Cassidy Lê
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 $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$ where $\mathbf{S}$ =diag$(\mu_1(1-\mu_1), \dots, \mu_n(1-\mu_n))$. 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.

(a) Given $\sigma(x) = \frac{1}{1+e^{-x}}$, we can evaluate the derivative and use some algebra to write it in the form of $\sigma(x)$

$$\begin{aligned}
\sigma'(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{e^x}\left(\frac{e^x}{e^x+1}\right)^2 \\
&= \frac{e^x}{(e^x+1)^2} \\
&= \sigma(x)\left(\frac{1}{e^x+1}\right).
\end{aligned}$$

An alternate form of the sigmoid function is $\sigma(x) = \frac{e^x}{e^x+1}$. If we use this form of $\sigma(x)$ to determine $1 - \sigma(x)$, we get

$$\begin{aligned} 1 - \sigma(x) &= 1 - \frac{e^x}{e^x + 1} \\ &= \frac{e^x + 1}{e^x + 1} - \frac{e^x}{e^x + 1} \\ &= \frac{1}{e^x + 1}. \end{aligned}$$

Therefore,

$$\sigma'(x) = \sigma(x)[1 - \sigma(x)]._{QED}$$

(b) We can start by considering the negative log likelihood for logistic regression:

$$n\ell\ell(\theta) = -\sum_{i=1}^{N} y_i \log \sigma(\theta^T x_i) + (1 - y_i) \log[1 - \sigma(\theta^T x_i)].$$

Then, take the gradient with respect to θ. By derivative rules, we can differentiate each term within the sum, so

$$\nabla_\theta n\ell\ell(\theta) = -\sum_{i=1}^{N} \frac{d}{d\theta} y_i \log \sigma(\theta^T x_i) + \frac{d}{d\theta}(1 - y_i) \log[1 - \sigma(\theta^T x_i)].$$

Because $\frac{d}{dx} \log x = \frac{1}{x}$,

$$= -\sum_{i=1}^{N} \left[\frac{y_i}{\sigma(\theta^T x_i)} - \frac{1 - y_i}{1 - \sigma(\theta^T x_i} \right] \sigma'(\theta^T x_i).$$

From part (a), we know that $\sigma'(x) = \sigma(x)[1 - \sigma(x)]$, so substitute this in to get

$$\begin{aligned} &= -\sum_{i=1}^{N} \left[\frac{y_i}{\sigma(\theta^T x_i)} - \frac{1 - y_i}{1 - \sigma(\theta^T x_i} \right] \sigma(\theta^T x_i)[1 - \sigma(\theta^T x_i)] x_i \\ &= -\sum_{i=1}^{N} \left[\frac{y_i - y_i \sigma(\theta^T x_i) - \sigma(\theta^T x_i) + y_i \sigma(\theta^T x_i)}{\sigma(\theta^T x_i)[1 - \sigma(\theta^T x_i)]} \right] \sigma(\theta^T x_i)[1 - \sigma(\theta^T x_i)] x_i \\ &= -\sum_{i=1}^{N} [y_i - \sigma(\theta^T x_i)] x_i \\ \nabla_\theta n\ell\ell(\theta) &= \sum_{i=1}^{N} [\sigma(\theta^T x_i) - y_i] x_i. \end{aligned}$$

I was able to get to this form. However, I checked my answers against the solutions and I think the question wanted the solution in a different form. So I used the solutions to help me put it in the desired form: Let $\mu = \sigma(\theta^T x_i)$ and x_i is the transpose of the ith row of matrix X. Then,

$$\nabla_\theta n\ell\ell(\theta) = X^T(\vec{\mu} - \vec{y}).$$

(c) The Hessian of the log likelihood for logistic regression is defined as $H\ell\ell(\theta) = \frac{d}{d\theta}\nabla_\theta^T$. Therefore,

$$
\begin{aligned}
H\ell\ell(\theta) &= \frac{d}{d\theta}[X^T(\vec{\mu} - \vec{y}]^T \\
&= \frac{d}{d\theta}(\vec{\mu} - \vec{y})^T X \\
&= \frac{d}{d\theta}\vec{\mu}^T X \\
&= X^T \mathrm{diag}[\vec{\mu}(1 - \vec{\mu})] X \\
H\ell\ell(\theta) &= X^T S X.
\end{aligned}
$$

H_θ is positive semi-definite if S is positive semi-definite. In order for S to be positive semi-definite, its eigenvalues must be non-negative. Since S is a diagonal matrix, its eigenvalues are its diagonal entries, which are determined by the sigmoid function. Since the sigmoid function is bounded above 0 and below 1, we know that

$$\sigma(x)[1 - \sigma(x)] \geq 0.$$

Therefore, S is positive semi-definite, which in turn indicates that H_θ is positive semi-definite.$_{QED}$

■

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.

Since the density $\mathbb{P}(x;\sigma^2)$ must equal 1, it must be true that

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

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

To comput this, consider

$$Z^2 = \int_a^b \int_a^b \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) dxdy.$$

Convert the coordinates from Cartesian to polar coordinates, so $x = r\cos\theta$, $y = r\sin\theta$, and $dxdy = drd\theta$. With this conversion, we know that $x^2 + y^2 = r^2\cos^2\theta + r^2\sin^2\theta = r^2(\cos^2\theta + \sin^2\theta) = r^2$. Therefore, Z^2 is

$$Z^2 = \int_0^{\infty} \int_0^{2\pi} r \cdot \exp\left(-\frac{r^2}{2\sigma^2}\right) drd\theta$$

$$= 2\pi \int_0^{\infty} r \cdot \exp\left(-\frac{r^2}{2\sigma^2}\right) dr.$$

Let $u = \exp\left(-\frac{r^2}{2\sigma^2}\right)$ and $du = -\frac{1}{\sigma^2} r \cdot \exp\left(-\frac{r^2}{2\sigma^2}\right) dr$. Then,

$$Z^2 = 2\pi\Big[-\sigma^2 u\Big]_0^{\infty}$$

$$= -2\pi\sigma^2\left[\exp\left(-\frac{r^2}{2\sigma^2}\right)\right]_0^{\infty}$$

$$= -2\pi\sigma^2(0-1)$$

$$Z^2 = 2\pi\sigma^2$$

Therefore, $\boxed{Z = \sigma\sqrt{2\pi}.}$ ∎

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,

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

is equivalent to the ridge regression problem

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

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 $||^\star||_2$ where 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 $\hat{y} = \theta^T \mathbf{x}$ with $\mathbf{x}_0 = 1$, we compute $\hat{y} = \theta^T \mathbf{x} + b$. 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.

(a) By the probability distribution formula, $\mathcal{N}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Substituting this into the maximum posteriori problem, we get

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w_o - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{w_j^2}{2\tau^2}\right).$$

Re-arranging this using logarithmic rules, we get

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left(-\frac{(y_i - w_o - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2} - \log\sigma\sqrt{2\pi}\right) + \sum_{j=1}^{D} \left(-\frac{w_j^2}{2\tau^2} - \log\sigma\sqrt{2\pi}\right)$$

$$= \arg\max_{\mathbf{w}} -\left((N+D)\log\sigma\sqrt{2\pi} + \sum_{i=1}^{N} \frac{(y_i - w_o - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2} + \sum_{j=1}^{D} -\frac{w_j^2}{2\tau^2}\right).$$

Since $-(N+D)\log\sigma\sqrt{2\pi}$ is a constant, it does not affect or change the optimal values $\mathbf{w}^*$. Under the same rationale, scaling our problem by $2\sigma^2$ will not affect $\mathbf{w}^*$. Thus,

$$= \arg\max_{\mathbf{w}} -\left(\sum_{i=1}^{N} (y_i - w_o - \mathbf{w}^T\mathbf{x}_i)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2\right).$$

Since the maximum of a function is equivalent to the minimum of the negative of that given function,

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - w_o - \mathbf{w}^T\mathbf{x}_i)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2.$$

Let $\lambda = \frac{\sigma^2}{\tau^2}$. Then,

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - w_o - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{D} w_j^2$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - w_o - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda ||\mathbf{w}||_2^2 \cdot_{QED}$$

(b) In order to find the closed form solution $\mathbf{x}^*$ to the ridge regression problem, find the gradient of $||A\mathbf{x} + \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2$ with respect to x, so

$$\nabla_x \left[(A\mathbf{x} + \mathbf{b})^T (A\mathbf{x} + \mathbf{b}) + (\Gamma\mathbf{x})^T (\Gamma\mathbf{x}) \right] = \nabla_x \left[(\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x} \right]$$

$$= \nabla_x \left[\mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b} + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x} \right]$$

$$= 2A^T A\mathbf{x} - 2A^T \mathbf{b} + 2\Gamma^T \Gamma \mathbf{x}.$$

Then, set the gradient equal to zero, which results in

$$(A^T A + \Gamma^T \Gamma)\mathbf{x} = A^T \mathbf{b}$$

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

(Note to the grader: I looked at the solution key for this problem in order to further simplify the solution to the following.) Let $\Gamma = \mathbf{I}\sqrt{\lambda}$. Then, the closed form solution is

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

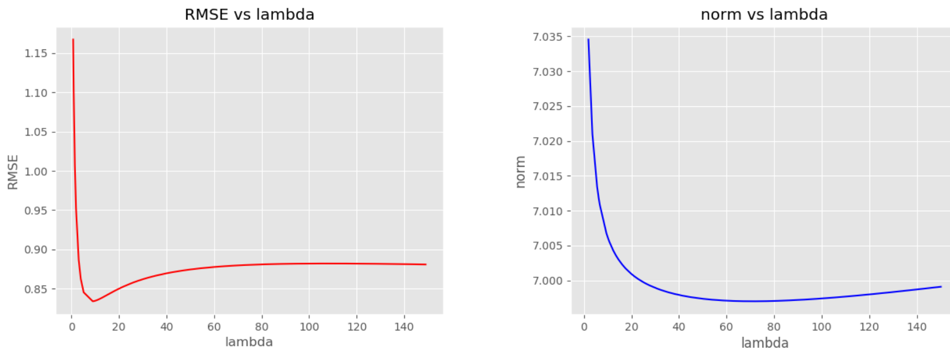

Figure 1: RMSE and Norm Graphs

(c)

(d) (Note to the grader: I had to use the solution for this problem to help assist me in order to finish it.) In order to find the closed form solution $\mathbf{x}^*$ to the ridge regression problem, find the gradient of $||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2$ with respect to x, so

$$\nabla_x \left[(A\mathbf{x} + b\mathbf{1} - \mathbf{y})^T (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x}) \right] = \nabla_x \left[(\mathbf{x}^T A^T + b\mathbf{1}^T - \mathbf{y}^T)(A\mathbf{x} - b\mathbf{1} - \mathbf{y}) + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} \right]$$

$$= \nabla_x \left[\mathbf{x}^T A^T A\mathbf{x} + 2b\mathbf{1}^T A\mathbf{x} - 2\mathbf{1}^T\mathbf{y} + b^2 n + \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} \right]$$

$$= 2A^T A\mathbf{x} + 2bA^T\mathbf{1} - 2A^T\mathbf{y} + 2\Gamma^T\Gamma\mathbf{x}.$$

Then, set the gradient equal to zero, which results in

$$0 = 2A^T A\mathbf{x} + 2bA^T\mathbf{1} - 2A^T\mathbf{y} + 2\Gamma^T\Gamma\mathbf{x}$$

$$= 2\mathbf{1}^T A\mathbf{x} - 2\mathbf{1}^T\mathbf{y} + 2bn.$$

(Used the solution to help finish the rest of the problem.) Solve this for b^* to get

$$b^* = \frac{\mathbf{1}^T(\mathbf{y} - A\mathbf{x})}{n}.$$

Then, substitute this into the equation where the gradient is set equal to zero, so

$$0 = (A^T A + \Gamma^T\Gamma)\mathbf{x} + \left(\frac{\mathbf{1}^T(\mathbf{y} - A\mathbf{x})}{n} \right) A^T\mathbf{1} - A^T\mathbf{y}$$

$$= (A^T A + \Gamma^T\Gamma)\mathbf{x} + \frac{1}{n} A^T\mathbb{1}\mathbb{1}^T - \frac{1}{n} A^T\mathbb{1}\mathbb{1}^T A - A^T$$

$$\left[A^T A + \Gamma^T\Gamma - \frac{1}{n} A^T\mathbb{1}\mathbb{1}^T A \right]\mathbf{x} = A^T - \frac{1}{n} A^T\mathbb{1}\mathbb{1}^T A$$

$$\left[A^T\left(\mathbf{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)A+\Gamma^T\Gamma\right]\mathbf{x}=A^T\left(\mathbf{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)$$

$$\mathbf{x}^*=\left[A^T\left(\mathbf{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right)A+\Gamma^T\Gamma\right]^{-1}A^T\left(\mathbf{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^T\right),$$

where $\mathbf{I}$ is the identity matrix, $\mathbb{1}$ is a vector of all ones, and $y \in \mathbb{R}^n$.
When computing the bias term of the previous problem using this close form, we get the same results.

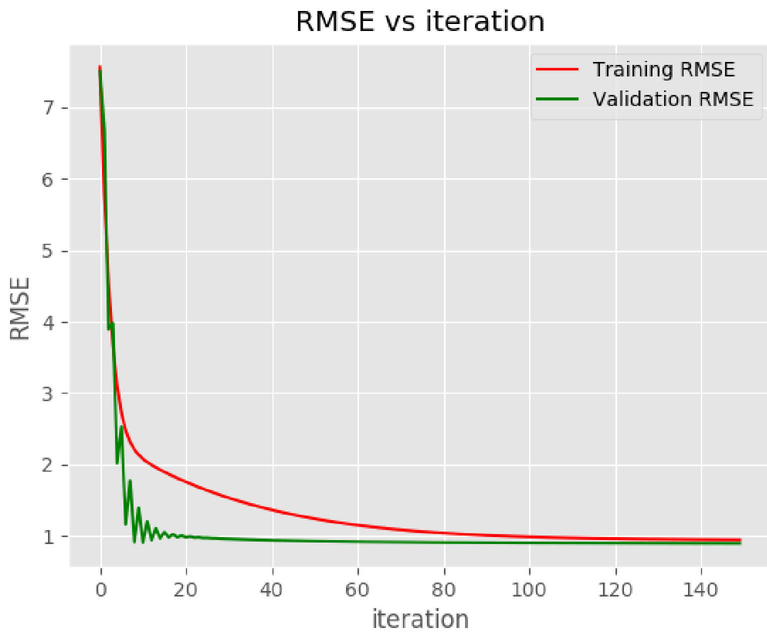

Figure 2: Convergence Graph

(e)

■