Examen - Probabilités et statistiques

8 novembre 2020

Instructions

La solution complète doit être envoyée sous forme d'un seul fichier PDF avec le nom "Exam_PS_nom_prenom.pdf" à l'adresse giancarlo.fissore@inria.fr.

Date limite: dimanche 15 novembre, minuit.

Le PDF doit contenir le code complet et COMMENTÉ; les résultats numériques et les tracés/histogrammes/images doivent être visibles sur le fichier. Si vous utilisez Jupyter Notebook, vous pouvez facilement exporter tout le code et les tracés sous forme de fichier PDF; si vous utilisez d'autres outils, vous devez exporter les résultats de la même manière. Les exercices crayon et papier peuvent être numérisés et inclus dans le même fichier PDF (de nombreux utilitaires de fusion PDF sont disponibles en téléchargement ou en ligne) ou directement saisis sous forme numérique.

Les exercices Bonus ne sont pas obligatoires; ils donnent des points supplémentaires.

1 Computing Pi avec Montecarlo

Les algorithmes de Montecarlo utilisent l'échantillonnage de variables aléatoires pour calculer des quantités d'intérêt.

Partant de l'observation que le rapport de l'aire du cercle unitaire à l'aire du carré qui l'entoure est $\frac{\pi}{4}$, nous pouvons écrire un algorithme montecarlo pour calculer une valeur approximative pour π.

- Aire du cercle unitaire (cercle de rayon 1, centré à l'origine):

$$A_c = \pi r^2 \tag{1}$$

- Aire du carré entourant le cercle unitaire
  (carré avec sommets $(1, 1), (-1, 1), (-1, -1), (1, -1)$):

$$A_s = 4r^2 \tag{2}$$

- Ratio:
$$\frac{A_c}{A_s} = \frac{\pi}{4} \tag{3}$$

Algorithme:

- Échantillonner N valeurs de x entre -1 et 1, uniformément
- Échantillonner N valeurs de y entre -1 et 1, uniformément
- Calculer le rapport des aires comme le nombre de points (x, y) tombant dans le cercle unitaire divisé par les N points totals:

$$\frac{\text{\# de points à l'intérieur du cercle}}{N} \sim \frac{\pi}{4} \tag{4}$$

Exercice. Implémentez l'algorithme ci-dessus et affichez une valeur approximative pour Pi.

Remarque. Les points appartenant au cercle unitaire satisfont l'équation suivante:

$$x^2 + y^2 \leq 1 \tag{5}$$

Exercice bonus: tracez le carré, le cercle et les points échantillonnés (différentes couleurs pour les points à l'intérieur et à l'extérieur du cercle).

2 Distributions continues

2.1 Fonction de distribution cumulative (FDC)

La FDC d'une distribution de probabilité continue est défini comme

$$C(z) = \int_{-\infty}^{z} p(x)dx \tag{6}$$

Propriétés utiles de la FDC:

- $C(+\infty) = 1$ (découle de la normalisation)
- la fonction n'est pas décroissante

Plus d'informations sur la FDC: `https://fr.wikipedia.org/wiki/Fonction_de_r%C3%A9partition`

2.2 Méthode de la transformée inverse

Compte tenu des propriétés de la FDC, nous pouvons définir un algorithme pour échantillonner à partir de distributions de probabilités arbitraires en transformant d'échantillons uniformes.

Algorithme. Échantillonnage à partir de $p(x)$:

- Calculer la FDC de $p(x)$: $C(z)$
- Calculer l'inverse de la FDC: $F(x) = C^{-1}(x)$
- Échantillonnez un nombre aléatoire uniforme u dans l'intervalle $[0, 1]$
- Calculer un échantillon x distribué selon $p(x)$ en utilisant la formule: $x = F(u)$

Plus de détails: `https://fr.wikipedia.org/wiki/M%C3%A9thode_de_la_transform%C3%A9e_inverse`

2.3 La distribution exponentielle

Considérons la distribution exponentielle avec le paramètre λ:

$$p(x) = \frac{1}{\lambda} e^{-x/\lambda}, \qquad x \geq 0 \tag{7}$$

Exercices crayon et papier.

1. Calculez la moyenne μ et la variance σ^2 de la distribution.
2. Calculer la FDC correspondante; nous l'appelons $C(z)$.

   Rappelez-vous: $p(x)$ est défini pour des valeurs positives de x.
3. Calculez l'inverse de la FDC: $F(x) = C^{-1}(x)$
4. Fournissez une explication concise du fonctionnement de l'algorithme d'échantillonnage par transformation inverse.

Problèmes de codage. Définissez $\lambda = 3$.

5. Implémenter une fonction qui calcule des échantillons à partir de la distribution exponentielle à l'aide de l'algorithme d'échantillonnage par transformation inverse. Afficher un histogramme de la distribution.
6. Comparez votre fonction d'échantillonnage à la fonction *numpy* pour échantillonner des distributions exponentielles. Construisez 2 histogrammes et superposez la courbe théorique de $p(x)$.
7. Considérons la fonction $t(x) = x^2$. Supposons que la variable x soit distribuée selon une distribution exponentielle avec le paramètre $\lambda = 3$, calculez la moyenne m de t sur $N_s = 10000$ échantillons. Répétez le processus N_t fois et donnez une estimation de l'erreur. Quelle est la valeur minimale de N_t nécessaire pour obtenir une erreur plus petite que $\epsilon = 3 \cdot 10^{-3}$?

3 SVD et MNIST

Nous utiliserons le jeu de données MNIST.

1. Calculer la SVD de l'ensemble de données complet X
2. Visualisez les 10 premiers composants sous forme d'images.
3. Calculez les projections de tous les chiffres (0 à 9) sur les 3 premiers composants.
4. Produisez 2 nuages de points (scatter plots) dans les directions 1-2, 2-3. Choisissez une couleur différente pour chaque chiffre.

Bonus. Produire un nuage de points 3D dans les directions 1-2-3.