DSC 495: Project

Some instructions:

- Your final report should not just be in the form of code, but should also contain a description of what you are doing. Here is a suggested outline:
  - Introduction/Motivation/Brief description of problem
  - Methodology
  - Implementation
  - Results and testing
- Show through extensive testing that your implementations are producing sane outputs.
- The project must be submitted on your github repository as `.ipynb` files (Jupyter notebooks) in the `project` folder; see the problem set for instructions. The `.ipynb` files should have all the outputs.
- All the plots must be clearly titled and labeled.
- The notebooks should be self-contained. The instructor will not execute the code unless to check for correctness.

Logistics Due dates:

- Project choice: November 14.
- Final project: December 5.

Grading rubric:

Category	Points
Documentation/Methodology	20
Implementation	40
Testing	30
Comments/References	10

Here are some ideas for projects. They are only suggested guidelines and not hard-and-fast rules. The tasks listed below should be incorporated in the report.

1. **X-ray reconstructions** The recovery of images from data collected using x-ray devices can be obtained by solving a least squares problem of the form

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2,$$

where x denotes the grayscale image in vectorized form, $A \in \mathbb{R}^{m \times n}$ is a matrix which represents the physics of the x-ray machine, and b represents the data collected by the machine. Your tasks are the following:

   (a) Download the mat files from the Finnish Inverse Problems society webpage. You can choose from: lotus root, walnut, or carved cheese datasets. The mat files should contain the data b and the matrix A and a ground truth for comparison.

(b) Solve the least squares problem using Conjugate Gradient on the normal form

$$(A^T A + \lambda^2 I)x = A^T b.$$

You can use the conjugate gradient solver from `scipy.sparse.linalg.cg`. Here I is the $n \times n$ sparse identity matrix.

(c) Solve the least squares problem for different values of λ. The values of lambda should span several orders of magnitude (e.g., $10^{-3} - 10^{1}$). You can use the ground truth for each problem to verify the performance.

2. **Clustering** Clustering is an important unsupervised learning task in machine learning, to organize the data points into one or more clusters. Here are some suggested tasks

   (a) Implement any algorithm of your choice for clustering. Explain this method in your words.

   (b) Compare the algorithms for 1-2 datasets (e.g., Iris species, MNIST handwriting database). Use metrics such as a confusion matrix to quantify clustering accuracy.

   (c) Experiment with algorithmic parameters.

   Datasets and Implementations of other algorithms can be found in scikit-learn website.

3. **Visualizing Earth's temperature** NOAA's web page has detailed historical records of earth's temperature from 1880-present. In this project, you will develop an interactive jupyter notebook to visualize this dataset. Create an interactive Jupyter notebook for visualizing the spatiotemporal temperature distribution. Here are some ideas for tasks; you can look at https://www.ncdc.noaa.gov/cag/ for inspiration.

   (a) Drop down menus that plots the time series of temperature history that gives the user a choice from 10 different cities and then let's them select the time frame (e.g., annually, monthly, for a given range);

   (b) Plot the global temperature for a given point in time (use `geopandas` or `basemap`);

   (c) Plot the global mean temperature (annual, monthly, etc).

   You can use `netCDF4` for loading the dataset; note that the data has missing entries which you will have to handle appropriately. Interactive Jupyter notebooks can be constructed using Jupyter widgets.

4. Wikipedia is a common data source for language processing tasks. These data sets can be pretty large, so we can choose to work with this dataset which only includes the article title and summary. Download the raw dataset.

   (a) Determine which words and adjacent pairs of words are most common in the wikipedia summaries.

   (b) Visualize the collection of words from a few sample article summaries using the wordcloud package.

   (c) For each article summary construct a list of words in the article, with the associated number of occurrences of each word. Then given a fixed "reference" article, find the five other articles which have the most similar word distribution.

   (d) Note: there are multiple ways to compare distributions of words, you can pick what you think is best. Also, you may want to exclude certain common words, such as "is", "and", etc. The ".vocab" file in the dataset may aid in this.

5. **Your own project** Propose your own project. Ideally, this should be related to your research. Send the instructor an email including the following details: what you are planning to implement, what resources (e.g., libraries, datasets) you will use. The expectations for the content and the report should be in line with one of the project listed above. You should have the instructor's written approval before submitting the project. Feel free to make an appointment with the instructor to discuss the project proposal.