TP Module - Probabilité et Statistique
Master 1 - Informatique - Université Paris-Sud

Lundi 20 Janvier 2020
Durée : 2h00
Documents autorisés : supports et notes de cours

Pour ce TP, il vous est demandé de rendre un fichier notebook jupyter clairement commenté et exécutable. Vous enverrez le fichier à l'adresse aurelien.decelle@u-psud.fr à la fin de la séance. Le sujet est composé de deux parties indépendantes, vous pouvez traiter les questions dans l'ordre qui vous va.

1 Variables aléatoires continues :

Dans cette partie, on va s'intéresser aux variables aléatoires distribuées selon la loi uniforme. Commençons par regarder la distribution uniforme,

$$p(x) = \begin{cases} \frac{1}{b-a} \text{ si } x \in [a,b] \\ 0 \text{ sinon} \end{cases}$$

pour $a < b$. Tout d'abord :

1. Rappeler (ou calculer) la moyenne et la variance de cette distribution.

On prend pour la suite $a = 0$ et $b = 1$.

2. Exprimer l'aire sous la courbe de $p(x)$ pour $x \in [0, p]$ à l'aide d'une intégrale et donner sa valeur.
3. Expliquer comment générer des nombres suivant une loi de Bernoulli ($x = 0$ avec probabilité $1 - p$ et $x = 1$ avec probabilité p) en utilisant la loi uniforme.
4. Généraliser l'approche au cas suivant. On cherche à générer : $x = 0$ avec $p_0 = 0.1$, $x = 1$ avec $p_1 = 0.3$, $x = 2$ avec $p_3 = 0.4$ et $x = 4$ avec $p_4 = 0.2$, toujours en utilisant uniquement la loi uniforme.
5. Montrer à l'aide d'un histogramme ou en calculant les fréquences de chaque évènement que votre procédure fonctionne correctement (vous générerez $N = 10000$ échantillons).

On va maintenant étudier la loi de probabilité d'une loi uniforme transformée. A partir de deux variables aléatoires u et θ, chacune suivant la loi uniforme de paramètre $a = 0$ et $b = 1$, construisez les nouvelles variables suivantes :

$$z_0 = \sqrt{-2\ln(u)}\cos(2\pi\theta) \tag{1}$$
$$z_1 = \sqrt{-2\ln(u)}\sin(2\pi\theta) \tag{2}$$

6. En générant 2×10000 variables uniformes et en construisant les variables transformées, faites un histogramme des valeurs obtenues.
7. Déduire de cet histogramme la loi de ces variables et afficher en sur-impression de l'histogramme la densité de probabilité correspondante.

On va maintenant regarder une façon intéressant de calculer la valeur de π. On va donc générer des pairs de variables aléatoires indépendantes (x, y) où x et y sont distribuées selon la loi uniforme de paramètre $a = 0$ et $b = 1$. On pourra se représenter la densité $p(x, y)$ à l'aide d'un repère $(0, x, y)$, et imaginer qu'un tirage de pair revient à mettre un point dans le carré donné par les coordonnées $(0, 0)$, $(1, 0)$, $(1, 1)$ et $(0, 1)$ (voir la fig. 1). Commencençont par les préliminaires :

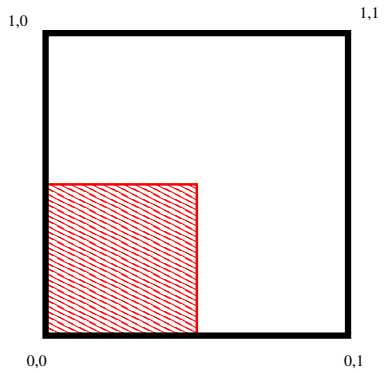

FIGURE 1 – .

8. Quelle est la probabilité que $0 < x < 1/2$ **ET** $0 < y < 1/2$? (donc que la pair (x, y) tombent dans la zone rouge hachurée)
9. On veut mettre en place un compteur qui sera incrémenté à chaque fois qu'une pair (x, y) tombe dans la zone rouge hachurée. Si je tire N pairs, quel sera le nombre moyen de pairs qui sera tombé dans cette zone ? Vérifier la réponse à l'aide d'une expérience numérique.

On va définir une nouvelle zone correspondant au cercle de centre $(1/2, 1/2)$ et de rayon $r = 1/2$.

10. Quelle est la probabilité qu'une pair (x, y) tombe dans le cercle ?
11. Si je tire N pairs, quel sera le nombre moyen de pairs qui tombera dans le cercle.
12. Mettez en place une procédure où vous aler tirer resepectivement $N = 10, 10^2, 10^3, 10^4, 10^5$ et 10^6 pairs. Pour chaque cas, vous calculerez la moyenne et l'erreur statistique sur la moyenne.
13. Montrer à l'aide d'un graphique, avec quelle allure la valeur moyenne estimée se rapproche de la vraie valeur moyenne.

2 SVD et le jeu de données MNIST

Pour cette partie vous allez travailler avec la base de données MNIST (déjà utilisée au TP4). Cette base d'images représente des chiffres manuscrits scannés, allant de 0 à 9 en nuance de gris : chaque pixel prend une valeur réelle $p_i \; in[0 : 1]$.

FIGURE 2 – Exemples d'images de la base MNIST

On cherchera ici à établir une procédure simple permettant d'identifier un chiffre d'un autre à l'aide de la SVD. On regardera donc les 4 et les 5.

2.1 Décomposition en SVD et identification des 5

A l'aide des lignes suivantes, récupérer tous les chiffres 4 dans un tableu et les 5 dans un autre.

```
X4 = train_set[0][np.where(train_set[1]==4)].T
X5 = train_set[0][np.where(train_set[1]==5)].T
print(X.shape)
```

Le travail est le suivant

1. Récupérer les $k = 10$ directions propres de l'ensemble des 4 d'une part, et des 5 d'autre part.
2. Pour une image de 4, montrer l'image obtenue si on utilise seulement $k = 10$ directions propres de l'ensemble des 4. Faire la même chose avec une image de 5.
3. Pour une image de 5, montrer l'image obtenue si on utilise seulement $k = 10$ directions propres de l'ensemble des 5. Faire la même chose avec une image de 4.
4. Calculer l'erreur de reconstruction d'un 4 et d'un 5 sur les directions propres des 4.
5. Calculer l'erreur de reconstruction d'un 4 et d'un 5 sur les directions propres des 5.
6. En déduire une façon de déduire directement d'une image si c'est un 4 ou un 5.
7. Effectuer une comparaison sur 1000 images pour $k = 10$ et $k = 50$.