AMATH 582: HOME WORK 4

RAPHAEL LIU

Department of Appliede Mathematics, University of Washington, Seattle, WA
raph651@uw.edu

Abstract. In this assignment, a data set of house voting records in 1984 United States is given to us. As a data scientist, our task is to classify the party affiliation (Republican or Democrat) using different strategies: Spectral Clustering and Semi-Supervised Learning.

1. Introduction and Overview

The house voting records data set contains voting records of 435 members of the House on 16 bills. The 1 output feature is the class affiliation: Republican or Democrat. There are 267 members of the democratic party and 168 members of the republican party. The 16 input features are our bills. All the features are Boolean valued, either 'y' or 'n', except some attributes are missing and denoted as '?'. Our goal in this homework is to fit our voting features data using 2 different methods: Spectral Clustering and Semi-Supervised Learning. To optimize our models and classification, we need to find the optimal parameters accordingly. Below is a list of tasks to complete.

(1) Import and preprocess the data set. Construct the output vector y by assigning labels {-1, +1} to members of different parties. Then construct the input vectors x corresponding to the voting records of each member by replacing 'y' votes with +1, 'n' votes with -1 and '?' with 0.

(2) Clustering Algorithm: Construct the unnormalized graph Laplacian matrix on X using the weight function $\eta(t) = exp(-\frac{t^2}{2\sigma^2})$ with variance parameter σ and compute its second eigenvector (i.e., the Fiedler vector) which denoted as q1. Take sign(q_1) as classifier and compute its classification accuracy after comparison with y:

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

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 the findings.

(3) Semi-Supervised Learning: Now consider the unnormalized Laplacian with optimal parameter σ^*. Given an integer $M \geq 1$ consider the Laplacian embedding

$$F(x_j) = ((q_0)_j, (q_1)_j, ..., (q_{M-1})_j) \in \mathbb{R}^M,$$

where q_j denote the eigenvecotrs of the Laplacian matrix. Write $F(X) \in \mathbb{R}^{435 \times M}$ for the Laplacian embedding of X, ie., the matrix whose j-th row is $F(x_j)$.

Date: March 11 2022.

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 $\mathbf{y}$,

$$A_{ij} = F(X)_{ij}, \quad i = 0, ..., J-1 \quad j = 0, ...M-1$$
$$\mathbf{b}_i = \mathbf{y}_i, \quad 0, ..., J-1$$

Use linear regression (least squares) to find

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^M}{\arg\min} ||A\beta - \mathbf{b}||_2^2$$

then take $\hat{\mathbf{y}} = \text{sign}(F(X)\hat{\beta})$ as the predictor of classes of all points in X. Provide a table summarizing the accuracy of $\hat{\mathbf{y}}$ as your classifier for $M = 2, 3, 4, 5, 6$ and $J = 5, 10, 20, 40$.

$$\text{SSL accuracy} = 1 - \frac{1}{435} \times \text{number of missclassified members.}$$

Discuss the findings.

2. Theoretical Background

2.1. **Spectral Clustering.**

Clustering: The unsupervised task of dividing the input data X into meaningful groups. Or, say the task of finding meaningful structures in X where we have no labels.

Split X into K clusters, $X = \bigcup_{k=0}^{K-1} C_k \quad \text{s.t. } C_k \bigcap C_l = \emptyset$

Similarity Graphs: Consider our data set $X = x_1, ..., x_{N-1} \in \mathbb{R}^d$ & asymmetric matrix $W \in \mathbb{R}^{N \times N}$ with non-negative entries $w_{ij} \geq 0$.

We can then define a weighted undirected graph $G = X, W$, where the $X_j \in \mathbb{R}^d$ are the vertices of G & the entries w_{ij} of W denote weights that are associated to edges that connect x_i to x_j.

- If $w_{ij} = 0$ for some i, j then x_i & x_j are not connected by an edge
- Since W is symmetric then $w_{ij} = w_{ji}$, so the edge is not directed
- The value of w_{ij} represents the strength of the connection

Now we consider a particular family of weighted graphs called **Proximity Graphs**: Let η be a non-negative, non-increasing & continuous function (weight function), we then take $w_{ij} = \eta(||x_i - x_j||_p)$ for $1 \leq p \leq \infty$

A typical example is simply to choose $\eta = exp(-\frac{t^2}{2\sigma^2})$ & $p = 2$ which leads to $w_{ij} = exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2})$

Another popular choice for η is $\eta(t) = \begin{cases} 0 & \text{if } t \geq r \\ 1 & \text{if } t < r \end{cases}$

- Define the degree vector $d \in \mathbb{R}^N$, $d_j = \sum_{i=0}^{N-1} w_{ji}$, sum of rows of W.
- Define the diagonal degree matrix $D = \text{diag}(d)$
- Define the (Unnormalized) graph Laplacian $\tilde{L} = D - W$
- As well as the normalized graph Laplacian $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$

2.2. **Semi-Supervised Learning.**

For semi-supervised learning (SSL), we have labeled data (sub)set and the unlabeled data (sub)set. Here we will focus on the particular case of SSL with graph Laplacian regularization (aka manifold regularization.)

Consider we have inputs $\mathbb{R}^{d\times N} \to X = \{x_0, x_1, ..., x_{N-1}\}$, along with **some** outputs, $\mathbb{R}^M \to y = \{y_0(x_0), y_1(x_1), ..., y_{M-1}(x_{M-1})\}$. We call the pairs $\{(x_0, y_0), ..., (x_{M-1}, y_{M-1})\}$ the labeled date (sub)set and the set $\{x_M, ...x_{N-1}\}$ the unlabeled data (sub)set.

Semi-Supervised Regression (SSR) is a particular case of SSL where the output of $y(x)$ are real valued then $y = (y_0(x_0), ..., y_{m-1}(x_{m-1})) \in \mathbb{R}^M$. We proceed analogously to Kernel regression, ie, we wish to find a function $f(x) = \sum_{j=0}^{J} c_j F_j(x)$ so that $f(x_j) \approx y(x_j)$ for $M \le j \le N-1$. ie, we want f to approximate the output only on the unlabeled set. The feature map F_j is picked using Graph Laplacian embedding.

$$F(x_j) = ((q_0)_j, (q_1)_j, ..., (q_{M-1})_j) \in \mathbb{R}^M,$$

where q_j denote the eigenvecotrs of the Laplacian matrix. Write $F(X) \in \mathbb{R}^{435\times M}$ for the Laplacian embedding of X, ie., the matrix whose j-th row is $F(x_j)$.
Given an integer $J \ge 1$ 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 $\mathbf{y}$,

$$A_{ij} = F(X)_{ij}, \quad i = 0, ..., J-1 \quad j = 0, ...M-1$$
$$\mathbf{b}_i = \mathbf{y}_i, \quad 0, ..., J-1$$

Finally solve the linear regression problem:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^M}{\arg\min} ||A\beta - \mathbf{b}||_2^2$$

3. Algorithm Implementation and Development

The programming language used here is Python 3. The procedure for solving the tasks are given below, separately.

(1) First preprocess the data set. Label different party affiliation to {-1, +1} to construct vector $\mathbf{y}$. Then construct the input vectors $\mathbf{x_j}$ corresponding to the voting records, labeling the response to {-1, +1, 0}.

(2) Define a Gaussian weight function with parameter σ to be tuned. Start with $\sigma = 1$. Generate the similarity graph W. Summing the rows to get d and then define the diagonal degree matrix $D = \text{diag}(d)$. Construct graph Laplacain $L = D - W$. Decompose L to get eigenvectors and eigenvalues. Sort the eigenvalues with eigenvectors. Use the second eigenvector q_1 as our Laplacian embedding.

(3) The classifier is taken to be the sign of Laplacian embedding. Record the miss-classified number of instances, and calculate accuracy with comparision with y. Since the classifier doesn't tell which is republican or democrat, the score might be 1-score in cases, whichever higher.

(4) Now sweep parameter σ on interval (0,4] with 20 different data points. Same as the previous step, compute and record the accuracy. Note the optimal parameter σ where maximum accuracy is reached. Denote this value as σ^*. Plot accuracy as a function of σ.

(5) To visualize the data, use the optimal σ^* and regenerate the Laplacian embedding $F(X)$. Re-ordered X with 168 republican first then the 267 democrats. Plot the scatter graph of the embedding and the re-ordered embedding. Note the behavior of clustering.

(6) Moving to semi-supervised learning, make array of M=[2,3,4,5,6] and J=[5,10,20,40]. Iterate each pair of M,J-the Laplacian embedding $F(X)$ is taken to be the first M eigenvectors, the A matrix is taken be the first J rows of $F(X)$, as well as $\mathbf{b}$ be the first J rows of $\mathbf{y}$. Use sklearn's linear regression model to fit and solve for β. Act the Laplacian embedding $F(X)$ on β to get the classification results. Compute and record the SSL accuracy. Record the maximum accuracy and the optimal pair of (M^*, J^*).

(7) To visualize the data, use the optimal σ^* and (M^*, J^*) to compute Laplacian embedding $F(X)$. Plot the scatter graph of $F(X)$, re-ordered $F(X)$, and re-ordered classifier $\hat{\mathbf{y}}$. Note the behavior of clustering.

4. Computational Results

The data set is preprocessed and labeled correspondingly. For the next part, $\sigma = 1$ is set to be the starting point. Apply spectral clustering and computation of accuracy. A accuracy of 0.87816 is recorded. Then iterate the parameter σ on the interval (0,4]. This results in an optimal $\sigma^* = 2.400$ and maximum clustering accuracy 0.88276. A plot of accuracy function of σ is shown below: 1

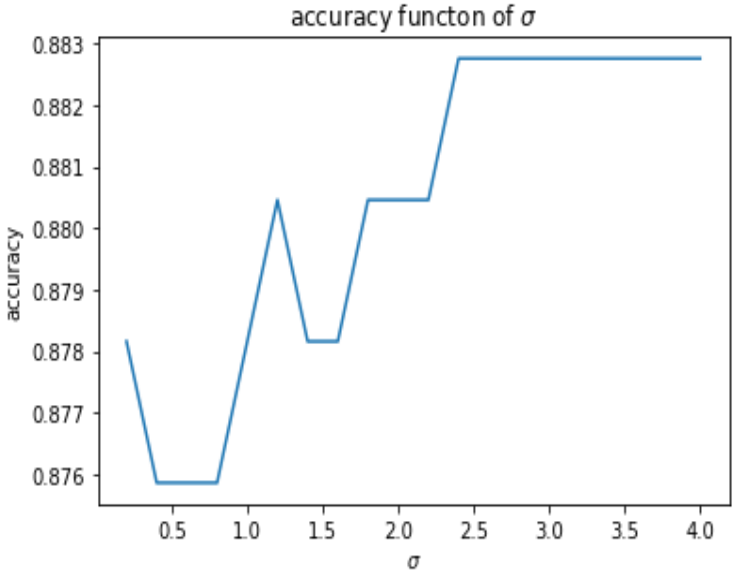

Figure 1. Plot of clustering accuracy over $\sigma \in (0, 4]$

Now we choose the parameter σ^* and plot the scatter graph of Laplacian embedding, along with the re-ordered one. 2The re-orderd Laplacian embedding show an obvious pattern for clustering: the blue dots tend to stay positive-valued, while the orange dots tend to stay negative-valued.

The Semi-Supervised Learning part first requires us to define pairs of parameter M,J. Next, iterate over each pair of them and record the corresponding accuracy.1 The maximum SSL accuracy is found to be 0.85977 at M=6, J=10.

Finally, the visualization is accomplished by setting M,J,σ to be the optimal and plotting the classification (Laplacian embedding acting on $\mathbf{b}$), re-ordered classification, and re-ordered classifier.3After re-ordering, the scatter pattern shows a clear clustering pattern that the two clusters tend to stay opposite-valued. The re-ordered classifier shows that even though some points are miss-classified, the majority is split into the correct clusters.

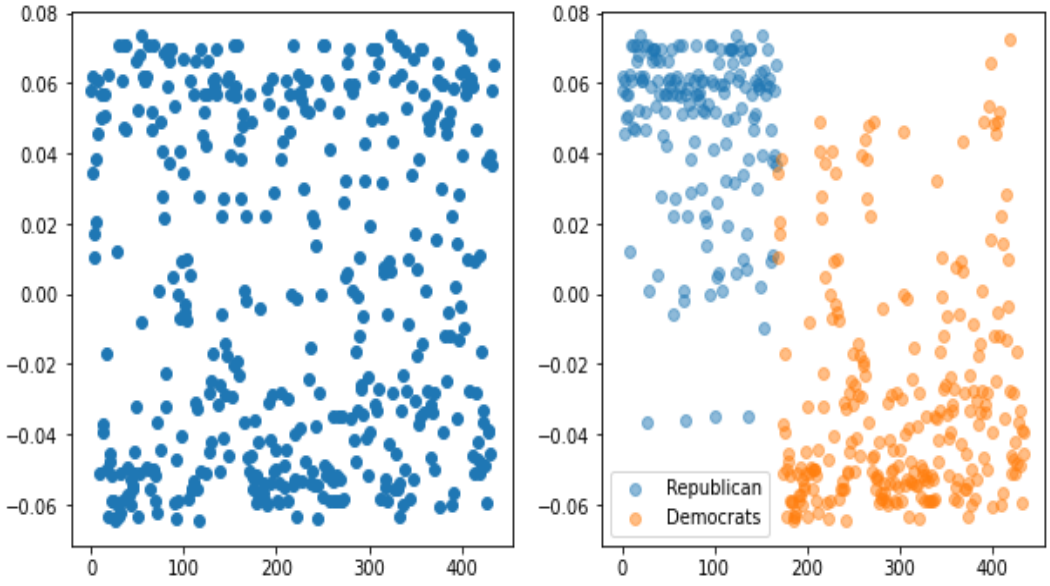

FIGURE 2. Scatter graph of Laplacian embedding (left) and re-ordered Laplacian embedding (right)

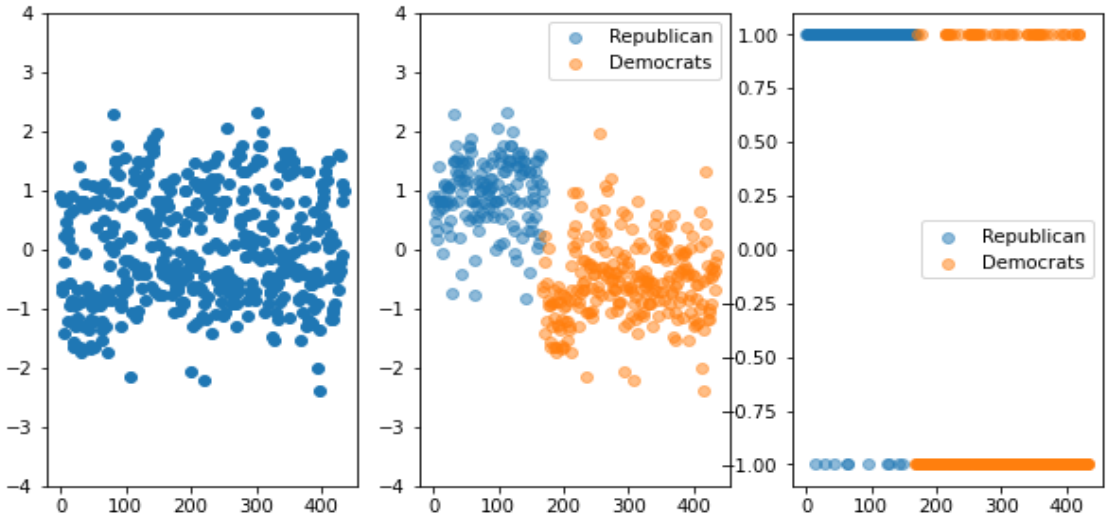

FIGURE 3. Scatter graph of classification (left), re-ordered classification (middle), and re-ordered classifier (right)

Accuracy	J=5	J=10	J=20	J=40
M=2	0.61379	0.61379	0.61379	0.61379
M=3	0.61379	0.61379	0.61379	0.61379
M=4	0.61379	0.61379	0.61379	0.61379
M=5	0.80690	0.61379	0.61379	0.61379
M=6	0.80920	0.85977	0.61379	0.61379

TABLE 1. Prediction of scores by each of the three models

5. Summary and Conclusions

In order to make promising classification for the house voting record data, there are few steps most importanly to obtain the results. These steps include preprocessing data set, tuning parameter for gaussian weight function, spectral clustering, tuning parameters for number of eigenvectors to keep and number of labeled points for semi-supervised learning, and last conducting SSR.

Acknowledgements

The author is thankful to Prof. Hosseini for detailed discussion about spectral clustering and semi-supervised learning.