Paper submissions are preferred. You can submit the homework in class, or in the box of the second floor in Packard (there is a box for EE378B). You can use any programming language. Submit your code as part of the homework. Tables and figures are preferred to help to illustrate your results. There are several questions for each homework. Make sure that it is easy for readers to figure out which sub-questions you are answering.

This homework is about collaborative filtering. We will use small a dataset collected by GroupLens, a research group at University of Minnesota. The data file can be found on the class webpage.

The file contains 100,000 ratings given by $m = 943$ users to $n = 1682$ movies. The format of the file is as follows. Each row corresponds to a single rating information, given in the following format:

\[
\text{user id — item id — rating — timestamp}
\]

whereby rating is an integer between 1 and 5. (Timestamps will not be used for this homework.)

We will denote by $R_{ij}$ the rating given by user $i \in [m]$ to movie $j \in [n]$, and by $\Omega_{\text{tot}} \subseteq [m] \times [n]$ the set of available ratings. Performance of various methods will be evaluated using the prediction root mean square error (RMSE), which is estimated on a set $S \subseteq [m] \times [n]$ by letting

\[
\text{RMSE} = \sqrt{\frac{1}{|S|} \sum_{(i,j) \in S} (\hat{R}_{ij} - R_{ij})^2},
\]

where $\hat{R}_{ij}$ denotes the algorithm estimate, and $|S|$ the size of set $S$. In order to compute the RMSE, you should follow a ‘cross-validation’ procedure. More precisely: (i) Hold out 20% of the data, chosen uniformly at random. Call this set $S$. (ii) Compute predictions for these using the remaining 80% of the data $\Omega = \Omega_{\text{tot}} \setminus S$; (iii) Estimate the RMSE using the formula $\text{RMSE}$. As a sanity check, the resulting RMSE should not be very sensitive to $S$ (as long as $S$ is random).

(a) Use as a predictor the movie mean rating. What is the resulting RMSE? Repeat the same exercise with the user mean rating.

(b) Call $A$ the matrix whose entry $(i,j)$ is the mean rating for movie $j$. Subtract the movie mean rating from the data, and fill residual entries with 0’s. Call the resulting matrix $M^\Omega$. Project $M^\Omega$ on rank $r$ matrices using singular value decomposition, and call the result $P_r(M^\Omega)$. In other words, $P_r(M^\Omega)$ is obtained from $M^\Omega$ by setting to zero all the singular values except the top $r$ ones. Implement the predictor

\[
\hat{R} = A + \frac{1}{\alpha} P_r(M^\Omega),
\]

where $\alpha \geq 0$ is a regularization parameter. Compute the resulting RMSE for several values of $\alpha$ and $r \in \{5, 10, 15\}$ determine by cross-validation a value that achieves close-to-optimal prediction accuracy.

Compare this with the theoretical value in absence of noise $\alpha = |\Omega|/(mn)$. 

1
(c) Compare the error achieved at the last point with the one obtained with synthetic data. Namely, generate random matrices $X \in \mathbb{R}^{m \times n}$ (with $m = 943$ and $n = 1682$) by letting $X = U V^T / 2$, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ have i.i.d. entries $U_{ia}, V_{ja} \sim \text{Uniform}([+1,1])$, and $r = 20$ (the factor 2 is introduced so that the entries of $X$ are roughly on the same scale as those of $M$ above).

For $\delta \in \{0.05, 0.1, \ldots, 0.95\}$, reveal a random subset $\Omega$ of the entries, whereby $P\{(i,j) \in \Omega\} = \delta$. Denoting by $Y = P_{\Omega}(X) / \delta$ the rescaled observed entries, estimate $X$ by $\hat{X}_r(Y) = P_r(Y)$.

Plot the resulting root mean square error, averaged over $n_{\text{samp}} = 20$ realizations, as a function of $\delta \in \{0.05, 0.1, \ldots, 0.95\}$. Compare the results with the one obtained at the previous point for real data, and with the theory developed in class.

(d) We next reconsider the GroupLens data. Write a program that fits a rank $r = 10$ matrix to the data, by minimizing the cost function (notice that $M_{ij}$ equals the rating minus movie mean):

$$L(X, Y) = \|P_{\Omega}(M - XY^T)\|_F^2 + \lambda\|X\|_F^2 + \lambda\|Y\|_F^2$$

over $X \in \mathbb{R}^{m \times r}$ and $Y \in \mathbb{R}^{n \times r}$, with $\lambda = 20$. Note that

$$\|P_{\Omega}(M - XY^T)\|_F^2 = \sum_{(i,j) \in \Omega} (M - XY^T)_{ij}^2.$$  

The algorithm to be implemented is ALTERNATE LEAST SQUARES, which iteratively minimizes the cost over $X$, then over $Y$, then again over $X$ and so on. Each step is a least squares (quadratic programming) problem.

Plot the prediction RMSE after $t \in \{1, 2, \ldots, 100\}$ iterations, for $\lambda \in \{5, 10, 15, \ldots, 30\}$. 
