MATH 547/ BIOINF 547: Mathematics of Data

Date: April 19, 2019

Due date: April 28, 2019

Problem Set 4: The goal of the following problems is to refresh your memory about PCA, MDS and SVD.

1. **Principal component analysis (PCA)**: Take any digit data ('0',...,'9'), or all of them, from this website: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.digits/` (data info here, full gzipped data download here), and perform PCA.
   (a) Set up data matrix Set up data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^n$ and compute the sample mean μ_n and form $\mathbf{Y} = \mathbf{X} - e\mu_n^T$.
   (b) Compute top k SVD of $\mathbf{Y} = \mathbf{USV}^T$.
   (c) Plot eigenvalue curve, i.e. i vs. $\frac{\lambda_i(\mathbf{M})}{\sum \lambda_i(\mathbf{M})}$ $(i = 1, ..., k)$,with top$-k$ eigenvalue λ_i for sample covariance matrix $\mathbf{M} = \frac{1}{n}\mathbf{YY}^T$, which gives you explained variation of data by principal components.
   (d) Use *imshow* to visualize the mean and top$-k$ principle components as *left* singular vectors $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_k]$.
   (e) For $k = 1$, sort the image data $(\mathbf{x}_i)(i = 1, ..., n)$ according to the top right singular vectors, $\mathbf{v}_1$, in an ascending order. For $k = 2$, make a scatter plot $(\mathbf{v}_1, \mathbf{v}_2)$ and show those images on a grid in such a plane (e.g. Figure 14.23 in book Elements of Statistical Learning [1]).

2. **Multidimensional scaling (MDS) of cities**: Visit the following website to perform the following exercise. `http://geobytes.com/citydistancetool/`
   (a) Input a few cities (no less than 7), and collect the pairwise air traveling distances shown on the website into a matrix $\mathbf{D}$.
   (b) Make your own codes for the MDS algorithm for $\mathbf{D}$;
   (c) Plot the normalized eigenvalues $\frac{\lambda_i}{\sum \lambda_i}$ in a descending order of magnitudes, analyze your observations (did you see any negative eigenvalues? if yes, why?).
   (d) Make a scatter plot of those cities using top 2 or 3 eigenvectors, and analyze your observations.

3. **Singular Value Decomposition (SVD)**: One of the best references for the SVD is Chapter 2 in the book Matrix Computations (Golub and Van Loan, 3rd edition [2]).
   (a) **Existence**: Prove the existence of the SVD. That is, show that if $\mathbf{A}$ is an $m \times n$ real valued matrix, then $\mathbf{A} = \mathbf{USV}^T$, where $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{V}$ is an $n \times n$ orthogonal matrix, and $\mathbf{S} = \text{diag}(\sigma_1, \sigma_1, ..., \sigma_p)$ (where $p = min\{m, n\}$) is an $m \times n$ diagonal matrix. It is conventional to order the singular values in decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Determine to what extent the SVD is unique. (See Theorem 2.5.2, page 70 in Golub and Van Loan).
   (b) **Best rank$-k$ approximation - Frobenius norm**: Show that the SVD also provides the best rank$-k$ approximation for the Frobenius norm, that is, $\mathbf{A}_k = \mathbf{US}_k\mathbf{V}^T$ satisfies

$$\|\mathbf{A} - \mathbf{A}_k\|_F = \min_{rank(B)=k} \|\mathbf{A} - \mathbf{B}\|_F$$

References

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[2] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.