AMSC808N/CMSC828V Maria Cameron

Project 2. Due Thursday, Nov. 5

The goal of this project is to practice techniques for data clustering, matrix factorization, and matrix completion on real datasets. This is a group project as Project 1. *Each student should develop her/his own codes. One report per group should be submitted. Each student's codes should be linked to the pdf file with the report.*

Dataset 1: an incomplete spreadsheet of movie ratings. Data file: `MovieRankings36`. Format: CSV (can be opened and edited e.g. using Numbers (Mac OS), Excel (Windows)). Columns correspond to the following movies:

1. "Home Alone";
2. "The Lion King";
3. "The Princess Bride";
4. "Titanic";
5. "Beauty and the Beast";
6. "Cinderella";
7. "Shrek";
8. "Forrest Gump";
9. "Aladdin";
10. "Ferris Bueller's Day Off";
11. "Finding Nemo";
12. "Harry Potter and the Sorcerer's Stone";
13. "Back to the Future";
14. "UP";
15. "The Breakfast Club";
16. "The Truman Show";
17. "Avengers: Endgame";
18. "The Incredibles";
19. "Coraline";
20. "Elf"';

Rows correspond to users (my family, my family friends, some of you). Scores range from 1 (bad) to 5 (great). **Add your row there** (it is fine to make a repeated row and it is important to know which row is yours) and also feel free to add more rows with your friends' ratings.

Dataset 2: The collection of 139 documents on two topics: `US:Indiana:Evansville` (id:10567) and `US:Florida` (id:11346) . Each document is a "bag of words" described

with respect to a dictionary of 18446 words. This dataset is downloaded from TechTC - Technion Repository of Text Categorization Datasets. Go down to *Availability and usage*. Read under *Preprocessed feature vectors* right above it. Then click on TechTC-300, and download `Preprocessed feature vectors: techtc300_preprocessed.zip`. It contains many folders. I suggest to play with the data from `Exp_10567_11346`. For your convenience, I prepared a data file `vectors.txt` by removing extra symbols from `vectors.dat` and wrote a Matlab function `readdata.m` that reads it and outputs the matrix M (M_{ij} is the count of word j in document i) and the ground truth classification vector y ($y_i = -1$: Indiana, $y = 1$: Florida). You can pick another folder if you wish.

Programming: pick any language you wish. High-level language is preferable. All requested algorithms should be programmed from scratch. SVD (with option 'econ') should be computed using standard SVD solver (command `svd` in Matlab).

Problems

1. **Dataset 1: movie rankings. Task: NMF.** Select a complete submatrix of the movie ranking matrix. Choose a reasonable number of clusters k and use the k-means algorithm to cluster the users. Try to interpret the result. For the same k, compute the NMF $A \approx WH$ where W has k columns, using

   (a) Projected gradient descend.

   (b) Lee-Seung scheme.

   (c) Start with projected gradient descend, continue with Lee-Seung.

   Plot the Frobenius norm squared vs iteration number for each solver. Which one do you find to be the most efficient?

2. **Dataset 1: movie ranking. Task: matrix completion.** Do matrix completion in two ways:

   (a) Low-rank factorization. Compute it using alternating iteration (see my lecture notes or Bindel's Lecture 8, Section 2). Experiment with different values of λ. Which λ gives the most reasonable result?

   (b) Nuclear norm trick and the iteration $M^{j+1} = S_\lambda(M^j + P_\Omega(A - M^j))$ (Bindel's Lecture 8, Section 3). Experiment with different values of λ.

   In both cases, experiment with different values of λ. Which λ gives the most reasonable result? Compare these two approaches for matrix completion. Which one gives more sensible results? Which one is easier to use? Which one do you find more efficient?

3. **Dataset 2: text documents. Task: CUR factorization.** Program the CUR algorithm as described in M. Mahoney and P. Drineas, *CUR matrix decompositions for improved data analysis*, PNAS, vol. 106, no. 3, 697–702. Run it for k from 2 to 10 and for $c = r = ak$ for $a = 1, 2, \ldots, 8$. Since it is a randomized algorithm, perform 100 runs for each combination of k and a. Plot the mean ratio $\|M - CUR\|_F / \|M - M_k\|_F$ versus k for each a. Also plot $\|M - M_k\|_F$ versus k for each a. Pick a reasonable combination of k and a.

4. **Dataset 2: text documents. Task: text categorization.** Pick a reasonable combination of k and a. Look up the k words with the maximal leverage (see `features_idx.txt`). What are they? Are they suitable for classification of these docs to Florida and Indiana? If not, invent a criterion for selecting columns of M (columns correspond to words), select 10,000 columns of M, and make two plots:

   - Compute SVD of the new M and plot the data projected onto first two principal components (first two right singular vectors) coloring the according to the ground truth class.
   - Select 5 columns of the new M with the maximal leverage scores. Compute the SVD for this subset of columns and plot the dat projected onto the first two principal components. Color them according to the ground truth.

   If the data from Florida and Indiana are not separable, modify your column selection criterion and repeat. Write summary of your findings.