Project: Unsupervised Learning (600.692)

Instructor: René Vidal

April 17, 2017

Please submit a one page proposal describing (1) the problem you intend to solve, (2) the algorithms you intend to use, and (3) the datasets and metrics by which you plan to evaluate the algorithms. The algorithms should primarily be algorithms we have discussed in class and/or extensions of them in the GPCA book or closely related literature. Due to limited time for project presentation (about 5 presentations of 10-15 min), the project should be done in groups of three students, but some exceptions will be accepted if you provide a good justification.

You are strongly encouraged to do the movie recommendation project described below, but you may choose an alternative project of your own involving an equivalent amount of work. Please read below for specific details.

1. **Movie Recommendation Grand Challenge.** The goal of this project is to develop a new recommender system for movies. You will be given a matrix of incomplete user ratings for different movies and you will use various low-rank matrix completion methods to predict the ratings for other movies. You will also use various subspace clustering methods to discover genres and low-rank matrix completion algorithms to predict recommendations per genre to see if this gives better recommendations than without using the genres. Specifically, you will do the following:

   (a) Download either the NETFLIX or the Movielens dataset. Study the README files to get familiarized with the database content. Extract all movies corresponding to at least two different genres and define three different datasets: the small dataset #1 will contain movies associated with genre #1, the small dataset #2 will contain movies associated with genre #2, the medium dataset will contain movies associated with genres #1 and #2, and the large dataset will contain all movies from all genres. Note that the NETFLIX dataset does not contain genre information, while the Movielens dataset contains genre information.

   (b) Apply at least three different matrix completion algorithms to the small #1, small #2, medium and large datasets and report the prediction error. One of the three algorithms must be the following

   $$\min_{U,V,d} \|P_\Omega(X - UV^\top)\|_F^2 + \lambda_u \|U\|_F^2 + \lambda_v \|V\|_F^2, \tag{1}$$

   where $X \in \mathbb{R}^{D\times N}$ is the given data matrix, Ω is the set of observed entries, P_Ω is the projection map onto the space of observed entries, $U \in \mathbb{R}^{D\times d}$ and $V \in \mathbb{R}^{N\times d}$ are low-rank factors, and $\lambda_u > 0$ and $\lambda_v > 0$ are tunable parameters. You are free to design any optimization algorithm, e.g., gradient descent, alternating proximal gradient, etc., but you should ensure convergence. Note you are asked to optimize over d.

   (c) Use the best matrix completion algorithm from part (b) to complete the medium dataset. Apply at least three different subspace clustering algorithms to the so-completed medium dataset and report the clustering error. Apply the best matrix completion algorithm from part (b) to each one of the clusters produced by each subspace clustering method and report the prediction error. How does the prediction error compare with the prediction errors you obtained for dataset #1, dataset #2, and medium dataset? Please discuss your results. You may also want to re-cluster the data given the new completion results and so on. Use the best matrix completion algorithm from part (b) to complete the large dataset. Apply all three subspace clustering methods to the so-completed large dataset. Report qualitative clustering results for NETFLIX or quantitative clustering results for Movielens. Apply the best matrix completion algorithm from part (b) to each one of the clusters produced by each method, report the prediction errors and compare them with those obtained in part (b). Does unsupervised genre information improve the quality of the recommendations?

   (d) Discuss the advantages and disadvantages of each matrix completion and subspace clustering method that you encountered.

2. **Submit your own project**. The project topic should follow into one or more of the following categories:

   (a) **Theory:** Extend the theoretical results for the correctness of the methods discussed in class to new domains. For example, do theoretical analysis of alternating minimization for matrix completion (e.g., prove convergence, prove correctness of recovered matrix); do theoretical analysis of convex optimization method for matrix completion with noise (e.g., prove correctness of recovered matrix); do theoretical analysis of sparse subspace clustering with noise (e.g., prove correctness of recovered matrix); extend low-rank matrix completion to the case where the data is not missing uniformly at random; extend algebraic subspace clustering to the case of affine subspaces (with rigorous analysis of the correctness); extend matrix completion to the case where the data points are drawn from a union of subspaces; or extend the theoretical results for the correctness of SSC to include graph connectivity.

   (b) **Algorithms:** Propose new algorithms for solving one of the problems discussed in class. For example, how can you scale the SSC algorithm from 1,000 points in dimension 1,000 to 100,000 points of dimension 100,000?

   (c) **Evaluation:** Evaluate the performance of methods discussed in class on synthetic and/or real data. For example, evaluate and compare all methods for matrix completion, PCA with corrupted entries, PCA with outliers, on synthetic data; perform an exhaustive comparison of 5 subspace clustering algorithms as a function of the dimension of the data, the dimension of the subspaces, the number of subspaces, the amount of noise, the amount of corruptions, etc; or compare 5 subspace clustering algorithms for face clustering or motion segmentation.

   (d) **Applications:** Use a combination of the methods discussed in class to solve a problem in your own area of research. E.g., develop and evaluate face recognition algorithm with variations not only in illumination, but also in pose, expression, with occlusions, etc.

Submission instructions. Please submit a 1-page project proposal by April 24th and 11:59PM. Please submit a report containing title, abstract, introduction, problem description, proposed solution, experiments, conclusions, and references by May 12 at 9AM. Project presentations are scheduled from 10:30-12 noon on May 9th.