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

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, PRINTEMPS 2013

Rémi Lajugie
remi.lajugie@ens.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) On considère $h > 0$ 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, si l'on considère un n-échantillon d'entraînement $(X_1, Y_1), \ldots, (X_n, Y_n)$.

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

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

3) On va maintenant implémenter la règle de classification par plus proches voisins.

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).

c) Séparez votre ensemble d'entraînement en un ensemble d'entrainement 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
s

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

4) 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'entrainement plusieurs fois de manière aléatoire et regardez le comportement du nombre de plus proches voisins sélectionné. Que remarquez vous ?

3. Pour réviser quelques notions importantes du cours : Non-consistance des plus proches voisins si k est choisi indépendemment de n

Le titre de l'exercice est trompeur, nous allons seulement nous borner à démontrer la non consistance de la règle du plus proche voisin.

5) a) Rappelez la définition de la consistance d'une règle de classification $g : \mathbb{R}^d \longrightarrow \{0, 1\}$.

b) Rappelez brièvement une condition suffisante assurant la consistance dans le cas des méthodes par moyennage local.

6) On considère le cas de la classification binaire où l'on dispose d'un n-échantillon $D_n = (X_i, Y_i)$ avec $\mathcal{X} = [0, 1], \mathcal{Y} = \{0, 1\}$. On considère que les entrées X_i sont distribuées selon une certaine distribution P admettant une densité f par rapport à la mesure de Lebesgue sur $\mathcal{X}$ et que les étiquettes Y_i sont distribuées en respectant $\forall x \in \mathcal{X}, \eta(x) = \mathbb{P}(Y = 1|X = x) = \alpha > \frac{1}{2}$.

7) Donnez l'expression du classifieur de Bayes dans ce cas et le risque associé.

8) On commence par considérer un classifieur quelconque g. Montrez que le risque associé à peut s'exprimer comme (on remarquera qu'un classifieur binaire quelconque peut s'écrire comme $g(x) = 1_{g(x)=1}$) :

$$R(g) = \alpha - (2\alpha - 1)\mathbb{E}[g(x)].$$

9) On va maintenant montrer que g_1, l'estimateur du plus proche voisin n'est pas une règle consistante.

a) Montrez 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 de Bernoulli $B_i(x)$ valant 1 si l'exemple i de l'échantillon est le plus proche voisin de x et 0 sinon, exprimez g_1 comme une somme de variables aléatoires. Que vaut $\forall x, S(x) = \sum_{i=1}^{n} B_i(x)$?

c) Quelle est la moyenne des Y_i ? Donnez l'expression de l'espérance de g_1.

10) Calculez le risque de g et en déduire que la méthode des 1-ppv n'est pas consistante.

11) On considère maintenant toujours le problème de classification binaire sur $\mathcal{X}$ avec le classifieur g_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 montrez que le risque de l'estimateur g_K des K plus proches voisins peut s'exprimer en fonction de la probabilité pour une variable binomiale de paramètres K et α (en d'autres termes, U peut s'écrire comme la somme de K variables de Bernoulli de paramètre α).

b) Montrez que le 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 g est strictement plus petite que1 par ce qui précède.

Morale de l'histoire. *De manière plus générale, il n'y a pas de choix universel 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 ou en révision de l'examen 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 la suite de règles des k_n plus proches voisins.*