Patrick Hutecker
Math189R SP19
Homework 2
Monday, Feb 11, 2019

CONSULTED SOLUTIONS FOR $1.b, 1.c, 3.a - e$.

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)

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

b) Then negative log likelihood for logistic regression is

$$NLL(\mathbf{w}) = -\sum_{i=1}^{N} y_i \log(\sigma(\mathbf{w}^\top \mathbf{x}_i)) + (1-y_i)\log(1-\sigma(\mathbf{w}^\top \mathbf{x}_i)),$$

it then follows that

$$\begin{aligned}
\nabla_{\mathbf{w}} NLL(\mathbf{w}) &= -\sum_{i=1}^{N} y_i[1-\sigma(\mathbf{w}^T\mathbf{x}_i)]\mathbf{x}_i - (1-y_i)\sigma(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i \\
&= -\sum_{i=1}^{N} y_i\mathbf{x}_i - y_i\sigma(\mathbf{w}^\top\mathbf{x}_i)\mathbf{x}_i - \sigma(\mathbf{w}^\top\mathbf{x}_i)\mathbf{x}_i + y_i\sigma(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i \\
&= \sum_{i=1}^{N}(\sigma(\mathbf{w}^\top\mathbf{x}_i) - y_i)\mathbf{x}_i. \\
&= \sum_{i=1}^{N}(\mu_i - y_i)\mathbf{x}_i.
\end{aligned}$$

It is also worth mentioning that this form can be further reduced into dot products of vectors - but this form is much clearer.

c) The formula for the Hessian is

$$\mathbf{H} = \nabla_{\mathbf{w}}(\nabla_{\mathbf{w}} NLL(\mathbf{w}))^\top,$$

let us substitute the formula we got for $\nabla_{\mathbf{w}} NLL(\mathbf{w})$ in part *b*. We get

$$\begin{aligned}
\mathbf{H} &= \nabla_{\mathbf{w}}(\sum_{i=1}^{N} \mathbf{x}_i^\top(\mu_i^\top - y_i^\top)) \\
&= \sum_{i=1}^{N} \mathbf{x}_i^\top(\mu_i(1-\mu))^\top \mathbf{x}_i \\
&= \sum_{i=1}^{N} \mathbf{x}_i^\top(\mu_i(1-\mu))\mathbf{x}_i \\
&= \mathbf{X}^\top\mathbf{S}\mathbf{X}.
\end{aligned}$$

As desired. Now we wish to show that $\mathbf{H}$ or rather $\mathbf{X}^\top\mathbf{S}\mathbf{X}$ is positive semidefinite.

We know by the textbook that we can assume that $0 < \mu_i < 1$ for all i. The elements of $\mathbf{S}$ are then strictly positive. By definition, $\mathbf{S}$ is then positive semidefinite - as it is a diagonal of positive values.

Now, consider $\mathbf{H}$ or $\mathbf{X}^\top\mathbf{S}\mathbf{X}$. Consider v_1 which is some column vector of the appropriate dimension. Then, consider

$$\mathbf{v}_1^\top(\mathbf{X}^\top\mathbf{S}\mathbf{X})\mathbf{v}_1.$$

It follows that $\mathbf{X}\mathbf{v}_1$ can be expressed as some other column vector $\mathbf{v}_2$. Furthermore, it then follows that $\mathbf{v}_2^\top = \mathbf{v}_1^\top \mathbf{X}^\top$. We can substitute $\mathbf{v}_2$ into our original expression giving

$$\mathbf{v}_2^\top \mathbf{S} \mathbf{v}_2,$$

since $\mathbf{v}_2$ is some column vector of the appropriate dimension and since we have shown that $\mathbf{S}$ is positive semidefinite it must follow that $\mathbf{v}_2^\top \mathbf{S} \mathbf{v}_2 \geq 0$, but by our original expression this implies that $\mathbf{v}_1^\top \mathbf{H} \mathbf{v}_1 \geq 0$ for an arbitrary column vector $\mathbf{v}_1$. But this is the definition of positive semidefinite - showing that $\mathbf{H} \succeq 0$.

$\blacksquare$

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.

For $\mathbb{P}(x;\sigma^2)$ to be a valid density, it must follow that $\int_{-\infty}^{\infty}\mathbb{P}(x;\sigma^2)dx = 1$. Let us simplify the left hand side and solve for Z

$$\begin{aligned}
\mathbb{P}(x;\sigma^2) &= \int_{-\infty}^{\infty}\frac{1}{Z}\exp(-\frac{-x^2}{2\sigma^2})dx \\
&= \sqrt{(\int_{-\infty}^{\infty}\frac{1}{Z}\exp(-\frac{x^2}{2\sigma^2})dx)(\int_{-\infty}^{\infty}\frac{1}{Z}\exp(-\frac{y^2}{2\sigma^2})dy)} \\
&= \frac{1}{Z}\sqrt{\int_{\infty}^{\infty}\int_{\infty}^{\infty}\exp(-\frac{x^2+y^2}{2\sigma^2})dxdy} \\
&= \frac{1}{Z}\sqrt{\int_0^{2\pi}d\theta\int_0^{\infty}r\exp(\frac{-r^2}{2\sigma^2})dr} \\
&= \frac{1}{Z}\sqrt{2\pi\sigma^2(\exp\frac{-r^2}{2\sigma^2}\Big|_0^{\infty})} \\
&= \frac{1}{Z}\sqrt{2\pi\sigma^2(1-0))} \\
&= \frac{1}{Z}\sqrt{2\pi}\sigma.
\end{aligned}$$

It then follows that $\frac{1}{Z}\sqrt{2\pi}\sigma = 1$, meaning that $Z = \sqrt{2\pi}\sigma$ for $\mathbb{P}(x;\sigma^2)$ to be 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)

(a) We know that

$$\mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}),$$

substituting into the expression gives

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

Since we are looking to maximize this expression in terms of $\mathbf{w}$ it is pointless to consider terms which are not dependant of $\mathbf{w}$ as they will simply contribute a scaling in our maximization. We can therefore write

$$
\begin{aligned}
&= \arg\max_{\mathbf{w}} -(\sum_{i=1}^{N} \frac{(y_i - w_0 + \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} - \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}) \\
&= \frac{1}{2\sigma^2} \arg\max_{\mathbf{w}} -(\sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^\top \mathbf{x}_i))^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2).
\end{aligned}
$$

As mentioned before, the term $\frac{1}{2\sigma^2}$ does not affect the maximization as it simply scales the solution. Furthermore, the maximum value of a negated expression is equivalent to the minimum value of the original expression. We can also substitute in $\lambda = \frac{\sigma^2}{\tau^2}$ and the inner product for $\sum_{j=1}^{D} w_j^2$ giving

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

as desired.

(b) We can minimize the expression by taking the gradient and using the transpose expression of the inner product

$$
\begin{aligned}
\nabla_{\mathbf{x}}(||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2) &= \nabla_{\mathbf{x}}((A\mathbf{x} - \mathbf{b})^\top (A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x})) \\
&= \nabla_{\mathbf{x}}((\mathbf{x}^\top A^\top - \mathbf{b}^\top)(A\mathbf{x} - \mathbf{b}) + \mathbf{x}^\top \Gamma^\top (\Gamma\mathbf{x})) \\
&= \nabla_{\mathbf{x}}(\mathbf{x}^\top A^\top A\mathbf{x} - \mathbf{b}^\top A\mathbf{x} - \mathbf{x}^\top A^\top \mathbf{b} + \mathbf{b}^\top \mathbf{b} + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x}) \\
&= 2A^\top A\mathbf{x} - 2A^\top \mathbf{b} + 2\Gamma^\top \Gamma\mathbf{x}.
\end{aligned}
$$

To find the minimum we must find the roots of this expression meaning that

$$
2A^\top A\mathbf{x} - 2A^\top \mathbf{b} + 2\Gamma^\top \Gamma\mathbf{x} = 0,
$$

which gives

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

or

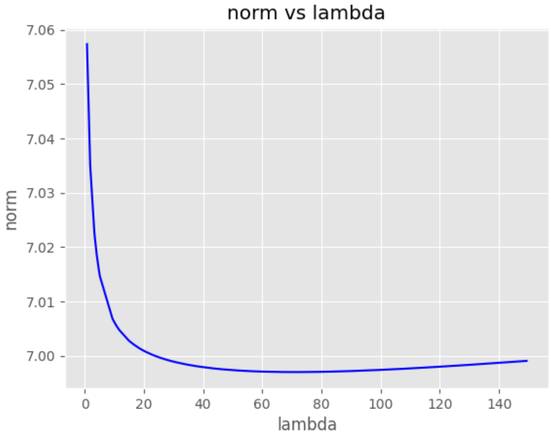
$$\mathbf{x}^* = (A^\top A + \Gamma^\top \Gamma)^{-1} A^\top \mathbf{b}.$$

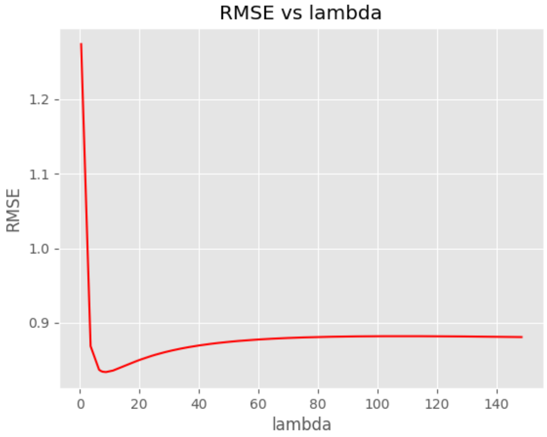
(c) The optimal regularization parameter is $\lambda^* = 8.7970$.

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.

The plots are:





■

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.

(d) To find the optimal solution - we must find the roots of the gradient of the expression. Consider

$$\begin{aligned}
&= \nabla_{\mathbf{x}}((\mathbf{x}^\top A^\top + \mathbf{1}b^\top - \mathbf{y}^\top)(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x}) \\
&= \nabla_{\mathbf{x}}(\mathbf{x}^\top A^\top A\mathbf{x} + \mathbf{1}b^\top A\mathbf{x} - \mathbf{y}^\top A\mathbf{x} + \mathbf{x}^\top A^\top b\mathbf{1} + b^2 n - y^\top b\mathbf{1} - \mathbf{x}^\top A^\top \mathbf{y} - \mathbf{1}b^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x}) \\
&= 2A^\top A\mathbf{x} + 2bA^\top \mathbf{1} - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x}.
\end{aligned}$$

and

$$\begin{aligned}
&= \nabla_b(\mathbf{x}^\top A^\top A\mathbf{x} + \mathbf{1}b^\top A\mathbf{x} - \mathbf{y}^\top A\mathbf{x} + \mathbf{x}^\top A^\top b\mathbf{1} + b^2 n - y^\top b\mathbf{1} - \mathbf{x}^\top A^\top \mathbf{y} - \mathbf{1}b^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \Gamma^\top \Gamma \mathbf{x}) \\
&= 2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top \mathbf{y} + 2bn.
\end{aligned}$$

If we solve for $b*$ we get

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

Plugging this into our expression for the gradient in terms of $\mathbf{x}$ we get

$$\begin{aligned}
0 &= (A^\top A + \Gamma^\top \Gamma)\mathbf{x} + (\frac{\mathbf{1}^\top(\mathbf{y} - A\mathbf{x})}{n})A^\top \mathbf{1} - A^\top \mathbf{y} \\
&= (A^\top A + \Gamma^\top \Gamma)\mathbf{x} + \frac{1}{n}A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} - \frac{1}{n}A^\top \mathbf{1}\mathbf{1}^\top A\mathbf{x} - A^\top \mathbf{y}
\end{aligned}$$

which can then be expressed as

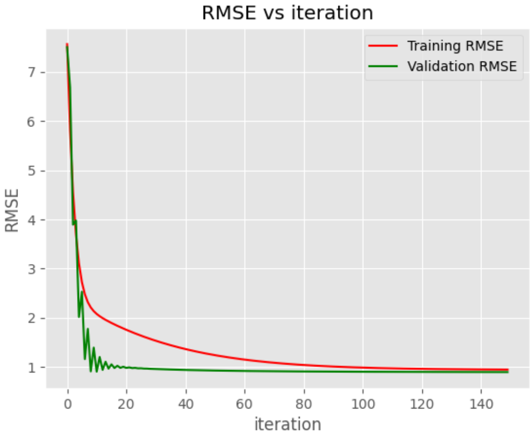
$$(A^\top A + \Gamma^\top \Gamma - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top A)\mathbf{x} = (A^\top - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top)\mathbf{y}$$

$$\mathbf{x}^* = (A^\top A + \Gamma^\top \Gamma - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top A)^{-1}(A^\top - \frac{1}{n} A^\top \mathbf{1}\mathbf{1}^\top)\mathbf{y}.$$

After computing the bias term for the previous problem we get a small error.

Difference in bias is 3.6555E-11

Difference in weights is 6.6146E-11



(e)

Difference in bias is 1.5387E-01

Difference in weights is 7.9798E-01

■