Math189R SP19
Homework 2
Monday, Feb 11, 2019

Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

The starter files can be found under the Resource tab on course website. The graphs for problem 3 generated by the sample solution could be found in the corresponding zipfile. These graphs only serve as references to your implementation. You should generate your own graphs for submission. Please print out all the graphs generated by your own code and submit them together with the written part, and make sure you upload the code to your Github repository.

1 (**Murphy 8.3**) Gradient and Hessian of the log-likelihood for logistic regression.

(a) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Show that

$$\sigma'(x) = \sigma(x)\left[1 - \sigma(x)\right].$$

(b) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.

(c) The Hessian can be written as $\mathbf{H} = \mathbf{X}^\top \mathbf{S}\mathbf{X}$ where $\mathbf{S} = \text{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Derive this and show that $\mathbf{H} \succeq 0$ ($A \succeq 0$ means that A is positive semidefinite).

Hint: Use the **negative** log-likelihood of logistic regression for this problem.

Part a: Given $\sigma(x) = \frac{1}{1+e^{-x}}$, we can find $\sigma'(x)$ as the following.

$$\begin{aligned}
\sigma'(x) &= \frac{d}{dx}(\sigma(x)) \\
&= \frac{d}{dx}(\frac{1}{1+e^{-x}}) \\
&= e^{-x}(1+e^{-x})^{-2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{1+e^{-x}-1}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot (1 - \frac{1}{1+e^{-x}}) \\
&= \sigma(x)[1-\sigma(x)]
\end{aligned}$$

Part b: By definition, the negative log likelihood for logistic regression is the following

$$nll(\theta) = -\sum_i y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))$$

Next, we can find the gradient of nll with respect to θ.

$$\begin{aligned}
\nabla_\theta nll(\theta) &= -\sum_i \frac{y_i}{\sigma(\theta^T x_i)} \cdot \sigma'(\theta^T x_i) - \frac{(1 - y_i)}{1 - \sigma(\theta^T x_i)} \cdot -\sigma'(\theta^T x_i) \\
&= -\sum_i y_i(1 - \sigma(\theta^T x_i)) - (1 - y_i)\sigma(\theta^T x_i)x_i \\
&= -\sum_i y_i x_i - y_i \sigma(\theta^T x_i)x_i - \sigma(\theta^T x_i)x_i + y_i \sigma(\theta^T x_i)x_i \\
&= \sum_i (\sigma(\theta^T x_i) - y_i)x_i \\
&= \sum_i (\mu_i - y_i)x_i \\
&= X^T(\mu - y)
\end{aligned}$$

where $\mu_i = \sigma(\theta^T x_i)$ and x_i is the i^{th} column of X^T.

Part c: From part b, we have $nll(\theta))^T = \nabla_\theta [X^T(\mu - y)]^T$. Then, we can use this to find the Hessian matrix as the following.

$$\begin{aligned}
H_\theta &= \nabla_\theta (\nabla_\theta nll(\theta))^T \\
&= \nabla_\theta [X^T(\mu - y)]^T \\
&= \nabla_\theta (\mu^T X - y^T X) \\
&= \nabla_\theta \mu^T X = \nabla_\theta \sigma(X\theta)^T X \\
&= X^T \text{diag}(\mu(1 - \mu))X \\
&= X^T S X
\end{aligned}$$

where $S = \text{diag}(\mu_1(1 - \mu_1), \ldots, \mu_n(1 - \mu_n))$.

Note that since S is a diagonal matrix, its eigenvalues are its diagonal entries, and it is positive semi-definite if the diagonal entries are greater than 0. Also note that $0 < \sigma(x) < 1$ for any x. Since $\mu_i = \sigma(\theta^T x_i)$, we can see that $\mu_i(1 - \mu_i) = \sigma(\theta^T x_i)(1 - \sigma(\theta^T x_i)) \geq 0$. Thus, S is positive semi-definite. Thus, H_θ must also be semi-definite.

■

2 (**Murphy 2.11**) Derive the normalization constant (Z) for a one dimensional zero-mean Gaussian

$$\mathbb{P}(x;\sigma^2) = \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

such that $\mathbb{P}(x;\sigma^2)$ becomes a valid density.

By definition, a valid density function must integrate to one.

$$\int_{\mathbb{R}} p(x;\sigma^2)dx = \int_{\mathbb{R}} \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)dx = \frac{1}{Z}\int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx = 1$$

Thus, multiplying Z over, we get.

$$Z = \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx$$

Now we can consider Z^2 to evaluate the integral.

$$\begin{aligned}
Z^2 &= \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx \int_{\mathbb{R}} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy \\
&= \iint_{\mathbb{R}^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)dxdy \\
&= \int_0^{\infty}\int_0^{2\pi} \exp\left(-\frac{r^2}{2\sigma^2}\right)rd\theta dr \\
&= 2\pi\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right)rdr \\
&= 2\pi(-\sigma^2)\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right)\left(-\frac{r}{\sigma^2}\right)dr \\
&= -2\pi\sigma^2\exp\left(-\frac{r^2}{2\sigma^2}\right)\Big|_0^{\infty} \\
&= -2\pi\sigma^2(0-1) \\
&= 2\pi\sigma^2
\end{aligned}$$

Thus, we have $Z^2 = 2\pi\sigma^2$ which means $Z = \sqrt{2\pi\sigma^2} = \sqrt{2\pi}\sigma$.

■

3 (**regression**). In this problem, we will use the online news popularity dataset to set up a model for linear regression. In the starter code, we have already parsed the data for you. However, you might need internet connection to access the data and therefore successfully run the starter code.

We split the csv file into a training and test set with the first two thirds of the data in the training set and the rest for testing. Of the testing data, we split the first half into a 'validation set' (used to optimize hyperparameters while leaving your testing data pristine) and the remaining half as your test set. We will use this data for the remainder of the problem. The goal of this data is to predict the **log** number of shares a news article will have given the other features.

(a) (**math**) Show that the maximum a posteriori problem for linear regression with a zero-mean Gaussian prior $\mathbb{P}(\mathbf{w}) = \prod_j \mathcal{N}(w_j|0, \tau^2)$ on the weights,

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i|w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j|0, \tau^2)$$

is equivalent to the ridge regression problem

$$\arg\min \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^\top \mathbf{x}_i))^2 + \lambda ||\mathbf{w}||_2^2$$

with $\lambda = \sigma^2/\tau^2$.

(b) (**math**) Find a closed form solution $\mathbf{x}^\star$ to the ridge regression problem:

$$\text{minimize: } ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

(c) (**implementation**) Attempt to predict the log shares using ridge regression from the previous problem solution. Make sure you include a bias term and *don't regularize the bias term*. Find the optimal regularization parameter λ from the validation set. Plot both λ versus the validation RMSE (you should have tried at least 150 parameter settings randomly chosen between 0.0 and 150.0 because the dataset is small) and λ versus $||\boldsymbol{\theta}^\star||_2$ where $\boldsymbol{\theta}$ is your weight vector. What is the final RMSE on the test set with the optimal $\lambda^\star$?

(continued on the following pages)

■

3 (continued)

(d) (**math**) Consider regularized linear regression where we pull the basis term out of the feature vectors. That is, instead of computing $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x}$ with $\mathbf{x}_0 = 1$, we compute $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x} + b$. This corresponds to solving the optimization problem

$$\text{minimize: } ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

Solve for the optimal $\mathbf{x}^\star$ explicitly. Use this close form to compute the bias term for the previous problem (with the same regularization strategy). Make sure it is the same.

(e) (**implementation**) We can also compute the solution to the least squares problem using gradient descent. Consider the same bias-relocated objective

$$\text{minimize: } f = ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

Compute the gradients and run gradient descent. Plot the ℓ_2 norm between the optimal $(\mathbf{x}^\star, b^\star)$ vector you computed in closed form and the iterates generated by gradient descent. Hint: your plot should move down and to the left and approach zero as the number of iterations increases. If it doesn't, try decreasing the learning rate.

Part a: First, we apply the probability distribution $\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ to the maximum a posteriori problem, in which we get

$$\arg\max_w \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{w_j^2}{2\tau^2}\right)$$

By property of logarithm, we can simplify to

$$\arg\max_w \sum_{i=1}^{N} \left(-\frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma}\right) + \sum_{j=1}^{D} \left(-\frac{w_j^2}{2\tau^2} - \log\sqrt{2\pi\tau}\right)$$

$$= \arg\max_w - \left((N+D)\log\sqrt{2\pi\tau} + \sum_{i=1}^{N} \frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}\right)$$

Since constants do not affect the optimization problem, we can drop the term $-(N+D)\log\sqrt{2\pi\tau}$. To facilitate calculation, we can also scale our problem by $2\sigma^2$ without changing the optimal w^*.

Also, since maximizing a function is equivalent to minimizing its negative, we have:

$$\arg\min_w \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2$$

Let $\lambda = \frac{\sigma^2}{\tau^2}$, we have

$$\arg\min_w \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda \sum_{j=1}^{D} w_j^2.$$

Then, we can rewrite as

$$\arg\min_w \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda \|w\|^2$$

which results equivalent to the ridge regression problem, as desired.

Part b: First, we find the gradient of the objective function f with respect to x.

$$\begin{aligned}\nabla_x f &= \nabla_x \left((Ax-b)^T (Ax-b) + (\Gamma x)^T (\Gamma x) \right) \\ &= \nabla_x \left(x^T A^T A x - 2x^T A^T b + b^T b + x^T \Gamma^T \Gamma x \right) \\ &= \nabla_x \left[x^T A^T A x - 2x^T A^T b + b^T b + x^T \Gamma^T \Gamma x \right] \\ &= 2A^T A x - 2A^T b + 2\Gamma^T \Gamma x\end{aligned}$$

Next, we can set $\nabla_x f = 0$ and solve for x

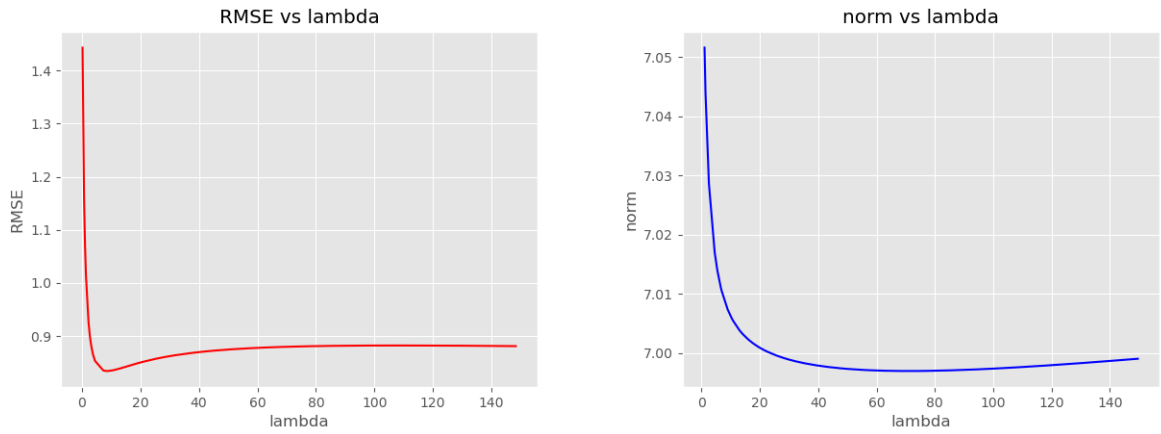
$$2A^T A x - 2A^T b + 2\Gamma^T \Gamma x = 0$$

$$x^* = (A^T A + \Gamma^T \Gamma)^{-1} A^T b$$

Let $\Gamma = \sqrt{\lambda} I$, then we can see that the closed form optimal solution for the ridge regression form is

$$x^* = (A^T A + \lambda I)^{-1} A^T b.$$

Part c:



The optimal regularization parameter is 8.3910.
The RMSE on the validation set with the optimal regularization parameter is 0.8340.
The RMSE on the test set with the optimal regularization parameter is 0.8628.

Part d: To solve for the closed form solution, we first expand the objective function as the following:

$$\begin{aligned} f &= \|Ax + b1 - y\|_2^2 + \|\Gamma x\|_2^2 \\ &= (Ax + b1 - y)^T(Ax + b1 - y) + (\Gamma x)^T(\Gamma x) \\ &= (x^T A^T + b1^T - y^T)(Ax + b1 - y) + x^T\Gamma^T\Gamma x \\ &= x^T A^T Ax + 2b1^T Ax - 2y^T Ax - 2b1^T y + b1^2 n + y^T y + x^T\Gamma^T\Gamma x \end{aligned}$$

Then, we can find the gradient of f with respect to x and b.

$$\nabla_x f = 2A^T Ax + 2bA^T 1 - 2A^T y + 2\Gamma^T\Gamma x$$

$$\nabla_b f = 21^T Ax - 21^T y + 2bn$$

Next, we set $\nabla_b f = 0$ and solve for b^*, which results in

$$b^* = \frac{1^T(y - Ax)}{n}$$

Now, we can plug b^* back to the gradient with respect to x and set it equal to 0 solve for x^*.

$$\begin{aligned} (A^T A + \Gamma^T\Gamma)x + \left(\frac{1^T(y - Ax)}{n}\right) A^T 1 - A^T y = 0 \\ (A^T A + \Gamma^T\Gamma)x + \frac{1}{n}A^T 11^T y - \frac{1}{n}A^T 11^T Ax - A^T y = 0 \\ \left[A^T A + \Gamma^T\Gamma - \frac{1}{n}A^T 11^T A\right] x = A^T y - \frac{1}{n}A^T 11^T y \\ \left[A^T\left(I - \frac{1}{n}11^T\right)A + \Gamma^T\Gamma\right] x = A^T\left(I - \frac{1}{n}11^T\right) y \end{aligned}$$

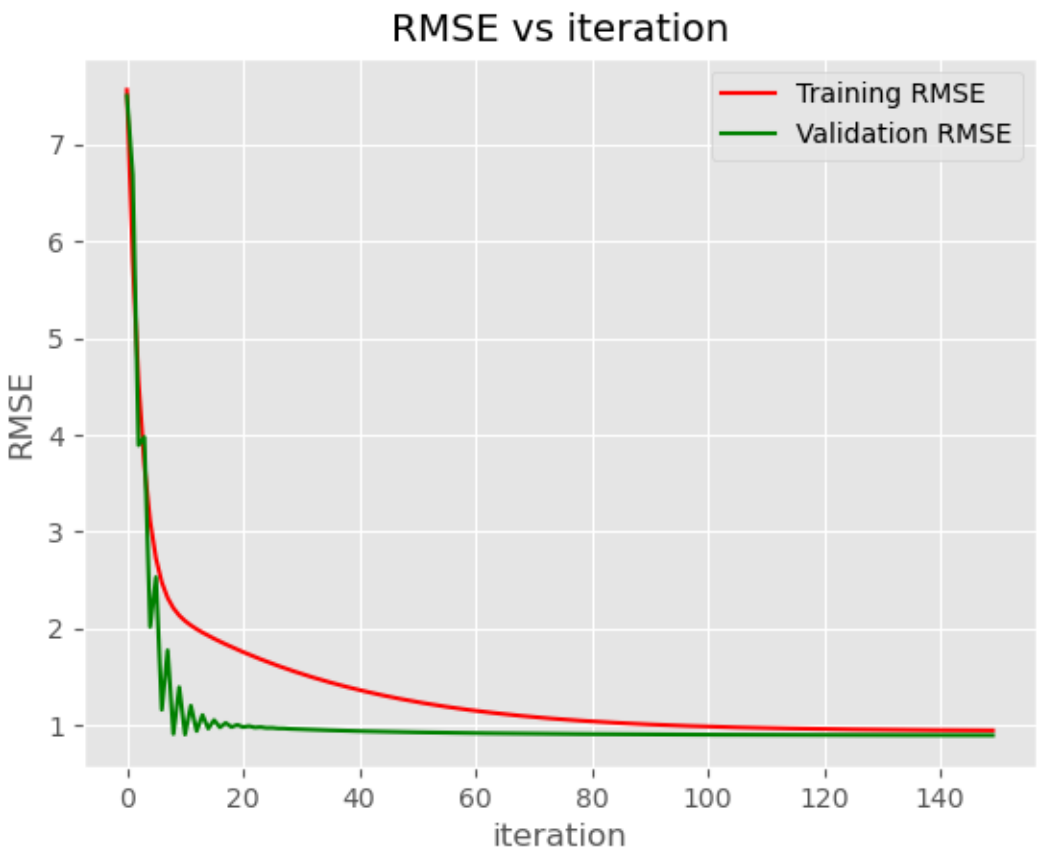
Thus, $x^* = \left[A^T\left(I - \frac{1}{n}11^T\right)A + \Gamma^T\Gamma\right]^{-1} A^T\left(I - \frac{1}{n}11^T\right) y$. where I is the identity matrix and 1 is the one vector.

Then, we can calculate results using the closed form solution and find the difference between results from part c.

Difference in bias is 2.3657E-11
Difference in weights is 3.1597E-11

We can see that the differences are so small that they are negligible.

Part e:



Difference in bias is 1.5388E-01
Difference in weights is 8.0269E-01 ■