Name: Gosuddin Siddiqi

AMATH 521 Homework Set 2

Due: Monday October 30th, on Canvas.

***Problem 1**: More Prox*

Recall that

$$\text{prox}_{\alpha f}(z) = \underset{x}{\text{argmin}} \frac{1}{2\alpha}\|x - z\|^2 + f(x) \tag{1}$$

Suppose f is convex and $\alpha > 0$.

(1) Show $\text{prox}_{\alpha f}$ is a single-valued mapping (i.e. there's a unique solution to (1)).

The equation $\text{prox}_{\alpha f}$ comprises of two parts.
$\frac{1}{2\alpha}|||x - z|||^2$ (1)
and $f(x)$ (2)

Part (1) is strictly convex and Part (2) is given as convex. A combination of these two creates a convex function. As we know that a strictly convex function has just one global minimizer. The composition of these two function would also result in having just one global minimizer.

Since the Part (1) is strongly convex, it also has a quadratic lower bound. This enforces the function to have extreme values ($\pm\infty$) when $|||x||| \to \pm\infty$, exhibiting that the sum is coercive.

Hence the composition of these two parts, having one unique global minimizer, makes $\text{prox}_{\alpha f}$ have a single-valued mapping.

(2) Compute $\text{prox}_{\alpha f}$, for $f(x) = \beta\|x\|_1 + (1 - \beta)\|x\|_2^2$.

Solution:

Given, $f(x) = \frac{1}{2\alpha}|||x - z|||^2 + \beta\|x\|_1 + (1 - \beta)\|x\|_2^2$.

We can treat every point iteration as separate function as the every iteration is independent of previous or next data point. This makes our problem equivalent to solving n equations.

Solving for the First-order derivative of $f(x)$,

$\nabla f(x) = \frac{1}{\alpha}(|||x - z|||) + \beta + 2(1 - \beta)$

Equating to zero to minimize the derivate and rearranging the terms, we get, When $x > 0$,

$x = \frac{z-\alpha\beta}{1+2\alpha-2\alpha(1-\beta)}$

This implies that $z > \alpha\beta$. ...(1)

Similarly when $x < 0$,

$x = \frac{z+\alpha\beta}{1+2\alpha-2\alpha(1-\beta)}$

This implies that $z < -\alpha\beta$. ...(2)

When $x = 0$,

$|z| \leq \alpha\beta$ ...(3)

Compiling $(1), (2), (3)$,

$\text{prox}_{\alpha f}(z) =$

$$\begin{cases} z > \alpha\beta, & x_i > 0 \\ z < -\alpha\beta, & x_i < 0 \\ |z| \leq \alpha\beta, & x_i = 0 \end{cases}.$$

(3) Compute $\text{prox}_{\alpha f}$, for $f(x) = \|x\|_2$. (no square)

Given $f(x) = \frac{1}{2\alpha}|||x - z|||^2 + \|x\|_2$.

This is differentiable when $\|x\| > 0$

Differentiating, we get

$\nabla f(x) = \frac{(x-z)}{\alpha} + \frac{x}{|x|}$

Equating the above $\nabla f(x)$ to zero and rearranging,

$x = z\frac{\|x\|}{(\|x\|+\alpha)}$

Thus, $\|z\| = \|x\| + \alpha$ ...(1)

When $\|x\| > 0$, $\|z\| > \alpha$ ...(2)

We get $\|x\| = 0$ when $\|z\| < \alpha$ ...(3)

$\text{prox}_{\alpha f}(z) =$

$$\begin{cases} z(1 - \frac{\alpha}{\|z\|}), & \|z\| > \alpha \\ 0, & \|z\| \leq \alpha \end{cases}.$$

(4) Compute $\text{prox}_{\alpha f}$, for $f(x) = \|x\|_2^2 + \delta_{\mathbb{R}^n_+}(x)$, where

$$\delta_{\mathbb{R}^n_+}(x) = \sum_{i=1}^{n} \begin{cases} 0, & x_i \geq 0 \\ \infty, & x_i < 0 \end{cases}.$$

Again since the $f(x)$ is independent of the previous or the next iteration in the dataset, we can treat the function as n independent minimization problem.

$f(x)$ is composed of three parts and the last part $\delta_{\mathbb{R}^n_+}(x)$ enforce the minimization problem to have a projection in the positive of x, as $f(x)$ is ∞ when x is negative.

Now, when $z \leq 0$, the minimization of the $f(x)$ depends on minimizing the $|x|$, and the best possible value is when $x = 0$

When $z > 0$, minimized value can be achieved when $x > 0$

The derivative when equated to zero and rearranged we get,

$x = \frac{z}{1+2\alpha}$

This can be compiled as,

$\text{prox}_{\alpha f}(z) =$

$$\begin{cases} \frac{z}{1+2\alpha}, & z > 0 \\ 0, & z \leq 0 \end{cases}.$$

Problem 2: *Bias-Variance Decomposition*

In the lecture, we showed that generalized error could be decomposed into irreducible error, error from bias, and error from variance, using linear regression as a specific example. Consider more general setting:

- $\{y_i, a_i\}_{i=1}^m$ is the training data
- the model is
$$y_i = f(a_i) + \epsilon_i, \quad i = 1, \ldots, m$$
where ϵ_i are i.i.d. random variables with mean 0 and variance σ_ϵ^2.
- f is the true data generating mechanism, and $\hat{f}$ is the estimated model.
- $\{y_0, a_0\}$ is a test data point, which also satisfies
$$y_0 = f(a_0) + \epsilon_0.$$

Based on this information please show that,

$$\begin{aligned}\text{Err}_{a_0} =& \mathbb{E}\left[\left(y_0 - \hat{f}(a_0)\right)^2\right] \\ =& \sigma_\epsilon^2 + \left(\mathbb{E}\left[\hat{f}(a_0)\right] - f(a_0)\right)^2 + \mathbb{E}\left[\left(\hat{f}(a_0) - \mathbb{E}\left[\hat{f}(a_0)\right]\right)^2\right]\end{aligned}$$

Solution:

In our solution, let $f = f(a_0)$ and $\hat{f} = \hat{f}(a_0)$

For any random variable X, we know that $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Rearranging the equation, we get,
$\mathbb{E}[X^2] = Var[X] + \mathbb{E}[X]^2 \quad \ldots(1)$

Since f is a true data generating mechanism, we can conclude that $\mathbb{E}(f) = f \quad \ldots(2)$

From given, we know that, $y = f + \epsilon$ and that its mean is zero ($\mathbb{E}[\epsilon] = 0$). This implies,
$\mathbb{E}[y] = \mathbb{E}[f + \epsilon] = \mathbb{E}[f] = f \quad \ldots(3)$

It is given that, $Var[\epsilon] = \sigma^2$

We can say that,
$Var[y] = \mathbb{E}[(y - \mathbb{E}[y])]^2 = \mathbb{E}[(y - f)^2] = \mathbb{E}[(f + \epsilon - f)^2] = \mathbb{E}[(\epsilon)^2] \ldots(4)$

Using (1),
$\mathbb{E}[(\epsilon)^2] = Var[\epsilon] + \mathbb{E}[\epsilon]^2 = \sigma^2$

Now, $Err_a 0 = Err_f = \mathbb{E}[(y - \hat{f})^2]$ can be expanded as follows
$= \mathbb{E}[(y^2 + \hat{f}^2 - 2y\hat{f})]$
$= [(y^2)] + \mathbb{E}[(\hat{f}^2)] - \mathbb{E}[(2y\hat{f})]$

Using (1),we get,
$= Var[y] + \mathbb{E}[y]^2 + Var[\hat{f}] + \mathbb{E}[\hat{f}]^2 - 2f\mathbb{E}[\hat{f}]$
Using (3) and rearranging, we get,
$= Var[y] + Var[\hat{f}] + (f^2 - 2f\mathbb{E}[\hat{f}] + \mathbb{E}[\hat{f}]^2)$
$= Var[y] + Var[\hat{f}] + (\mathbb{E}[\hat{f}] - f)^2$
Using (4), we can say that,
$= \sigma_\epsilon^2 + Var[\hat{f}] + (\mathbb{E}[\hat{f}] - f)^2$
The above equation is composed of three parts, (1) irreducible error, (2), variance, and (3) bias.

The final equation can be represented into the desired format using (3) and substituting back $f = f(a_0)$ and $\hat{f} = \hat{f}(a_0)$,
$= \sigma_\epsilon^2 + \left(\mathbb{E}\left[\hat{f}(a_0)\right] - f(a_0)\right)^2 + \mathbb{E}\left[\left(\hat{f}(a_0) - \mathbb{E}\left[\hat{f}(a_0)\right]\right)^2\right]$

***Problem 3**: Newton's Method*

Consider logistic regression problem,

$$\min_x f(x) := \sum_{i=1}^{m} \{\log(1 + \exp(\langle a_i, x\rangle)) - b_i \langle a_i, x\rangle\} + \frac{\lambda}{2}\|x\|_2^2$$

(1) What is the Hessian of this objective, $\nabla^2 f(x)$?

Let, $\langle a_i, x\rangle = u$

From the previous assignment we computed the $\nabla f(x)$ which is equal to,
$\nabla f(x) = F^T(\frac{1}{1+exp(-u)} - b) + \lambda x$ ...(1)

We know that $(\frac{1}{1+exp(-u)})$ is a sigmoid function. Let this be $\sigma(x)$.
Differentiating $\sigma(x)$, we get $\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$

Thus the Hessian for the equation (1), with a diagonal matrix R formed with the entries from $\frac{d}{dx}\sigma(x)$ can be written as,
$\nabla f^2(x) = F^T RF + \lambda$

(2) Implement Newton's method in **Jupyter Notebook**. Use the validation dataset to select λ. Report your best test error.

***Problem 4**: FISTA*

Using the data set from **Problem 3**, pick your favorite penalty $g(x)$ (not limited by the ones we saw on class) and solve logistic regression problem,

$$\min_x \sum_{i=1}^{m} \{\log(1 + \exp(\langle a_i, x\rangle)) - b_i \langle a_i, x\rangle\} + \lambda\, g(x)$$

(1) What is your g? Calculate $\text{prox}_{\alpha g}$ based the g you provide.

(2) Implement the FISTA algorithm in **Jupyter Notebook**, and use validation to select the best λ. Report your best test error.

We will have a competition, the one with the best test error will receive bonus credits!