Arjun Natarajan
Math189R SP 19
Homework 2
Monday, February 10 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 $\mathbf{S} = \mathrm{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. 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.

a) so, $\sigma(x) = (1+e^{-x})^{-1}$.
Then,

$$\sigma'(x) = -1(1+e^{-x})^{-2} \cdot (-e^{-x})$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(\frac{e^{-x}}{1+e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$

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

■

b) So,

$$NLL(\theta) = -\sum_{i=1}^{N}\left[y_i \log\sigma(\theta^T x_i) + (1-y_i)\log(1-\sigma(\theta^T x_i))\right]$$

$$\nabla NLL(\theta) = -\sum_{i=1}^{N} y_i \cdot \frac{1}{\sigma(\theta^T x_i)}\sigma'(\theta^T x_i) + (1-y_i)\cdot\frac{1}{1-\sigma(\theta^T x_i)}\cdot -\sigma'(\theta^T x_i)$$

$$= -\sum_{i=1}^{N} \frac{y_i}{\sigma(\theta^T x_i)}\left(\sigma(\theta^T x_i)\left[1-\sigma(\theta^T x_i)\right]\right)x_i - \frac{(1-y_i)}{1-\sigma(\theta^T x_i)}\sigma(\theta^T x_i)\left(1-\sigma(\theta^T x_i)\right)x_i$$

$$= -\sum_{i=1}^{N} (y_i - y_i\sigma(\theta^T x_i) - \sigma(\theta^T x_i) + y_i\sigma(\theta^T x_i))x_i$$

$$= -\sum_{i=1}^{N} (y_i - \mu_i)x_i, \quad \text{where } \mu_i = \sigma(\theta^T x_i)$$

$$= X^T(\mu - y).$$

c) The Hessian is

$$\nabla^2 NLL(\theta) = \nabla_\theta \sum_{i=1}^{N}(\mu_i - y_i)x_i$$

$$= \sum_{i=1}^{N} \nabla_\theta(\sigma(\theta^T x_i) - y_i)x_i$$

$$= \sum_{i=1}^{N} \nabla_\theta(\sigma(\theta^T x_i))x_i$$

$$= \sum_{i=1}^{N} \sigma(\theta^T x_i)\left(1-\sigma(\theta^T x_i)\right)x_i x^T$$

$$= X^T \operatorname{diag}(\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))) X$$

$$= X^T S X,$$

where $S \triangleq \operatorname{diag}(\mu(1-\mu))$.

To show that the Hessian is positive semidefinite, we only need to show that S is. Since S is a diagonal matrix, its eigenvalues are just the diagonal entries, which are $\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))$,

and recall that σ is bounded by 0 and 1. Thus, it must be the case that each entry of S is greater than or equal to 0, so the Hessian must be positive semidefinite.

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.

So, we know that

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

Thus,

$$\frac{1}{Z}\int_{-\infty}^{\infty} e^{\frac{-x^2}{2\sigma^2}}\,dx = 1$$

$$Z = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\,dx.$$

Consulted the solutions here.

Then,

$$Z^2 = \iint_{\mathbb{R}^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}\,dx\,dy$$

$$= \int_0^{2\pi}\int_0^{\infty} e^{\frac{-r^2}{2\sigma^2}}\,r\,dr\,d\theta$$

$$= 2\pi\int_0^{\infty} e^{\frac{-r^2}{2\sigma^2}}\,r\,dr$$

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

$$= -2\pi\sigma^2\, e^{\frac{-r^2}{2\sigma^2}}\Big|_0^{\infty}$$

$$= 2\pi\sigma^2.$$

■

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

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,

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

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)

consulted solutions

■

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.

a) So, knowing that $N(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, our expression becomes

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

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

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

$$= \underset{w}{\text{argmax}} -\left[(N+D)(\log\sigma\sqrt{2\pi}) + \sum_{i=1}^{N} \frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}\right].$$

Since we are maximizing we can minimize the negative, and constants and scaling will not matter. So, we can write this as

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

$$= \underset{w}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda ||w||_2^2 .$$

■

b) Consider

$$\nabla_x f = \nabla_x \left[(Ax-b)^T (Ax-b) + (\Gamma x)^T (\Gamma x) \right]$$

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

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

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

Since we are minimizing,

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

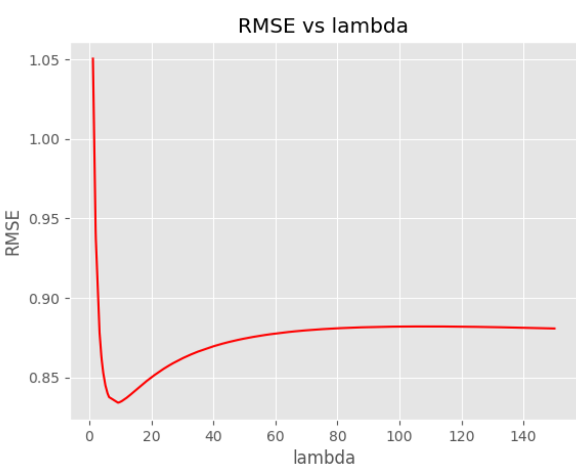
$$0 = 2A^T A x - 2A^T b + 2\Gamma^T \Gamma x,$$

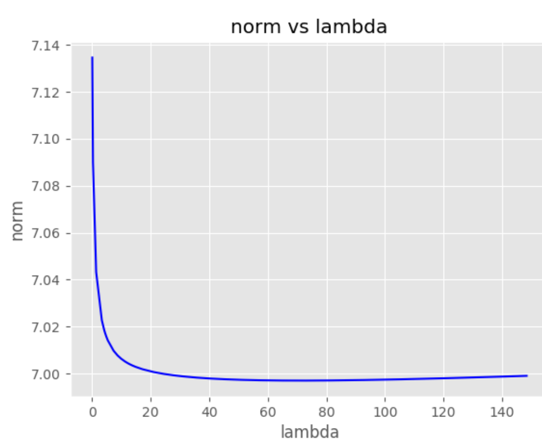
because $A^T b$ is symmetric, so

$$A^T b = (A^T A + 2\Gamma^T \Gamma) x$$

$$\boxed{x = (A^T A + 2\Gamma^T \Gamma)^{-1} A^T b} .$$

c) The optimal regularization parameter $\lambda^* = 9.2629$. The RMSE on the validation set with λ^* was 0.8342 and it was 0.8628 on the test set. Here are the plots





d) Expanding this yields,

$$f = (Ax + b\mathbb{1} - y)^T (Ax + b\mathbb{1} - y) + (\Gamma x)^T (\Gamma x)$$

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

$$= x^T A^T A x + x^T A^T b\mathbb{1} - x^T A^T y + b\mathbb{1}^T A x + b\mathbb{1}^T b\mathbb{1} - b\mathbb{1}^T y - y^T A x - y^T b\mathbb{1} + y^T y + x^T \Gamma^T \Gamma x$$

$$= x^T A^T A x + 2b\mathbb{1}^T A x - 2y^T A x - 2b\mathbb{1}^T y + b^2 n + y^T y + x^T \Gamma^T \Gamma x.$$

Minimizing this by setting $\nabla f_x, \nabla f_b = 0$ yields

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

and

$$2)\ 2\mathbb{1}^T A x - 2\mathbb{1}^T y + 2b^* n = 0$$

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

substituting b^* into 1),

$$A^T A x^* + \frac{\mathbb{1}^T (y - Ax)}{n} A^T \mathbb{1} - A^T y + \Gamma^T \Gamma x^* = 0$$

$$(A^T A + \Gamma^T \Gamma) x^* + \frac{1}{n} A^T \mathbb{1}\mathbb{1}^T y - \frac{1}{n} A^T \mathbb{1}\mathbb{1}^T A x^* - A^T y = 0$$

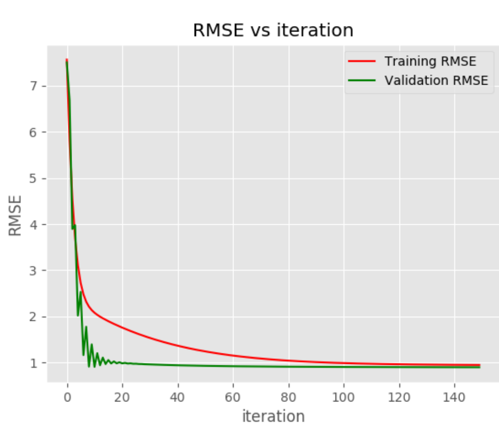
$$(A^T A + \Gamma^T \Gamma - \frac{1}{n} A^T \mathbb{1}\mathbb{1}^T A) x^* = A^T y - \frac{1}{n} A^T \mathbb{1}\mathbb{1}^T y$$

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

$$\underline{x^* = (A^T (I - \frac{1}{n}\mathbb{1}\mathbb{1}^T) A + \Gamma^T \Gamma)^{-1} A^T (I - \frac{1}{n}\mathbb{1}\mathbb{1}^T) y}.$$

when computing the difference in bias is 4.3644×10^{-10} and difference in weights is 5.765×10^{-10}. These are negligble, so both results are the same.

e)



The difference in bias is 1.53888×10^{-1}

The difference in weights is 8.0663×10^{-1}.