer les paramètres de la régression linéaire sur les seules données d'entraînement et affichez la droite obtenue sur le même graphique que les données. On pourra pour cela utiliser la commande `polyval` de Matlab.

5) Expliquer comment on peut utiliser l'équation normale de la régression linéaire ($\theta = (\mathbf{X}^T\mathbf{X})^\dagger\mathbf{X}^T\mathbf{X}$) pour estimer les coefficients d'une régression sur des polynômes de degré p (un indice : Souviens-toi de Vandermonde !).

6) Implémenter cette formule pour estimer les coefficients du polynôme de régression de 1 à 9 et représentez les fonctions obtenues sur le même graphe.
On pensera à normaliser la matrice de design. Etant donné une matrice de design $\mathbf{X} \in \mathbb{R}^{n\times d}$ comportant n données de dimension d, normaliser les données est en général une bonne pratique, notamment pour s'assurer d'une certaine stabilité numérique.

7) Calculer pour chacun de ces polynômes, l'erreur de régression sur les données d'entraînement. Que constatez vous ?

8) On va désormais étudier la capacité de généralisation (ou de prédiction) des fonctions de régression.

a) En utilisant les fonctions de régression estimées sur les données d'entraînement, calculer le risque empirique de régression sur les données de test.

b) Représenter l'évolution des erreurs d'entraînement et de test sur la même figure. Commentez les courbes obtenues.

9) Reprendre votre code en augmentant le nombre de données générées. Regardez l'évolution des erreurs de test et d'entraînement. Commentez le résultat.

3. Partie III : Erreurs et excès en tous genres

Le but de cette partie est d'étudier l'évolution empirique de certaines quantités théoriques vues en cours : l'excès de risque ainsi que l'erreur d'approximation par une classe de fonction

$\mathcal{G}$. On rappelle qu'on a vu en cours la formule de décomposition suivante pour le risque :

$$\underbrace{\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}\left(f^\star\right)}_{\text{excès de risque}} = \underbrace{\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}\left(f_S^\star\right)}_{\text{erreur d'estimation}} + \underbrace{\mathcal{R}\left(f_S^\star\right) - \mathcal{R}\left(f^\star\right)}_{\text{erreur d'approximation}}$$

Dans cette partie, on considère l'espace des fonctions de carré intégrable sur $[0,1]$ que l'on note $L_2([0,1])$. Cet espace est muni de son produit scalaire usuel $< f, g >= \int_{[0,1]} f(x)g(x)dx$ et de la norme associée.

10) On considère le problème de la régression polynomiale de degré au plus k. Commencer par calculer $\mathcal{R}\left(f^*\right)$ le risque de la fonction cible.

11) Considérer maintenant l'estimateur donné par l'équation normale pour la régression polynomiale. Calculer l'excès de risque $\mathcal{R}\left(\widehat{f}_S\right) - \mathcal{R}\left(f^\star\right)$. L'expression finale ne dépendra que des coefficients $(v_1, \ldots, v_k)$ de la régression. Pour le calcul, il peut être utile de se rappeler le résultat suivant : si $X \in \mathbb{R}^n$ et $Y \in \mathbb{R}^n$, on a $\mathbb{E}[X^T Y | D_n] = \mathrm{Tr}(\mathbb{E}[Y X^T | D_n])$. Le calcul, s'il est fait proprement, doit permettre une réponse plus aisée à la question suivante...

12) Trouver une expression analytique de la distance de f et d'un polynôme de coefficients $(\alpha_0, \ldots, \alpha_k)$. On cherchera à donner cette distance comme une forme quadratique (en utilisant le calcul matriciel !). En déduire $\mathcal{R}\left(f_S^\star\right)$. (NB : Dans cette question, si le besoin s'en fait ressentir, supposez que les matrices que vous rencontrez sont inversibles.)

13) Reprendre l'expression de l'erreur d'approximation et utiliser Matlab pour représenter son évolution à mesure que le degré des polynômes croît.

14) A partir des questions précédentes, représenter graphiquement l'évolution de l'erreur d'estimation.

15) Nous avions de la chance, nous connaissions la fonction cible de la régression dans notre cas, mais en pratique elle est inconnue. Si on suppose maintenant que l'on peut échantillonner autant que l'on veut le couple de variables (X, Y) quelle stratégie numérique aurait-on pu adopter pour estimer l'erreur d'approximation ?

16) Conclure.

Références

[1] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[2] Jean-Francois Le Gall. Integration, probabilites et processus aleatoires, disponible en ligne : `www.math.u-psud.fr/~jflegall/IPPA2.pdf`. *Ecole Normale Superieure*, 2006.

[3] Kevin P Murphy. *Machine learning : a probabilistic perspective*. MIT press, 2012.

[4] J.Y. Ouvrard. *Probabilites : Tome 1*. Cassini, 2007.