4. Coordinate Descent Ridge Regression

Coordinate descent (CD) cycles through successive minimizations of the objective function $F(\mathbf{w})$ with respect to each coordinate:

$$w_i = \text{argmin}_w F(w_1, \dots, w_{i-1}, w, w_{i+1}, \dots, w_p), i = 1, 2 \dots, p, 1, 2, \dots,$$

Here you will implement CD for ridge regression. The ridge regression objective function is

$$F_{ridge}(\mathbf{w}) = \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}))^2 + \lambda \sum_{i=1}^{p} |w_i|^2,$$

where $\lambda > 0$ and $\hat{y}_j(\mathbf{w}) = w_0 + \sum_{i=1}^{p} w_i x_{ij} = w_0 + \mathbf{x}^T \mathbf{w}_1$. As usual x_{ij} is the ij element of the $p \times n$ feature matrix $\mathbb{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$.

(a). Assume that the features have been adjusted so that they have zero mean, i.e., $\mathbb{X}\mathbf{1} = \sum_{j=1}^{n} \mathbf{x}_j = 0$, which will be true when you sphere the training data (see Problem 1). Show that the w_0 that minimizes F_{ridge} is $w_0^* = \frac{1}{n} \sum_{j=1}^{n} y_j$.

(b). Show that the partial derivative with respect to w_i, $i = 1, \dots, d$, of the data-dependent term in F_{ridge} is

$$\frac{\partial}{\partial w_i} \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}))^2 = a_i w_i - c_i$$

where

$$a_i = 2 \sum_{j=1}^{n} x_{ij}^2, \qquad c_i = 2 \sum_{j=1}^{n} x_{ij} (y_j - \hat{y}_j(\mathbf{w}_{-i})), \qquad i \geq 1$$

with $\hat{y}_j(\mathbf{w}_{-i})$ the predictor with coefficient w_i set to zero, i.e.,

$$\hat{y}_j(\mathbf{w}_{-i}) = w_0 + \sum_{k \neq i}^{p} w_k x_{kj}, \qquad i \geq 1.$$

where the summation is over $k = 1, \dots, d$, excluding index i.

(c). The optimality condition for w_i to minimize the ith coordinate of $F_{ridge}(\mathbf{w})$ is $\frac{\partial F(\mathbf{w})}{\partial w_i} = 0$, $i = 1, \dots, d$. By solving this equation for w_i show that the CD update of w_i is

$$\text{argmin}_{w_i} F_{ridge}(\mathbf{w}) = w_i = \frac{c_i}{a_i + 2\lambda}, \qquad i \geq 1.$$

(d). Here you will implement the coordinate descent ridge regression algorithm and apply it to the **sphered** Ames Iowa Housing data you treated in Problem 1. Make sure you include the y-intercept weight $w_0 = \frac{1}{n} \sum_{j=1}^{n} y_j$ in the data dependent term but exclude it from the penalty term in $F_{ridge}(\mathbf{w})$. Using the ridge regularization parameter $\lambda = 100$ and initializing your $\mathbf{w}$ to be a vector with all elements equal to 1, run your CD ridge regression for 2900 iterations (50 cycles).

Plot the evolution of the weights (a single plot of all weights except for w_0 over iteration), and also plot the learning curve with respect to average squared prediction error $\frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}^{train}))^2$ ($n = 2000$) on the training data. The label on the x-axis should be number of iterations.

Hint: Your weights should converge to the exact ridge weights implemented in Problem 1 when $\lambda = 100$.