Coordinate Descent

What is coordinate descent?

Coordinate Descent is an optimization technique where we optimize with respect to one parameter at a time rather than an entire set of parameters. This allow for more flexibility within an optimization process.

Why do we use coordinate descent?

We typically use coordinate descent when the alternative method of gradient descent is not possible. Most often the reason that gradient descent wouldn’t be possible is because the objective function is non-differentiable. Often, when a function is non-differentiable, if we break the equation down into singular-parameter steps (i.e coordinate descent) we are able to use a method like subgradients to still solve for individual derivatives.

How is coordinate descent different than gradient descent?

As explained above, coordinate descent iterates through each parameter optimizing only one at each step. Whereas in gradient descent we take the gradient wrt. the entire set of parameters and update all of the parameters in each step. Or in the the case of the closed form OLS we optimize all the parameters at once via solving for the gradient.

OLS via Coordinate Descent

Let’s look at an example of how we use could use coordinate descent for OLS (Linear Regression):

OLS Equation

Matrix Notation: $SSE = \frac{1}{2}(Y - \hat{Y})^2$

Summation Notation: $SSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta x_i)^2$

Expanded Summation Notation: $SSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{n}\theta_j x_i^j)^2$

We expand this equation for mathematical convenience. We are trying to target a single θ parameter. In order to do this we need to single out a single parameter. **So let's expand this equation even further for our convenience.**

Expanding....

$$SSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - (\sum_{j\neq k}^{n}\theta_j x_i^j + \theta_k x_i^k))^2$$
$$SSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - (\sum_{j\neq k}^{n}\theta_j x_i^j) - \theta_k x_i^k)^2$$

Next let's take the derivative wrt to a single parameter θ_k.

Using the chain rule ($F'(x) = f'(g(x)) \cdot g'(x)$)

$$f'(g(x)) = \sum_{i=1}^{n}(y_i - (\sum_{j\neq k}^{n}\theta_j x_i^j) - \theta_k x_i^k)$$
$$g'(x) = -x_i^k$$
$$SSE' = -\sum_{i=1}^{n} x_i^k(y_i - (\sum_{j\neq k}^{n}\theta_j x_i^j + \theta_k x_i^k))$$

Now that we have the derivative wrt. to θ_k we can solve for θ_k.

First, lets isolate our θ_k and move it outside the summation

.

$$SSE' = -\sum_{i=1}^{n} x_i^k(y_i - \sum_{j\neq k}^{n}\theta_j x_i^j) + \theta_k\sum_{i=1}^{n}(x_i^k)^2$$

When working with lasso regression we will need to normalize our data, the reason we do this is to simplify our equation. If our data is normalized then $\sum_{i=1}^{n}(x_i^k)^2 = 1$ meaning we can ignore it from our equation. **Resulting in the simpler form of:**

$$SSE' = -\sum_{i=1}^{n} x_i^k(y_i - \sum_{j\neq k}^{n}\theta_j x_i^j) + \theta_k$$

Now solve for θ_k:

$$\theta_k = \sum_{i=1}^{n} x_i^k (y_i - \sum_{j \neq k}^{n} \theta_j x_i^j)$$

Done.