TP/TD 1 : RÉGRESSIONS LINÉAIRES ET POLYNOMIALES

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, PRINTEMPS 2013

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 d'apprentissage. Conservez donc une trace de vos codes Matlab/Octave dans un fichier .m.

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.

1. Partie I : Rapide échauffement théorique

1) Pour rappel, un bruit Gaussien est une variable aléatoire généré selon une loi suivant une densité : $p(x) = \theta \exp(-\frac{x^2}{2\sigma^2})$ où σ^2 est une constante positive.

a) Calculez la constante de normalisation θ telle que p soit une densité de probabilité (de masse unité). On pourra commencer par démontrer la valeur de l'intégrale dite de Gauss : $\int_{-\infty}^{+\infty} \exp(-x^2)dx = \sqrt{\pi}$.

b) Quelle est la variance associée à une variable aléatoire ayant pour densité p ?

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

b) Rappelez l'expression de la fonction cible associée à cette perte. Quelle est la fonction cible de notre régression ?

2. Partie II : Simulations numériques

Etant donné une matrice de design $X \in \mathbb{R}^{n \times d}$ comportant n données de dimension d, il faut systématiquement normaliser ces données, c'est à dire, pour chaque

entrée : enlever la moyenne de la colonne à laquelle elle appartient et la diviser par l'écart type de cette colonne.
3) a) Générez 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).

b) Séparez 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".

4) a) Rappelez 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 ?

5) Commencez par faire une régression linéaire simple. On considère la classe de fonctions $\mathcal{F} = \{f, \exists \theta_1, \theta_2 \in \mathbb{R}, \forall x \in \mathbb{R}, f(x) = \theta_1 x + \theta_2\}$. Vous chercherez 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, vérifiez 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) Estimez 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.

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

7) Implémentez 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.

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

9) On va désormais regarder la capacité de généralisation (ou de prédiction) des fonctions de régression.

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

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

10) Reprenez 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 $\mathcal{F}$.

11) a) On considère l'ensemble de fonctions $\mathcal{F} = \{\theta_1 x + \theta_2, \theta_1, \theta_2 \in \mathbb{R}\} \subset L^2([0,1])$. Considérez une fonction de $g \in \mathcal{F}$, et donnez sa distance au sens $L^2([0,1])$ à f. Donnez ensuite l'expression de l'erreur d'approximation.

12) Considérez maintenant l'estimateur donné par l'équation normale pour la régression polynomiale. Calculez l'excès de risque en utilisant la formule du cours. L'expression finale ne dépendra que des coefficients $(v_1, \dots, v_k)$ d ela 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 $\mathbb{E}[X^T Y] = \mathrm{Tr}(\mathbb{E}[Y X^T])$. Le calcul, s'il est fait proprement, doit permettre une réponse plus aisée à la question suivante...

13) Toujours dans le cas où la classe d'hypothèses est l'ensemble des polynômes de degré k, essayez de trouver une expression analytique de la distance de f et d'un polynôme de coefficients $(\alpha_0, \dots, \alpha_k)$. On cherchera à donner cette distance comme une forme quadratique (en utilisant le calcul matriciel !). (NB : Dans cette question, si le besoin s'en fait ressentir, supposez que les matrices que vous rencontrez sont inversibles.)

14) Reprenez l'expression de l'erreur d'approximation et utilisez Matlab pour représenter son évolution à mesure que le degré des polynômes croît.

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

16) 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 ?

17) Conclure.