Assignment 8

CS-UH-2218: Algorithmic Foundations of Data Science

Assignments are to be submitted in groups of two or three*. Upload the solutions on NYU classes as one PDF file for theoretical assignments and separate source code files for each programming assignment. Submit only one copy per group. Clearly mention the participant names on each file you submit.*

Problem 1 (10 points)**.**
We had used the following simple result in the derivation of SVD: If A and B are two matrices s.t. $A\mathbf{v} = B\mathbf{v}$ for any vector v then $A = B$. Prove this result.
Hint: If $A \neq B$ show that there exist vectors $\mathbf{u}$ and $\mathbf{v}$ s.t. $\mathbf{u}^T A\mathbf{v} \neq \mathbf{u}^T B\mathbf{v}$.

Problem 2 (10 points)**.**
In the greedy algorithm we presented in class for computing best fit subspaces, we assumed that the best fit $(k+1)$-dimensional subspace for a data set contains the best fit k-dimensional subspace for the data set. Prove that this assumption is valid.
Hint: Let $\mathbf{v}_1, \cdots, \mathbf{v}_k$ be an orthonormal basis for the best fit k-dimensional subspace V. Let W be the best fit $(k+1)$-dimensional subspace. Show that it is possible to choose an orthonormal basis $\mathbf{w}_1, \cdots, \mathbf{w}_{k+1}$ for W such that $\mathbf{w}_{k+1} \perp \mathbf{v}_1, \cdots, \mathbf{v}_k$. Show that the subspace V' spanned by $\mathbf{v}_1, \cdots, \mathbf{v}_k$ and $\mathbf{w}_{k+1}$ is at least as good as W.

Problem 3 (10 points)**.**
In order to show that SVD can be used to compute the best fit hyperplane for a data set we needed the following result: the best fit k-dimensional hyperplane for a set of n data points in $\mathbb{R}^d$ passes through the centroid of the points. Prove this result.
Hint: Let h be the best fit k-dimensional hyperplane. Consider a subspace S of dimension $d-k$ orthogonal to h. Let p^* be the projection of a data point p on S. Note that the projection of h on S is a single point h^* and the distance between p and h is the same as the distance between p^* and h^*. Argue that h^* is the centroid of the projected points.

Problem 4 (10 points)**.**
The goal in this exercise is to design a simple movie recommendation system using Singular Value Decomposition. Use the dataset used in Assignment 2 Problem 3 (`ml-latest-small.zip` from `https://grouplens.org/datasets/movielens/latest/`).

Using SVD map each user and each movie to a vector in the "feature" or "concept" space as discussed in class. Then for each user u find a movie m that user u has not rated and so that the dot product of their corresponding vectors in concept space is maximized.

You can used Truncated SVD from `sklearn` to efficiently compute only the first few singular vectors. The algorithm by default uses Randomized SVD. See `http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html#sklearn.decomposition.TruncatedSVD`.