Devoir 2: Rapport

GIF-7005: Introduction à l'apprentissage machine

Stéphane Caron

17 Octobre 2018

Contents

Question 1

Dans cette question, l'objectif est d'estimer de manière non-parametrique la densité d'une loi de mélange entre deux lois normales.

Partie a

Dans la première partie, nous estimerons la densité de la loi mélange par la méthode de l'histogramme. La figure 1 illustre la densité de la loi en échantillonant 50 et 10 000 observations de la loi mélange.

Dans la figure 1, la ligne orange correspond à la courbe de densité réelle de la loi mélange. On remarque qu'avec 50 observations, l'estimation de la densité n'est pas très précise alors qu'avec 10 000 observations, on se rapproche beaucoup plus de la vraie densité.

Partie b

Dans cette partie, nous estimerons encore la densité de la loi de mélande, mais cette fois-ci avec une estimation par noyau *boxcar*. Cette méthode d'estimation n'est pas continue, mais elle évite de devoir poser une origine alors qu'une fenêtre est appliquée à chaque valeur du support. La figure 2 illustre l'estimation de la densité avec 50 et 10 000 observations selon différentes valeurs de *bandwidth* (h). Cette dernière valeur, contrôle la largeur de la fenêtre autour de laquelle nous allons considérer les données autour d'un point x du support.

À partir de la figure 2, on peut remarquer 2 choses. Premièrement, moins il y a d'observations, plus il y a de variations dans les estimations de la densité pour les valeurs du support. Cela se voit par les escaliers plus prononcés dans la figure avec 50 observations seulement. Deuxièment, plus la fenêtre d'estimation (*bandwidth*) est grande, plus la densité est constante sur le support. Cela fait du sens puisqu'avec une grande fenêtre, on considère beaucoup de données pour estimer la densité en un point. Si la fenêtre est petite, on donnne plus d'importance à la densité locale ce qui fait en sorte qu'on remarque davantage les deux cloches normales.

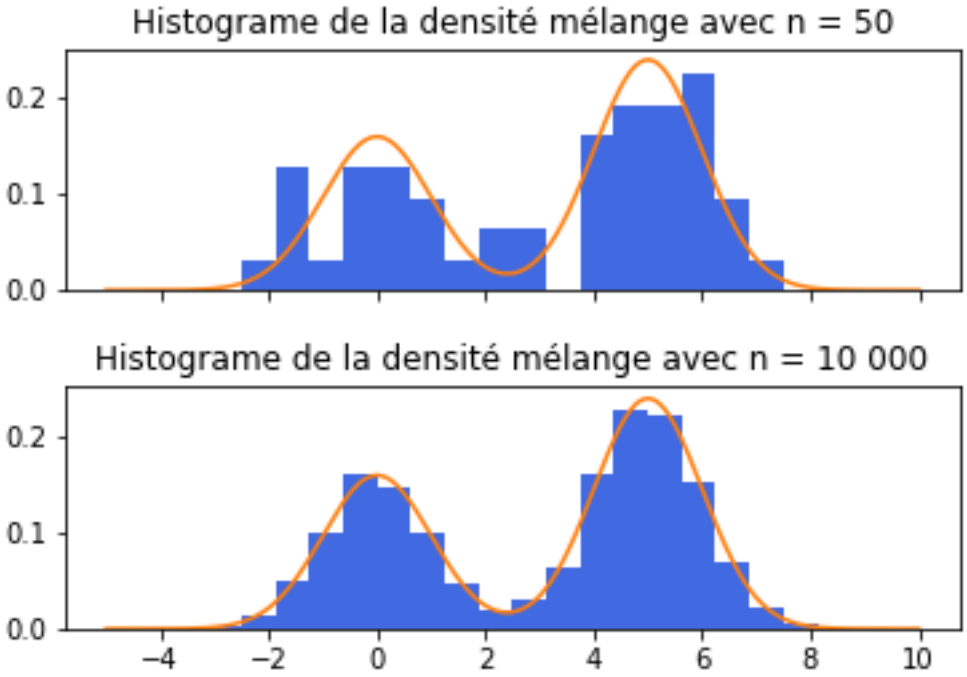

Figure 1: Estimation de la densité par histogramme avec 50 et 10 000 observations.

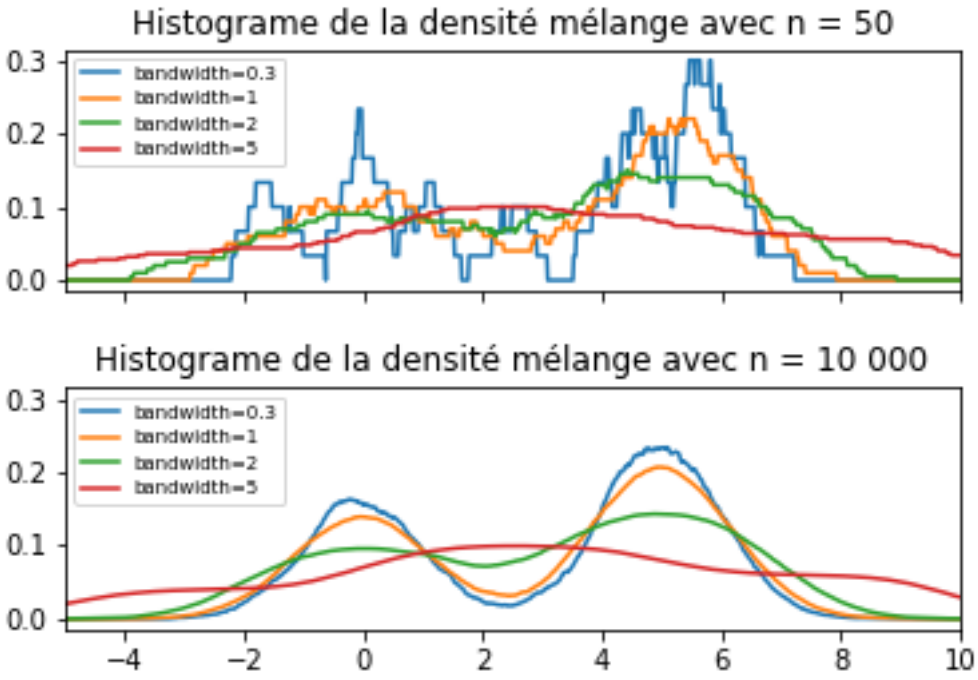

Figure 2: Estimation de la densité par noyau *boxcar* avec 50 et 10 000 observations selon différentes valeurs de *bandwidth*.

Partie c

Les méthodes d'estimation par noyau et la méthode des k-PPV sont tous des méthodes non-paramétriques utilisées pour estimer des densités. La principale différence entre les deux types de méthodes est que la fenêtre h, qui définit l'influence des données autour d'un point, est fixe pour les méthodes par noyau, mais flexible pour la méthode basée sur le k-PPV. Pour cette dernière, la fenêtre s'ajuste en fonction de la densité locale des données.

Par exemple, si une observation est située très loin dans l'espace par rapport aux autres observations, la méthode basée sur les k-PPV va tout de même chercher les k observations les plus proches. Dans ce cas-ci, la fenêtre h sera très grande. C'est une des raisons pourquoi que les méthodes par noyau sont mieux adaptées pour estimer des densités. Ces dernières vont considérer une fenêtre h fixe, ce qui fera en sorte que la densité sera pratiquement nulle pour cette observation.

Partie d

La méthode basée sur les k-PPV est relativement bien adaptée pour le classement et la régression. En effet, cette méthode est très simple et intuitive dans un contexte de modélisation. Pour ce qui est de l'estimation de la densité, cette méthode souffre du problème mentionné à la partie c).

Question 2

Partie a

Dans cette première partie, il faut faire le développement mathématique qui nous permettra de calculer l'ajustement à faire sur les paramètres w_i de notre discriminant linéaire. Pour ce faire, il faut calculer la dérivée de notre critère d'erreur par rapport à chacun de nos paramètres w_i.

Il faut donc calculer:

$$\nabla_w E = \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, ..., \frac{\partial E}{\partial w_D}\right]$$

Ces dérivées partielles nous indiqueront les ajustements à apporter sur chacun des paramètres w_i pour chacune des itérations effectuées lors de la descentre du gradient. La mise à jour pourra s'effectuer comme suit:

$$w_i = w_i + \Delta w_i$$

où $\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$ pour $i = 0, ..., D$.

Pour $i = 1, ..., D$, la dérivée partielle se développe comme suit:

$$\frac{\partial E}{\partial w_i} = \frac{\partial\left(\frac{1}{2}\sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]^2}{||\mathbf{x}^t||^2}\right)}{\partial w_i}$$

$$= -2 \times \frac{1}{2} \sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]}{||\mathbf{x}^t||^2} \times x_i^t$$

$$= -\sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]}{||\mathbf{x}^t||^2} \times x_i^t$$

où Y correspond à l'ensemble des données mal classées.

Pour $i = 0$, la dérivée partielle se développe de manière similaire comme suit:

$$\begin{aligned}\frac{\partial E}{\partial w_0} &= \frac{\partial\left(\frac{1}{2}\sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]^2}{||\mathbf{x}^t||^2}\right)}{\partial w_0} \\ &= -2 \times \frac{1}{2} \sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]}{||\mathbf{x}^t||^2} \\ &= -\sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]}{||\mathbf{x}^t||^2}\end{aligned}$$

Au final, on peut définir les ajustement à apporter aux paramètres w_i comme suit:

$$\Delta w_i = \begin{cases} \eta \times \sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]}{||\mathbf{x}^t||^2} \times x_i^t, & i \geq 1 \\ \eta \times \sum_{x^t \in Y} \frac{[r^t - (\mathbf{w}^T\mathbf{x} + w_0)]}{||\mathbf{x}^t||^2}, & i = 0 \end{cases}$$

Partie b

Partie c

Partie d

Partie e