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.

a)

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

b)

Logistic regression expression for gradient of log-likelihood is :

$$n\ell\ell(\boldsymbol{\theta}) = -\sum_i y_i \log \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) + (1-y_i)\log\left(1-\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)$$

Now just take the gradients using theta

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}^{n\ell\ell}(\boldsymbol{\theta})} &= -\sum_i y_i \frac{1}{\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)}\sigma'\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) + (1-y_i)\frac{1}{1-\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)}\left(-\sigma'\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right) \\
&= -\sum_i y_i\left(1-\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)\mathbf{x}_i - (1-y_i)\,\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i \\
&= -\sum_i y_i \mathbf{x}_i - y_i\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i - \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i + y_i\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\mathbf{x}_i \\
&= \sum_i \left(\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) - y_i\right)\mathbf{x}_i \\
&= X^\top(\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right) - \mathbf{y})
\end{aligned}$$

Each $\mathbf{x}_i$ is the i th column of X transposed.

C) Using prev results, find Hessian matrix

$$\begin{aligned}
\mathbf{H}_{\boldsymbol{\theta}} &= \nabla_{\boldsymbol{\theta}}\left(\nabla_{\boldsymbol{\theta}} n\ell\ell(\boldsymbol{\theta})\right)^\top = \nabla_{\boldsymbol{\theta}}\left[X^\top(\boldsymbol{\mu}-\mathbf{y})\right]^\top \\
&= \nabla_{\boldsymbol{\theta}}\left(\boldsymbol{\mu}^\top X - \mathbf{y}^\top X\right) \\
&= \nabla_{\boldsymbol{\theta}}\boldsymbol{\mu}^\top X = \nabla_{\boldsymbol{\theta}}\sigma(X\boldsymbol{\theta})^\top X \\
&= X^\top \operatorname{diag}(\boldsymbol{\mu}(1-\boldsymbol{\mu})X \\
&= X^\top S X
\end{aligned}$$

To prove this is semi-definite, we to prove that S's eigenvalues are nonnegative. S is a diagonal matrix, meaning its eigenvalues are its diagonal values. We can show this by showing $\boldsymbol{\mu}_i\left(1-\boldsymbol{\mu}_i\right) >= 0$. We can assert that: $\boldsymbol{\mu}_i\left(1-\boldsymbol{\mu}_i\right) = \sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\left(1-\sigma\left(\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)$. $\sigma(\cdot)(1-\sigma(\cdot))$ is always nonnegative since $\sigma(\cdot)$ is between 0 and 1. This shows that the Hessian matrix is positive 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.

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

which shows

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

Consider Z^2

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

$$Z^2 = 2\pi\sigma^2$$
$$Z = \sqrt{2\pi\sigma^2} = \sqrt{2\pi}\sigma$$

Probability density function integrate to 1 so becomes a valid density.

■

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.

a)

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

After applying the probability distribution $\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$,we get

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{w_j^2}{2\tau^2}\right)$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left(-\frac{\left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma\right) + \sum_{j=1}^{D} \left(-\frac{w_j^2}{2\tau^2} - \log\sqrt{2\pi}\sigma\right)$$

$$= \arg\max_{\mathbf{w}} - \left((N+D)\log\sqrt{2\pi}\sigma + \sum_{i=1}^{N} \frac{\left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}\right)$$

To maximize function, we minimize its negative.

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2$$

Replace λ with σ^2/τ^2

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \mathbf{w}^\top \mathbf{x}_i\right)^2 + \lambda \sum_{j=1}^{D} w_j^2$$

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

b) Find 0 value of gradient of f with respect to $\mathbf{x}$ in order to minimize $f = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \|\Gamma\mathbf{x}\|_2^2$

$$\begin{aligned}
\nabla_{\mathbf{x}} f &= \nabla_{\mathbf{x}} \left[(A\mathbf{x} - \mathbf{b})^\top (A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x})\right] \\
&= \nabla_{\mathbf{x}} \left[\left(\mathbf{x}^\top A^\top - \mathbf{b}^\top\right)(A\mathbf{x} - \mathbf{b}) + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x}\right] \\
&= \nabla_{\mathbf{x}} \left[\mathbf{x}^\top A^\top A\mathbf{x} - 2\mathbf{x}^\top A^\top \mathbf{b} + \mathbf{b}^\top \mathbf{b} + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x}\right] \\
&= 2A^\top A\mathbf{x} - 2A^\top \mathbf{b} + 2\Gamma^\top \Gamma\mathbf{x} \\
\nabla_{\mathbf{x}} f(0) &= \left(A^\top A + \Gamma^\top \Gamma\right)\mathbf{x} = A^\top \mathbf{b}
\end{aligned}$$

Closed form solution: $\mathbf{x}^\star = \left(A^\top A + \Gamma^\top \Gamma\right)^{-1} A^\top \mathbf{b}$

c)

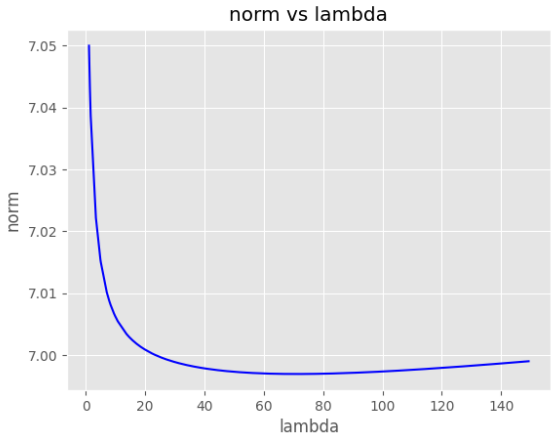

Figure 1: Norm vs Lambda

The optimal regularization parameter is 8.9946. The RMSE on the validation set with the optimal regularization parameter is 0.8341. The RMSE on the test set with the optimal regularization parameter is 0.8628.

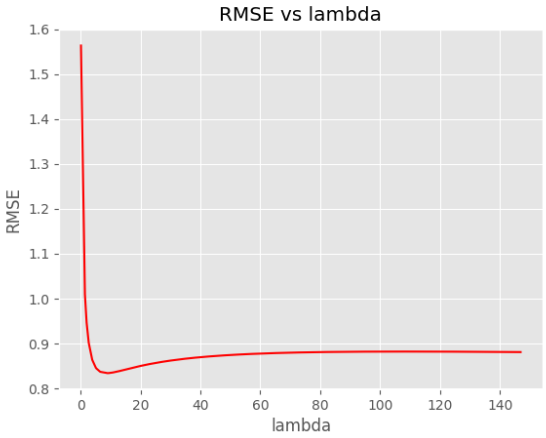

Figure 2: RMSE vs Lambda

d)

$$
\begin{aligned}
f &= \|A\mathbf{x} + b\mathbf{1} - \mathbf{y}\|_2^2 + \|\Gamma\mathbf{x}\|_2^2 \\
&= (A\mathbf{x} + b\mathbf{1} - \mathbf{y})^\top (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x}) \\
&= \left(\mathbf{x}^\top A\top + b\mathbf{1}^\top - \mathbf{y}^\top\right)(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x} \\
&= \mathbf{x}^\top A\top A\mathbf{x} + 2b\mathbf{1}^\top A\mathbf{x} - 2\mathbf{y}^\top A\mathbf{x} - 2b\mathbf{1}^\top\mathbf{y} + b^2 n + \mathbf{y}^\top\mathbf{y} + \mathbf{x}^\top\Gamma^\top\Gamma\mathbf{x}
\end{aligned}
$$

We know at Optimality is when$\nabla_{\mathbf{x}} f = 0$, so find when $\nabla_{\mathbf{b}} f$ is 0.

$$
\begin{aligned}
\nabla_{\mathbf{x}} f &= 2A^\top A\mathbf{x} + 2bA^\top\mathbf{1} - 2A^\top\mathbf{y} + 2\Gamma^\top\Gamma\mathbf{x} \\
\nabla_b f &= 2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top\mathbf{y} + 2bn \\
0 &= 2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top\mathbf{y} + 2bn
\end{aligned}
$$

$$
b^\star = \frac{\mathbf{1}^\top(\mathbf{y} - A\mathbf{x})}{n}
$$

Shows bias term is the mean value of output when x = 0.

Plugging $b^\star$ back to equation to solve for $\mathbf{x}^\star$.

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

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

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

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

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

where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is a vector of all ones, and $\mathbf{y} \in \mathbf{R}^n$.

Bias from sample code and part(c) are extremely similar

Difference in bias is 1.9661E-11
Difference in weights is 2.3683E-10

e)

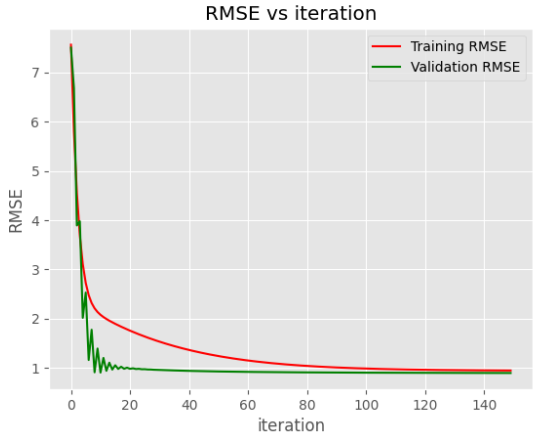

Figure 3: Convergence Plot

Difference in bias is 1.5386E-01
Difference in weights is 7.9580E-01

■