TME 2 - Estimation de densité

Données : velib

Télécharger l'archive jointe au sujet. Elle contient un mois de logs des stations velibs de Paris et les informations sur les stations. Le bout de code suivant permet de charger les données.

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.image as mpimg

fname = "velib.npz"
## fonction de deserialization de numpy
obj = np.load(fname)
## objets contenus dans le fichier
print(obj.keys())
## matrice 1217x#minutes nombre de velos disponibles
histo = obj['histo']
## matrice 1217x#minutes, pour chaque station nombre de velib pris a chaque minute
take = obj['take']
## infos stations statiques:
#### id_velib ->(nom,addresse,coord_y,coord_x,banking,bonus,nombre de places
stations = dict(obj['stations'].tolist())
## id_velib -> id matrice take, histo
idx_stations = dict(obj['idx_stations'].tolist())
stations_idx = dict(obj['stations_idx'].tolist()) ## id matrice -> velib
```

Les matrices `take` et `histo` contiennent respectivement le nombre de vélos pris et le nombre de vélos disponibles à une minute t depuis le 1er octobre 2016. Le dictionnaire `station` contient les informations de chaque stations. Le dictionnaire `idx_stations` permet de faire le lien entre une ligne i des matrices et l'identifiant correspondant de la station vélib (celui qui indexe le dictionnaire) : pour avoir les infos de la station de la première ligne : `stations[idx_stations[0]]`.

Visualisation

Le code suivant permet d'afficher la carte de Paris et les stations.

```
plt.ion()
parismap = mpimg.imread('paris-48.806-2.23--48.916-2.48.jpg')
## coordonnees GPS de la carte
xmin,xmax = 2.23,2.48   ## coord_x min et max
ymin,ymax = 48.806,48.916 ## coord_y min et max

def show_map():
    plt.imshow(parismap,extent=[xmin,xmax,ymin,ymax],aspect=1.5)
    ## extent pour controler l'echelle du plan
```

```
geo_mat = np.zeros((len(idx_stations),2))
for i,idx in idx_stations.items():
    geo_mat[i,0] =stations[idx][3]
    geo_mat[i,1]= stations[idx][2]
## alpha permet de regler la transparence
plt.scatter(geo_mat[:,0],geo_mat[:,1],alpha=0.3)
```

Expérimentations

L'objectif de ce TME est de prendre en main les algorithmes d'estimation de densité. Vous travaillerez successivement sur deux problèmes :

- Densité de l'offre statique des vélibs dans Paris : il s'agit d'estimer la densité $p(x,y)$ d'emplacements disponibles en fonction des coordonnées $(x,y) \in [xmin, xmax] \times [ymin, ymax]$, en ne prenant en compte que les informations statiques données par le dictionnaire `stations`.
- Densité de la demande à une station donnée en fonction du temps : il s'agit d'estimer la densité $p(t)$ de la probabilité qu'un vélo soit pris à une station en fonction de la minute t de la journée. Pour cela, vous pouvez extraire l'activité d'une seule journée (par exemple de 0 à 1440), ou des données agrégées de plusieurs journées (multiple de 1440).

Vous étudierez les deux algorithmes suivants :

- Méthode des histogrammes.
- Méthode à noyaux (Parzen pour commencer puis noyaux de votre choix).

Pour chacune des expériences, vous visualiserez les résultats obtenus en faisant une discrétisation de l'espace et en évaluant en chaque case la densité donnée par vos modèles (la méthode par histogramme vous donne directement la discrétisation à employer). Vous pouvez utiliser le code ci-dessous pour afficher votre discrétisation (dans le cadre de l'étude de l'offre statique) :

```
xx,yy = np.meshgrid(np.linspace(xmin,xma,steps),np.linspace(xmin,xmax,steps))
grid = np.c_[xx.ravel(),yy.ravel()]
res = monModele.predict(grid).reshape(steps,steps)
show_map()
plt.imshow(res,extent=[xmin,xmax,ymin,ymax],interpolation='none',\
           alpha=0.3,origin = "lower")
plt.colorbar()
plt.scatter(geo_mat[:,0],geo_mat[:,1],alpha=0.3)
```

Expérimentez selon différents réglages. Réflechissez en particulier aux questions suivantes :

- Que se passe-t-il pour une faible/forte discrétisation pour la méthode des histogrammes ?
- Peut-on se contenter d'une seule journée pour l'estimation de la demande ? Est-ce vrai sur toutes les stations ?
- Quel est le rôle des paramètres des méthodes à noyaux ?
- Observez vous des différences entre stations ? entre jours de semaines ?
- Comment choisir de manière automatique les meilleurs paramètres ?