MATH 829 - Spring 2016
Introduction to data mining and analysis
Homework #2

Instructions:

(1) You are allowed to work in groups of 1-4.
(2) Show (and briefly explain) all of your work to receive full credit.
(3) Submit your work before **Friday, March 11th, 2016**.

Theoretical part

Problem 1. Recall that the *singular value decomposition* (SVD) of $X \in \mathbb{R}^{n\times p}$ is $X = U\Sigma V^T$ where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{n\times p}$ is diagonal with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$.

a) Let $y \in \mathbb{R}^{n\times 1}$. Assuming $X^T X$ is invertible, show that

$$X\hat{\beta}^{\mathrm{LS}} = \sum_{i=1}^{p} u_i u_i^T y,$$

where $\hat{\beta}^{\mathrm{LS}}$ denotes the least squares solution of the system $y = X\beta$, and u_i denotes the i-th column of U.

b) If $\hat{\beta}^{\mathrm{ridge}}$ denotes the Ridge solution of the linear system $y = X\beta$ with parameter $\lambda > 0$, show that

$$X\hat{\beta}^{\mathrm{ridge}} = \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} u_i u_i^T y.$$

Problem 2. Let $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^{n\times 1}$. Show that the elastic net problem

$$\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \gamma_1\|\beta\|_1 + \gamma_2\|\beta\|_2^2 \qquad (\gamma_1, \gamma_2 > 0)$$

can be computed by solving a lasso problem with input data $X^* \in \mathbb{R}^{(n+p)\times p}$, $y^* \in \mathbb{R}^{(n+p)\times 1}$.
HINT: Let X^* be obtained by augmenting X by a multiple of the identity. Compute $X^*\beta$.

Problem 3. Suppose $X \in \mathbb{R}^{n\times p}$ has orthonormal columns and $y \in \mathbb{R}^{n\times 1}$. Show that the solution of the lasso problem

$$\hat{\beta}^{\mathrm{lasso}} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \alpha\|\beta\|_1 \qquad (\alpha > 0)$$

is obtained by soft-thresholding the least squares solution, i.e.,

$$\hat{\beta}_i^{\mathrm{lasso}} = \operatorname{sgn}(\hat{\beta}_i^{\mathrm{LS}})(|\hat{\beta}_i^{\mathrm{LS}}| - \alpha)_+ \qquad (i = 1, \dots, p),$$

where $\hat{\beta}^{\mathrm{LS}}$ denotes the least squares solution of the problem.

Python part

Problem 4.

a) Implement the coordinate descent method to solve the lasso problem.
b) Use `sklearn` to verify that your implementation is correct.
c) Use simulated data to estimate the average time your algorithm takes to solve the lasso problem as a function of p. Repeat the same experiment for different values of n/p.

Problem 5. The file `assay.csv` (available on Sakai) contains gene expression data ($p = 7,129$ genes) for $n = 49$ breast cancer tumor samples. The file `pheno.csv` contains a binary response variable for each sample (whether the sample tested positive or negative).

a) Split the dataset into a training and a test set. Train a Lasso model to predict the response variable from the gene expression data using cross-validation. Compute the prediction error on your test set.
b) Repeat the previous experiment using $N = 100$ random train/test pairs. Compute the average prediction error and its standard deviation.

Problem 6. The files `zip.train` and `zip.test` (available on Sakai) contain normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service. The images have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990). There are 7,291 observations in `zip.train` and 2,007 observations in `zip.test`.

a) Use the training set `zip.train` to train a model to predict the digits.
b) Compute the prediction error of your model on the test set `zip.test`.

Note: `sklearn.multiclass` has `OneVsRestClassifier` and `OneVsOneClassifier` objects if you want to use the one-vs-rest or one-vs-one classification strategies (see the `sklearn` documentation for more details).