Topics on High-Dimensional Data Analytics
ISYE 8803 -
Homework 5

Problem 1. Group lasso (50 points)

There are many regression problems in which the covariates have a natural group structure, and it is desirable to have all coefficients within a group become nonzero (or zero) simultaneously. A leading example is when we have qualitative factors among our predictors. We typically code their levels using a set of dummy variables or contrasts, and would want to include or exclude this group of variables together.

Consider a linear regression model involving J groups of covariates, where for $j = 1, \cdots, J$, the vector $\boldsymbol{Z}_j \in \mathbb{R}^{p_j}$ represents the covariates of group j. Our goal is to predict a real-valued response $Y \in \mathbb{R}$ based on the collection of covariates $(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_J)$. A linear model for the regression function $\mathbb{E}(Y|\boldsymbol{Z})$ takes the form $\sum_{j=1}^{J} \boldsymbol{Z}_j^T \boldsymbol{\theta}_j$, where $\boldsymbol{\theta}_j \in \mathbb{R}^{p_j}$ represents a group of p_j regression coefficients.

Given a collection of N samples $\{(y_i, \boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \cdots, \boldsymbol{z}_{iJ})\}_{i=1}^{N}$, the group lasso solves the convex problem

$$\min_{\boldsymbol{\theta}_j \in \mathbb{R}^{p_j}} \frac{1}{2} ||\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{Z}_j^T \boldsymbol{\theta}_j||_2^2 + \lambda \sum_{j=1}^{J} ||\boldsymbol{\theta}_j||_2 \tag{1}$$

where $||\boldsymbol{\theta}_j||_2$ is the euclidean norm of the vector $\boldsymbol{\theta}_j$.

This is a group generalization of the lasso, with the properties:

- depending on $\lambda \geq 0$, either the entire vector $\boldsymbol{\theta}_j$ will be zero, or all its elements will be nonzero,
- when $p_j = 1$ (continuous variables), then, we have $||\boldsymbol{\theta}_j||_2 = |\theta_j|$, so if all the groups are singletons, the optimization problem reduces to ordinary lasso.

a. **Block Coordinate Descent** (5 points): Show that the group lasso problem in (1) can be solved by iteratively solving, for $j = 1 \cdots, J$:

$$\min_{\boldsymbol{\theta}_j \in \mathbb{R}^{p_j}} \frac{1}{2} ||\boldsymbol{r}_j - \boldsymbol{Z}_j^T \boldsymbol{\theta}_j||_2^2 + \lambda ||\boldsymbol{\theta}_j||_2 \tag{2}$$

where $\boldsymbol{r}_j = \boldsymbol{y} - \sum_{k \neq j} \boldsymbol{Z}_k^T \boldsymbol{\theta}_k$.

b. **Decomposable Functions** (5 points): The problem in equation (2) can be decomposed as a sum of a convex and differentiable function g and a convex but not differentiable function h. Please identify these functions.

c. **Proximal Gradient Descent - Gradient Step** (10 points): To solve the problem in equation (2), we want to use proximal gradient descent. Show that the gradient step is:

$$\boldsymbol{a}_{j,k} = \boldsymbol{\theta}_{j,k} - t_k \nabla g(\boldsymbol{\theta}_{j,k}) = \boldsymbol{\theta}_{j,k} - t_k(-\boldsymbol{Z}_j^T(\boldsymbol{r}_j - \boldsymbol{Z}_j^T \boldsymbol{\theta}_{j,k})) \tag{3}$$

where t is a step size parameter.

d. **Proximal Gradient Descent - Parameters Step** (5 points): Write the proximal problem that needs to be solved next. For the remaining of the homework, you can use the fact that the solution to the proximal problem is:

$$\boldsymbol{\theta}_{j,k+1} = \left(1 - \frac{t_k \lambda}{||\boldsymbol{a}_{j,k+1}||_2}\right)_+ \boldsymbol{a}_{j,k+1} = \begin{cases} \left(1 - \frac{t_k \lambda}{||\boldsymbol{a}_{j,k+1}||_2}\right) \boldsymbol{a}_{j,k+1} & \text{if } t_k \lambda < ||\boldsymbol{a}_{j,k+1}||_2 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Notice that when $p_j = 1$ (continuous variables), we have that:

$$\theta_{j,k+1} = \left(1 - \frac{t_k \lambda}{|a_{j,k+1}|}\right)_+ a_{j,k+1} = \begin{cases} a_{j,k+1} - t_k \lambda & \text{if } t_k \lambda < a_{j,k+1} \\ 0 & -t_k \lambda \leq a_{j,k+1} \leq t_k \lambda \\ a_{j,k+1} + t_k \lambda & \text{if } t_k \lambda > a_{j,k+1} \end{cases} = S_{t_k \lambda}(a_{j,k+1}). \tag{5}$$

The houses dataset contains a collection of recent real estate listings in San Luis Obispo county and around it. The dataset is provided in RealEstate.csv. It contains the following fields:

- Price (continuous): the most recent listing price of the house (in dollars)
- Bedrooms (categorical): number of bedrooms
- Bathrooms (categorical): number of bathrooms
- Price/SQ.ft (continuous): price of the house per square foot
- Status (categorical): type of sale. Three types are represented in the dataset: Short Sale, Foreclosure and Regular

e. **Block coordinate proximal gradient descent** (25 points): Implement the block coordinate proximal gradient descent to solve the group lasso problem presented in equation (1).

  - Your dependent variable is the house price.
  - The first step is to create dummy variables corresponding to the categorical variables (bedrooms, bathrooms and status). To avoid multicollinearity issues use 0 bedrooms, 1 bathroom, and short sale as baselines, respectively.
  - To improve the results, your second step should be to standardize the data (all data should be between 0 and 1).

- Use $\lambda = 0.012$

Notice that you cannot use group lasso built in functions for this problem.

Present your results. What variables are more relevant to explain a house price?

Problem 2. Consensus optimization (50 points)
For this problem consider the image img.pb3.mat. This is a 31 × 31 sparse image, 620 of the 961 pixels have a value of 0. You can think of this image as a toy version of an MRI image that we are interested in collecting.
Suppose that, because of the nature of the machine that collects the MRI image, it takes a long time to measure each pixel value individually, but it is faster to measure linear combinations of pixel values. The machine produces three batches of 100 linear combinations, with the weights in the linear combination being random. Additionally, since the machine is not perfect, noise is added to the linear combinations.
In mathematical notation, we have

$$y_{ij} = \sum_{p=1}^{961} x_{ijp}\beta_p + \epsilon_{ij} \tag{6}$$

where $x_{ijp} \sim N(0,1)$ are the random weights, β_p is the vectorized image, $\epsilon_{ij} \sim N(0,25)$ is the measurement noise, $i = 1,2,3,\ j = 1,\cdots,100$ and $p = 1,\cdots,961$.
The observed values y_{ij}, and the random weights x_{ijp} can be found as MRI.mat.
Although the number of measurements $n = 300$ is smaller than the dimension of the image $p = 961$, since the image is sparse, it is possible to recover it. This is the idea behind compress sensing that we will study in module 7.
Here, we will recover the image by solving the following optimization problem:

$$\min_{\beta\in\mathbb{R}^p} \sum_{b=1}^{3} \frac{1}{2}||y_b - X_b\beta_b||_2^2 + \lambda||\beta||_1 \tag{7}$$

For this purpose, we will use alternating direction method of multipliers (ADMM).

a. (5 points) Show that the optimization problem in equation (7) can be written as:

$$\min_{\beta\in\mathbb{R}^p} \sum_{b=1}^{3} \left(\frac{1}{2}\beta_b^T X_b^T X_b\beta_b - y_b^T X_b\beta_b\right) + \lambda||\beta||_1 + K \tag{8}$$

where K is a constant that does not depend on β.
This problem can be written as:

$$\min_{\beta\in\mathbb{R}^p} \sum_{b=1}^{3} \left(\frac{1}{2}\beta_b^T X_b^T X_b\beta_b - y_b^T X_b\beta_b\right) + \lambda||\theta||_1 \quad s.t.\ \beta_b = \theta,\ b = 1,2,3 \tag{9}$$

b. (10 points) Write the augmented Lagrangian function for a given ρ.
It can be shown (you do NOT need to do it) that the augmented Lagrangian function can be simplified to:

$$L_\rho(\beta_b, \theta, u_b) = \sum_{b=1}^{3} \left(\frac{1}{2}\beta_b^T X_b^T X_b \beta_b - y_b^T X_b \beta_b \right) + \lambda ||\theta||_1 + \frac{\rho}{2} ||\beta_b - \theta + u_b||_2^2. \tag{10}$$

c. (10 points) For $\beta_b,\ b = 1, 2, 3$ since we have a quadratic objective show that:

$$\beta_b^{t+1} = (X_b^T X_b + \rho I)^{-1}(X_b^T y_b + \rho(\theta^t - u_b^t)) \tag{11}$$

Additionally, (you do NOT need to show this), we have that:

$$\theta^{t+1} = S_{\lambda/(\rho B)}(\overline{\beta}^{t+1} + \overline{u}^t) = \begin{cases} \overline{\beta}^{t+1} + \overline{u}^t - \dfrac{\lambda}{\rho B} & \text{if } \overline{\beta}^{t+1} + \overline{u}^t > \dfrac{\lambda}{\rho B} \\ 0 & \text{if } |\overline{\beta}^{t+1} + \overline{u}^t| \le \dfrac{\lambda}{\rho B} \\ \overline{\beta}^{t+1} + \overline{u}^t + \dfrac{\lambda}{\rho B} & \text{if } \overline{\beta}^{t+1} + \overline{u}^t < -\dfrac{\lambda}{\rho B} \end{cases} \tag{12}$$

where $\overline{\beta}^{t+1}$ and $\overline{u}^t$ denote averages over the blocks, and

$$u_b^{t+1} = u_b^t + (\beta_b^{t+1} - \theta^{t+1}). \tag{13}$$

d. (25 points) Implement the ADMM algorithm to obtain the MRI image (you need to develop your own code, you CANNOT use ADMM built in functions), use $\lambda = 1$ and $\rho = 0.5$. Please attach the image you obtain. Compare the true image with the recovered image.