Homework 4

AMATH 582/482, Winter 2022

Assigned Feb 25, 2022. Due on Mar 11, 2022 at midnight.

Directions, Reminders and Policies

Read these instructions carefully:

- **You are required to upload a PDF report to Canvas along with a zip of your code. Note the PDF and the zip should be uploaded separately.**
- The report should be a maximum of 6 pages long with references included. Minimum font size 10pts and margins of at least 1inch on A4 or standard letter size paper.
- Do not include your code in the report. Simply create a zip file of your main scripts and functions, without figures or data sets included, and upload the zip file to Canvas.
- Your report should be formatted as follows:
  - Title/author/abstract: Title, author/address lines, and short (100 words or less) abstract. This is not meant to be a separate title page.
  - Sec. 1. Introduction and Overview
  - Sec. 2. Theoretical Background
  - Sec. 3. Algorithm Implementation and Development
  - Sec. 4. Computational Results
  - Sec. 5. Summary and Conclusions
  - Acknowledgments (no more than four or five lines, also see the point below on collaborations)
  - References
- I suggest you use LaTeX(Overleaf is a great option) to prepare your reports. A template is provided on Canvas under the Syllabus tab. You are also welcome to use Microsoft Word or any other software that properly typesets mathematical equations.
- I encourage collaborations, however, everything that is handed in (both your report and your code) should be your work. You are welcome to discuss your assignments with your peers and seek their advice but these should be clearly stated in the acknowledgments section of your reports. This also includes any significant help or suggestions from the TAs or any other faculty in the university. You don't need to give all the details of the help you received, just a sentence or two.
- Your homework will be graded based on how completely you solved it as well as neatness and little things like: did you label your graphs and include figure captions. **The homework is worth 20 points. 10 points will be given for the overall layout, correctness and neatness of the report, and 10 additional points will be for specific technical things that the TAs will look for in the report itself.**
- **Late submissions will not be accepted on Canvas, send them to bamdadh@uw.edu directly. Late reports are subject to a 2 points/day penalty up to five days. They are no longer accepted afterwards. For example, if your report is three days late and you managed to get $16/20$, your final grade will be $16 - 6 = 10$.** Basically, you will lose 2% of your overall course grade for each day the report is late. So be careful.

Problem Description: Classifying Politicians

Your goal is to test the performance of spectral clustering and a simple semi-supervised regression algorithm on the 1984 house voting records data set. Download the data set called `house-votes-84.data`, and the description `house-votes-84.names`. The data set consists of voting records of 435 members of the House on 16 bills. There are 267 members of the democratic party and 168 members of the republican party. The voting record of each house member on the 16 bills will be our input $\mathbf{x}$ while the corresponding output/class y is that members party affiliation (republican or democrat embedded as ± 1).

Tasks

Below is a list of tasks to complete in this homework and discuss in your report.

1. Your first task is to import and preprocess the data set. Construct your output vector $\mathbf{y}$ by assigning labels $\{-1, +1\}$ to members of different parties. Then construct the input vectors $\mathbf{x}_j$ corresponding to the voting records of each member by replacing `'y'` votes with $+1$, `'n'` votes with -1 and `'?'` with 0. You do not need to center and normalize the data set in this case. This leads to a vector $\mathbf{y} \in \mathbb{R}^{435}$ and input matrix $X \in \mathbb{R}^{435\times 16}$. Note that we are using the `sklearn` convention for our X matrix.

2. **(Spectral Clustering)**: In this step you will mainly work with the matrix X and use $\mathbf{y}$ for validation of your clustering algorithm.

   (a) Construct the unnormalized graph Laplacian matrix on X using the weight function

   $$\eta(t) = \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

   with variance parameter σ and compute its second eigenvector (i.e., the Fiedler vector) which we denote as $\mathbf{q}_1$.

   (b) Take $\text{sign}(\mathbf{q}_1)$ as your classifier and compute its classification accuracy after comparison with $\mathbf{y}$:

   $$\text{clustering accuracy} = \frac{1}{435} \times \text{number of misclassified members.}$$

   <u>Note 1:</u> since this is an unsupervised learning task, your classifier $\text{sign}(\mathbf{q}_1)$ might disagree with $\mathbf{y}$ by a negative sign, that is, the $+1$ class may be assigned to -1 and vice versa. Don't forget to adjust for this.

   <u>Note 2:</u> "misclassified" here means a member who should have belonged to the opposite party/cluster.

   (c) Change the parameter σ in the range $(0, 4]$ and plot accuracy as a function of σ. Let σ^* denote the optimal variance parameter achieving maximum clustering accuracy. Discuss your findings.

3. **(Semi-supervised Learning)**: Now consider the unnormalized Laplacian from Step 2 with the optimal parameter σ^* that you picked in step 2c.

   (a) Given an integer $M \geq 1$ consider the Laplacian embedding

   $$F(\mathbf{x}_j) = \Big((\mathbf{q}_0)_j, (\mathbf{q}_1)_j, \ldots, (\mathbf{q}_{M-1})_j\Big) \in \mathbb{R}^M,$$

   where $\mathbf{q}_j$ denote the eigenvectors of the Laplacian matrix. Write $F(X) \in \mathbb{R}^{435\times M}$ for the Laplacian embedding of X, i.e., the matrix whose j-th row is $F(\mathbf{x}_j)$.

(b) Given an integer $J \geq 1$ consider the submatrix $A \in \mathbb{R}^{J \times M}$ and vector $\mathbf{b} \in \mathbb{R}^J$ consisting of the first J rows of $F(X)$ and $\mathbf{y}$,

$$\begin{aligned} A_{ij} &= F(X)_{ij}, \quad i = 0, \ldots, J-1, \quad j = 0, \ldots, M-1, \\ \mathbf{b}_i &= \mathbf{y}_i, \qquad\quad\;\; i = 0, \ldots, J-1. \end{aligned}$$

Warning: It is crucial here to use the original ordering of the data set. You should not be shuffling or re-ordering the rows of X.

(c) Use linear regression (least squares) to find

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^M} \|A\boldsymbol{\beta} - \mathbf{b}\|_2^2,$$

then take $\hat{\mathbf{y}} = \operatorname{sign}(F(X)\hat{\boldsymbol{\beta}})$ as the predictor of classes of all points in X.

Note: You do not need to add the column of 1's in A since $\mathbf{q}_0$ is already a constant vector so that β_0 accounts for the intercept.

(d) Provide a table summarizing the accuracy of $\hat{\mathbf{y}}$ as your classifier for $M = 2, 3, 4, 5, 6$ and $J = 5, 10, 20, 40$.

$$\text{SSL accuracy} = \frac{1}{435} \times \text{ number of misclassified memebrs.}$$

Discuss your findings.

Some comments and hints

Here are some pointers.

1. In Step 2, it is helpful to plot $\mathbf{q}_1$ by re-ordering X according to the original party affiliations. For example, plot the $\mathbf{q}_1$ values for the 168 republicans first then the 267 democrats. This will nicely visualize the behavior of $\mathbf{q}_1$ on the two clusters.
2. The above visualization trick is also helpful in Step 3, to visualize $F(X)\hat{\boldsymbol{\beta}}$ and the classifier $\hat{\mathbf{y}}$.