Homework 2

Due Monday, May 9

Send the homework via email to cfgranda@cims.nyu.edu, do **not** give in a hard copy

1. *Linear regression* Download the files for this problem here. They contain a training and test predictor matrix together with the corresponding response for three different simulated datasets.
   a. Each training dataset contains 200 examples and 300 predictors. Explain whether it would make sense to apply least-squares regression to learn a linear model from these data and justify your answer briefly.
   b. What error do you expect the lasso to have on the training set when λ is very small? Why?
   c. Plot the coefficient paths for the lasso, the elastic net and ridge regression for each of the datasets (using only the training data). Use either scikit-learn in Python or glmnet in R.
   d. Plot the training and test error of the different methods as a function of the regularization parameter for the three different datasets. Comment on the results, do they make sense considering the paths that you observed in the previous question?
   e. Explain briefly how you would go about selecting the regularization parameters if you only had access to the training data.
   f. Plot the correlation matrix $X^T X$ of the predictors for each of the three training datasets. Does this shed any light on the paths of the different algorithms?
   g. Assume that datasets 1 and 2 were generated using a linear model. Imagine that the sign of one of the coefficients corresponding to the relevant predictors were flipped, would you be able to observe this in the data? Justify your answer briefly.
2. *Movie ratings* Download the file for this problem here. The files contains two simulated datasets of movie ratings.
   a. In total, there are 100 movies and 100 users. The ratings are an integer between 0 and 10. The variable *samples_1* contains the indices of the observed ratings for the first dataset, whereas *data_1* contains the actual ratings. To be clear, if M denotes the original 100×100 matrix in Matlab notation

   $$\text{data\_1} = M\,(\text{samples\_1})\,. \tag{1}$$

   Use a low-rank model to estimate the missing ratings (remember to produce an estimate that is an integer between 1 and 10). To make the model robust to noise use an extra ℓ_2-norm term instead of an equality constraint to enforce data fidelity. Plot a histogram of the estimation errors (the true values are contained in the variable *true_values*) and compute their average.
   b. Do the same for the ratings in *data_2*, which are located at the positions contained in *samples_2*.
   c. You are not very satisfied with your result on the second dataset. A friend suggests that sometimes people will just give random ratings to random movies when they get bored. You

decide to model this behavior using a low-rank + sparse model. Use the model to estimate the missing ratings for the second dataset. Plot a histogram of the estimation errors and compute their average. To calculate the errors compare the low-rank component that you have learnt to *true_values* (note that the entries of *true_values* corresponding to *samples_2* do not necessarily equal to *data_2* because of the sparse rating outliers that you are trying to detect). In addition, plot the low rank and sparse components for different values of the regularization parameter.

d. How would you modify the low-rank model if your assumption was that a small number of users are providing random ratings for every movie they see? You don't need to implement the model.

3. *Real data (*Extra credit*)* Find a dataset online (for example in https://archive.ics.uci.edu/ml/datasets.html) and fit a sparse linear model either for regression or for classification. Divide the data into a training and test set and plot the error on both for different values of the regularization parameter.