TP/TD 2 : MÉTHODES PAR MOYENNAGE LOCAL ET VALIDATION CROISÉE

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, 2 OCTOBRE 2015

Jean-Baptiste Alayrac
jean-baptiste.alayrac@inria.fr

Résumé. Le but de ce Tp est de mettre en pratique les méthodes de moyennage local vues en cours et de sélectionner leurs paramètres à l'aide de la méthode de validation croisée.

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.

1. Démarrage en douceur : Rappel de cours

1) Soit $h > 0$ et le noyau Gaussien associé $\forall x \in \mathbb{R}^d, K_h(x) = \exp(-\frac{\|x\|_2^2}{h})$

a) Rappelez la définition de l'estimateur de Nadaraya-Watson pour la régression. Dans cet exercice on considère toujours qu'un n-échantillon d'entraînement $(x_1, y_1), \ldots, (x_n, y_n)$ nous est donné.

b) Faites tendre le paramètre h vers 0, quelle règle de décision vue en cours obtient on alors pour presque tous les points de l'espace de départ ? (On s'intéresse ici à la convergence ponctuelle de la fonction de décision associée aux noyaux K_h.)

2. Pour réviser quelques notions importantes du cours : Non-consistance de la règle du plus proche voisin.

Comme dit dans le titre nous allons chercher à démontrer que la règle du 1 plus proche voisin n'est pas consistante au sens vu dans le cours. Dans tout cet exercice, on appelle risque le risque au sens de Vapnik défini par $R(f) = \mathbb{E}[\ell(f(X), Y)]$.

2) Rappelez la définition de la consistance d'une règle d'apprentissage de classifieur $\widehat{f_n} = \mathcal{A}(D_n)$ où $\widehat{f_n} : \mathbb{R}^d \longrightarrow \{0, 1\}$.

Cadre de l'exercice : On considère le cas de la classification binaire (le risque est donc le risque de la perte $0-1$) où l'on dispose d'un n-échantillon $D_n = (x_i, y_i)$ avec $\mathcal{X} = [0, 1], \mathcal{Y} = \{0, 1\}$.

Ce n-échantillon est généré de manière i.i.d. comme suit : les x_i sont la réalisation (ou observation) de variables aléatoires X_i suivant une certaine distribution P admettant une densité p par rapport à la mesure de Lebesgue sur $\mathcal{X}$. Les étiquettes y_i sont la réalisation de variables aléatoires Y_i distribuées en respectant $\forall x \in \mathcal{X}, \eta(x) = \mathbb{P}(Y = 1|X = x) = \alpha > \frac{1}{2}$.

A partir de ce n-échantillon, on construit un classifieur $\widehat{f}_n = \mathcal{A}(D_n)$. Il est important de garder à l'esprit que ce classifieur $\widehat{f}_n$ peut être vu comme la réalisation d'une variable aléatoire dépendant de $D_n = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. Etudier la consistance de la régle de classification définie par $\widehat{f}_n$ revient à s'intéresser au comportement de la variable aléatoire $\widehat{f}_n$ quand n tend vers l'infini.

3) En considérant la distribution des données décrite ci-dessus, donner l'expression du prédicteur de Bayes et le risque associé.

4) On commence par considérer un classifieur quelconque f qui ne dépend pas a priori de l'echantillon d'apprentissage D_n. Soit X la variable aléatoire indépendante de D_n qui intervient dans la définition du risque et qui correspond aux données de test. Montrer que le risque associé à f peut s'exprimer comme (on remarquera qu'un classifieur binaire quelconque s'écrit $f(X) = \mathbb{1}_{\{f(X)=1\}}$) :

$$R(f) = \alpha - (2\alpha - 1)\mathbb{E}_X[f(X)].$$

5) On considère maintenant le classifieur construit par la règle du plus proche voisin $\widehat{f}^1$, et on va démontrer qu'il n'est pas consistant.

a) Montrer que chaque variable aléatoire Y_i est indépendante de $(X_1, \ldots, X_n)$ (on pourra se contenter d'une justification intuitive).

b) En introduisant des variables $B_i(X)$ valant 1 si l'exemple i de l'échantillon est le plus proche voisin de X et 0 sinon, exprimez $\widehat{f}^1(X)$ comme une somme de variables aléatoires. Que vaut $\mathbb{E}_X[\sum_{i=1}^n B_i(X)|X_1, \ldots X_n]$?
Remarque : il est naturel de se demander comment faire en pratique lorsqu'un point des données sur lesquelles on teste le classifieur est à égale distance de deux points d'apprentissage. Dans le formalisme de cet exercice, comme les variables ont une densité cela ne peut arriver qu'avec une probabilité nulle. En revanche, dans un contexte plus général, il est possible d'avoir avec une probabilité non nulle une nouvelle observation équidistante de deux points d'apprentissage. Plusieurs stratégies sont possibles : attribuer aléatoirement à la nouvelle observation le label de l'un ou l'autre des points dont elle est équidistant ou encore attribuer le label de l'observation de plus faible indice. Vous pouvez trouver de plus amples détails sur le sujet dans la section 11.2 de [1].

c) Quelle est l'espérance de Y_i conditionnellement à $(X_1, \ldots, X_n)$? Donnez l'expression de l'espérance de $\widehat{f}^1$ conditionnellement aux $(X_1, \ldots X_n)$ (l'espérance est prise par rapport à X et aux Y_i).

6) Donner le risque de $\widehat{f}^1$ et en déduire que la méthode des 1-ppv n'est pas consistante.

7) On considère maintenant toujours le problème de classification binaire sur $\mathcal{X}$ avec le classifieur $\widehat{f}^K$ des K plus proches voisins. Pour s'affranchir de certaines difficultés, on va supposer que K est impair.

a) En vous inspirant du raisonnement précédent montrer que le risque du classifieur $\widehat{f}^K$ des K plus proches voisins, conditionné sur les données d'entrainement $(X_1, \dots X_n)$ (donc l'espérance du risque de $\widehat{f}^K$ pris par rapport aux seuls Y_i) peut s'exprimer en fonction de la probabilité pour une variable binomiale U de paramètres K et α d'être plus grande que $\frac{K}{2}$ (On rappelle que U est une variable aléatoire binomiale de paramètre α et K si U peut s'écrire comme la somme de K variables de Bernoulli de paramètre α).

b) Montrer que l'espérance du risque associé au classifieur des K plus proches voisins dans ce cas est strictement plus grand que le risque de Bayes $1 - \alpha$. On pourra remarquer que l'espérance de $\widehat{f}^K$ est strictement plus petite que 1 par ce qui précède.

Morale de l'histoire. *De manière plus générale, il n'y a pas de choix universellement consistant du nombre de plus proches voisins, valable indépendamment du nombre de points dans l'ensemble d'entraînement. En revanche, il est possible de vérifier (en exercice à la maison par exemple) que si on considère une suite d'entiers k_n telle que $\lim_{n\to\infty} k_n = \infty$ et $\lim_{n\to\infty} k_n/n = 0$ alors les hypothèses du théorème de Stone sont vérifiées pour toutes les distributions. Ainsi pour la suite de règles des k_n plus proches voisins, on a la consistance universelle.*

3. Implémentation : K plus proches voisins et validation croisée

Pour ce TP on va utiliser des données réelles : des chiffres manuscrits, que vous pouvez télécharger à l'adresse `http://www.math.ens.fr/cours-apprentissage/mnist_digits.mat`. Vous êtes censés les avoir déjà manipulé dans le TP 0, mais si par amnésie vous ne vous rappelez plus comment on charge des fichiers sous Matlab/Octave, vous pouvez utiliser la commande `load`.
Pour la classification avec K classes, on appelle souvent matrice de confusion associée à des données $D_n = (x_t, y_t)$ la matrice $M \in \mathbb{N}^{K\times K}$ telle que $M_{i,j}$ soit le nombre d'éléments dont la vraie classe soit i et dont la classe prédite par le classifieur g soit j.
NB : Etant donné qu'il y a plus de 66000 images dans le jeu de données, on va travailler avec un sous ensemble de ces 66000 images afin de ne pas dépasser la mémoire disponible de votre ordinateur.
8) Commencez donc par reprendre contact avec les données. Elles sont composées d'un vecteur de labels `y` et d'images 28 pixels par 28 sous forme d'une matrice `x` de vecteurs linéarisés (chaque ligne de la matrice `x` correspond à une image).

a) Sélectionnez 6000 images au hasard dans le jeu de données. Dressez un histogramme des classes pour vérifier que cette sélection n'a pas déséquilibrez les classes (commande `hist`). Pour la suite on travaillera avec ce jeu de données restreints (cf remarque plus haut).

b) Mettez quelques images sous forme matricielle `reshape(x(i,:),28,28)` et affichez les.

c) Séparez les images en deux parties (dans les proportions $1/3, 2/3$ par exemple) : un ensemble d'entraînement et un ensemble de test.

9) On va maintenant implémenter la règle de classification par plus proches voisins. Pour cela, vous pourrez avoir besoin de la fonction Matlab/Octave sqdist que vous pouvez trouver à cette adresse : `http://www.math.ens.fr/cours-apprentissage/sqdist.m`. Cette fonction permet, étant donné deux matrices de design (attention la fonction prend en argument la transposée !) de calculer les distances euclidiennes au carré entre les points.

a) Ecrivez une fonction qui prenne en entrée le nombre de plus proches voisins désirés, les données d'entrainement et les données de test et ressorte la matrice de confusion sur l'ensemble de test.

b) Affichez l'erreur de classification sur les ensembles d'entraînement et de test en fonction du nombre de k, nombre de plus proches voisins pris en compte (attention la "complexité" est décroissante avec le nombre de voisins pris en compte). Vous pouvez faire varier k entre 1 et 20 par exemple.

c) Séparez votre ensemble d'entraînement en un ensemble d'entraînement réduit et un ensemble de validation (on appelle en général cette technique la "validation simple"). En utilisant le code précédent, écrivez une fonction qui va utiliser l'ensemble de validation pour sélectionner le meilleur paramètre nombre de voisins k au sens du nombre d'erreurs commises sur l'ensemble.

d) Séparez plusieurs fois de manière aléatoire votre ensemble d'entraînement. L'estimateur du nombre de plus proches voisins est il stable ?

10) On souhaite désormais sélectionner le nombre de plus proches voisins optimal par validation croisée. Vous allez, en utilisant les données d'entraînement implémenter la technique de la K-fold validation croisée pour K=8. Faites le partitionnement des données d'entraînement plusieurs fois de manière aléatoire et regardez le comportement du nombre de plus proches voisins sélectionné. Que remarquez vous ?

Références

[1] L. Devroye, L. Gyorfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* `http://www.szit.bme.hu/~gyorfi/pbook.pdf`, 1996.