CSC591/791: Spatial and Temporal Data Mining
Homework 2.
Released on: 02/03, Due on: 02/17, 11:45 pm

Student Names (including all group members):
Student IDs(including all group members):

Notes:

- Most questions are open-ended, so please craft an answer based on your reading/understanding.
- Some questions require reading papers and web resources. Likewise, you can discuss with fellow students. Whatever the case, answers should always be your own.
- You should always **cite** the source of all materials if they are used in your answer/solution/program. You may even cite urls.
- Answers should be precise and concise.
- PDF is the only accepted format. Embed all images/tables/graphs into your single pdf.
- Use Moodle for discussion in a way that does not violate academic integrity policies.

Q1. (30 points) For given satellite image (ilk-3b-1024.tif; 3-dimensional RGB, 1024x1024), do the following in RGB space:

(a) K-Means Clustering
(b) Model-based Clustering (GMM)
(c) Compare (a) and (b) on the following criteria:
    a. Finding optimal clusters
    b. Visualizing the clustered image (mark each cluster with different color) with the optimal number of clusters you have found
    c. Quality of results in terms of visual agreement with thematic classes (e.g., buildings, roads, waters, grass, trees, ...)

Hints:

- First, randomly sample (say 5% of) the image and use those samples to generate a model
- Use the model to predict cluster labels for all pixels in the image
- Change the color of pixels based on cluster assignment to visualize results
- You can use python/R and existing libraries
- You can read GeoTiff (and tons of other formats) images into R using "rgdal" and python using "gdal" libraries

Q2. (30 points) For given spatial dataset (cancer-data.csv), which contains x and y coordinates (in kilometers) with spatial resolution of 100 meters, do spatial clustering (DBSCAN). Through visual exploration, choose appropriate parameters for DBSCAN clustering. Answer the following:

- Report parameters
- Generate plot for raw data
- Generate plot for cluster results (mark each cluster with different color; mark noise points with separate color or symbol)

Hints:

- You can use any freely available implementations, but DBSCAN clustering in R (in "fpc" package) may be the easiest one.

Q3. (20 points) Compare and contrast C-DBSCAN algorithm [1] and DBCluC algorithm [2]. In addition to similarities and dissimilarities, comment on computational complexity, ease of implementation, ease of use. For each algorithm, list one spatial application that is best suited for the algorithm.

Q4. (20 points) Take a neutral image of your choice (please nothing in poor taste such as political or religious) and perform edge detection by *convolving* (yes, the exact same way Convolutional Neural Nets work) the "Sobel" operator through the image. This is a part of a data scientist's process called feature engineering and these representations are often used as inputs to a subsequent learning task. Show us:

- Original image
- X-gradient image (vertical edge detection)
- Y-gradient image (horizontal edge detection)

References:

[1] Carlos Ruiz, et al. C-DBSCAN: Density-Based Clustering with Constraints.
[2] Osmar R. Zaiane, et al. Clustering Spatial Data in the Presence of Obstacles: a Density-Based Approach.