CORRECTION TP/TD2 : METHODES PAR MOYENNAGE LOCAL ET VALIDATION CROISÉE

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

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

Pour toute question relative à ce corrigé (ou si vous repérez ce qui vous semble être une faute), vous pouvez m'envoyer un mail.

1. Partie I : Démarrage en douceur : Rappel de cours

1) a) Soit des données d'entrainement $(X_1, Y_1), \ldots, (X_n, Y_n)$. On rappelle les poids associés à chaque données d'entrainement X_i pour la méthode Nadaraya-Watson (vue dans le cours) :

$$W_i(x) = \frac{\exp\left(-\frac{\|x-X_i\|_2^2}{h}\right)}{\sum_j \exp\left(-\frac{\|x-X_j\|_2^2}{h}\right)}$$

On peut donc écrire η_h, l'estimateur de Nadaraya-Watson pour la régression avec le noyau gaussion de la sorte :

$$\eta_h(x) = \sum_{i=1}^{n} W_i(x) Y_i$$

b) On s'intéresse ici à savoir ce qu'il se passe lorsque le paramètre h tend vers 0. Comme un prédicteur est simplement une fonction de $\mathbb{R}^d$ dans l'espace de sortie, on peut s'intéresser ici à la convergence ponctuelle en chaque point de l'espace.
On peut considérer l'ensemble $D(x) = \text{argmin}_i \|x - X_i\|_2^2$. On considère maintenant un x tel que $D(x)$ soit un singleton. Dans ce cas, par une simple factorisation on peut écrire l'estimateur de Nadaraya-Watson comme :

$$\eta_h(x) = \frac{Y_{D(x)} + \sum_{j \neq D(x)} \exp\left(\frac{\|x-X_{D(x)}\|_2^2 - \|x-X_j\|_2^2}{h}\right) Y_j}{1 + \sum_{j \neq D(x)} \exp\left(\frac{\|x-X_{D(x)}\|_2^2 - \|x-X_j\|_2^2}{h}\right)}$$

En observant que $\forall j \neq D(x), \|x - X_{D(x)}\|_2^2 - \|x - X_j\|_2^2 < 0$ on a simplement que $\lim_{h \to 0} \eta_h(x) = Y_{D(x)} = \eta(x)$. Cet estimateur limite est celui du plus proche voisin. Notons enfin que si $D(x)$ n'est pas un singleton et consiste en la réunion d'indices $\{i_1, \ldots, i_K\}$, la fonction de décision associée aux noyaux K_h ne converge pas vers l'estimateur du plus proche voisin, mais, en reprenant le raisonnement précédent converge vers $\frac{1}{K} \sum_{i_k=i_1}^{i_K} Y_{i_k}$ (estimateur des K plus proche

voisins).

Conclusion. Si on suppose que l'ensemble des x tels que $D(x)$ n'est pas un singleton est de mesure nulle (ce qui est bien le cas dans le cas d'une distribution continue comme ici), la règle de décision induite par les K_h tend presque surement vers la règle du plus proche voisin.

2. Partie II : Non consistance de la règle du plus proche voisin

Durant tout l'exercice on notera $\mathcal{X} = [0, 1]$

2) a) On dit qu'un règle d'apprentissage est consistante pour une distribution $\mathbb{P}_{X,Y}$ générant les données d'entrainement D_n si on a :

$$\mathbb{E}_{D_n}\left[\mathcal{R}\left(\hat{f}_n\right)\right] \underset{n\to+\infty}{\longrightarrow} \min_{f\in\mathcal{F}} \mathcal{R}\left(f\right) =: \mathcal{R}\left(f^\star\right) =: R^\star$$

3) Dans le cas de la classification binaire, puisque l'on identifie l'espace de sortie $\mathcal{Y}$ à $\{0, 1\}$ il est usuel d'exprimer le prédicteur de Bayes comme une indicatrice, avec la notation $\eta(x) = \mathbb{P}(Y = 1|X = x)$,

$$f^\star = \mathbb{1}_{\{x\in\mathcal{X},\eta(x)\geq 1/2\}}$$

Avec un léger abus de notation on se permet d'écire que $\{x \in \mathcal{X}, \eta(x) \geq 1/2\} =: \{\eta(x) \geq 1/2\}$. En général on note donc $f^\star = \mathbb{1}_{\{\eta(x)\geq 1/2\}}$.

Dans notre cas, comme la fonction $\eta(x)$ est constante et est plus grande que $1/2$ sur $\mathcal{X}$ le prédicteur de Bayes correspond tout simplement au prédicteur qui prédit la classe 1 en tout point de $\mathcal{X}$. Le risque de Bayes est donc donné par :

$$\mathbb{E}_Y\left[\mathbb{1}_{\{Y=0\}}\right] = 1 - \alpha.$$

4) On remarque que n'importe quel classifieur binaire f peut s'écrire comme $\mathbb{1}_{\{f(x)=1\}}$. Gardant cette remarque en tête, on écrit la définition du risque

$$\begin{aligned}
R(f) &= \mathbb{E}_{X,Y}\left[\mathbb{1}_{\{f(X)\neq Y\}}\right] \\
&= \mathbb{E}_{X,Y}\left[\mathbb{1}_{\{f(X)=1\}}\mathbb{1}_{\{Y=0\}} + \mathbb{1}_{\{f(X)=0\}}\mathbb{1}_{\{Y=1\}}\right] \\
&= \mathbb{E}_{X,Y}\left[f(X)\mathbb{1}_{\{Y=0\}} + (1 - f(X)\mathbb{1}_{\{Y=1\}}\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_Y\left[f(X)\mathbb{1}_{\{Y=0\}} + (1 - f(X)\mathbb{1}_{\{Y=1\}} \mid X\right]\right] \\
&= \mathbb{E}_X\left[f(X)(1 - \eta(X)) + (1 - f(X))\eta(X)\right] \\
&= \alpha - (2\alpha - 1)\mathbb{E}_X\left[f(X)\right]
\end{aligned}$$

.

Remarque. : Le même genre de décomposition a été vue en cours pour le plugin.

5) a) On a d'une part que comme les paires (X_i, Y_i) sont iid :

$$\mathbb{P}(Y_i = 1|\ X_1, \ldots, X_n) = \mathbb{P}(Y_i = 1|\ X_i).$$

D'autre part comme la distribution conditionnelle des Y sachant X ne dépend pas de X on a :

$$\mathbb{P}(Y_i = 1|\ X_i) = \mathbb{P}(Y_i = 1) = \alpha.$$

Finalement on a bien l'indépendance de Y_i avec $(X_1, \dots X_n)$.

b) Avec les notations introduites, on a que $\forall x, \hat{f}_1(x) = \sum_{i=1}^n B_i(x) Y_i$. D'autre part on a simplement que $\forall x, \sum_{i=1} B_i(x) = 1$. D'où $\mathbb{E}_X \left[\sum_{i=1} B_i(X) \mid X_1, \cdots, X_n\right] = 1$.

c) Pour la suite on note $\vec{\mathbf{X}} = (X_1, \dots, X_n)$ et $\vec{\mathbf{Y}} = (Y_1, \dots, Y_n)$. L'espérance de Y_i conditionnelement à $\vec{\mathbf{X}}$ est simplement l'espérance de Y_i grâce à l'indépendance prouvée précédement. Or, l'espérance des Y_i vaut α (en écrivant $\mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1) = \mathbb{E}_X[\mathbb{P}(Y_i = 1 |\ X)] = \alpha$).
On a alors :

$$\begin{aligned}
\mathbb{E}_{X,\vec{\mathbf{Y}}}\left[\hat{f}_1(X) \mid \vec{\mathbf{X}}\right] &= \mathbb{E}_{X,\vec{\mathbf{Y}}}\left[\sum_{i=1}^n B_i(X) Y_i \mid \vec{\mathbf{X}}\right] \\
&= \sum_{i=1}^n \mathbb{E}_{X,\vec{\mathbf{Y}}}\left[B_i(X) Y_i \mid \vec{\mathbf{X}}\right] \\
&= \sum_{i=1}^n \mathbb{E}_X\left[B_i(X) \mid \vec{\mathbf{X}}\right] \mathbb{E}_{\vec{\mathbf{Y}}}\left[Y_i \mid \vec{\mathbf{X}}\right] \text{ (indépedance de } B_i(X) \text{ avec } Y_i \text{ sachant } \vec{\mathbf{X}}) \\
&= \sum_{i=1}^n \mathbb{E}_X\left[B_i(X) \mid \vec{\mathbf{X}}\right] \alpha \\
&= \alpha \mathbb{E}_X\left[\sum_{i=1}^n B_i(X) \mid \vec{\mathbf{X}}\right] \\
&= \alpha
\end{aligned}$$

.
6) On a donc $\mathbb{E}_{D_n}\left[\mathbb{E}_X\left[\hat{f}_1 \mid D_n\right]\right] = \alpha$. Ainsi on a $\mathbb{E}_{D_n}\left[\mathcal{R}(\hat{f}_1)\right] = \alpha - (2\alpha - 1)\alpha = 2\alpha(1-\alpha)$. Comme $\alpha > 1/2$, le risque de la méthode des 1-ppv est strictement supérieur au risque de la fonction cible pour tout n. La méthode n'est donc pas consistante pour la distribution considérée. Elle n'est donc pas universellement consistante.

7) a) Soit un n-échantillon. On introduit la notation $\mathcal{V}_k(x)$ comme étant les indices des k plus proches voisins de x dans l'ensemble d'entrainement (voisinage de x au sens des k plus proches voisins). On peut alors écrire que :

$$\forall x \in \mathcal{X},\ \hat{f}_k = \mathbb{1}_{\{\hat{\eta}_k(x) \geq 1/2\}} \text{ où } \hat{\eta}_k(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_k(x)} Y_i$$

On a alors que :

$$
\begin{aligned}
\mathbb{E}_{X,\vec{\mathbf{Y}}}\left[\hat{f}_k(X) \mid \vec{\mathbf{X}}\right] &= \mathbb{E}_{X,\vec{\mathbf{Y}}}\left[\mathbb{1}_{\{\hat{\eta}_k(x)\geq 1/2\}}(X) \mid \vec{\mathbf{X}}\right] \\
&= \mathbb{P}_{X,\vec{\mathbf{Y}}}\left(\hat{\eta}_k(X) \geq 1/2 \mid \vec{\mathbf{X}}\right) \\
&= \mathbb{P}_{X,\vec{\mathbf{Y}}}\left(\frac{1}{k}\sum_{i\in\mathcal{V}_k(X)} Y_i \geq \frac{1}{2} \mid \vec{\mathbf{X}}\right) \\
&= \mathbb{P}_{X,\vec{\mathbf{Y}}}\left(\sum_{i\in\mathcal{V}_k(X)} Y_i \geq \frac{k}{2} \mid \vec{\mathbf{X}}\right) \\
&= \mathbb{E}_{\vec{\mathbf{X}}}\left[\mathbb{P}_{\vec{\mathbf{Y}}}\left(\sum_{i\in\mathcal{V}_k(X)} Y_i \geq \frac{k}{2} \mid \vec{\mathbf{X}}, X\right)\right] \\
&= \mathbb{E}_{\vec{\mathbf{X}}}\left[\beta\right] \text{(Voir justification plus bas)} \\
&= \beta,
\end{aligned}
$$

.

où β est simplement la probabilité pour une variable binomiale U de paramètre k et α d'être plus grande que $\frac{k}{2}$. Une fois X fixée à un certain x, la somme des Y_i sur les indices de ce voisinage se comporte simplement comme une somme de k Bernouilli de paramètre α. Ceci justifie le passage à l'avant dernière ligne. Cette probabilité est une certaine constante positive qui ne dépend que de k, le nombre de plus proches voisins considérés..

b) Le risque du classifieur des k plus proches voisins vaut donc simplement $\alpha - (2\alpha - 1)\beta$. Ce risque ne dépend ni de $\vec{\mathbf{X}}$, ni de n. Or ce risque est strictement plus grand que le risque de Bayes $1 - \alpha$. On en déduit que la méthode des k plus proches voisins n'est pas consistante si le nombre de voisins est fixé indépendemment de la taille de l'échantillon.

3. Partie III : Implémentation, K plus proches voisins et validation croisée

8) **9**) **10**) Voir code matlab.