I used the keys to check my work. Red * next to anywhere I used it to fix mistakes.

Eli Weissler

Math189R SU17
Homework 2
Monday, May 22, 2017

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

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

(b) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression. what is X? → X and y or multid $\vec{X}$?

(c) The Hessian can be written as $\mathbf{H} = \mathbf{X}^\top \mathbf{S} \mathbf{X}$ where $\mathbf{S} = \operatorname{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. 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) $\sigma'(x) = \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = -(1+e^{-x})^{-2} \cdot \frac{d}{dx}[1+e^{-x}] = (1+e^{-x})^{-2} \cdot e^{-x}$

Now note that $[1-\sigma(x)] = 1 - \frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}}$

$\rightarrow \sigma'(x) = e^{-x}(1+e^{-x})^{-1} \cdot (1+e^{-x})^{-1} = \boxed{\sigma(x)[1-\sigma(x)]}$

b) Looking at any individual data point $^{(i)}$, its likelihood is given by

$\text{Likelihood} = \sigma(\vec{\theta}^T \vec{x}^{(i)})^{y_i} (1 - \sigma(\vec{\theta}^T \vec{x}^{(i)}))^{1-y_i}$

If we let $\sigma(\vec{\theta}^T \vec{x}^{(i)}) \equiv \mu_i$ (i.e. the mean or 'chance of doing it' for that data point)

Then $\text{likelihood} = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$

And the negative log likelihood is given by

$$NLL = -\left(y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\right)$$

The NLL of the whole distribution is additive (likelihood is mult), so

$$NLL_{all} = -\sum_1^N y_i \log \mu_i + (1-y_i)\log(1-\mu_i)$$

To find the gradient, consider $\frac{\partial NLL}{\partial \theta_j}$

$$\frac{\partial NLL}{\partial \theta_j} = -\sum_1^N \frac{y_i}{\mu_i}\frac{\partial \mu_i}{\partial \theta_j} - \frac{1-y_i}{1-\mu_i}\frac{\partial \mu_i}{\partial \theta_j} = -\sum_1^N \frac{\partial \mu_i}{\partial \theta_j}\left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right)$$

Now note that

$$\frac{\partial \mu_i}{\partial \theta_j} = \frac{\partial \sigma(\sum \theta_i x_i^{(i)})}{\partial \theta_j} = x_j^{(i)} \sigma'(\theta^T x^{(i)}) = x_j^{(i)} \mu_i (1-\mu_i)$$

Plugging back in yields

$$\frac{\partial NLL}{\partial \theta_j} = -\sum x_j^{(i)} \mu_i(1-\mu_i)\left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right) = -\sum x_j^{(i)}[(1-\mu_i)y_i - \mu_i(1-y_i)]$$

$$= -\sum x_j^{(i)}[y_i - \mu_i y_i - \mu_i + \mu_i y_i] = -\sum x_j^{(i)}[y_i - \mu_i]$$

So the gradient is given by

$$\boxed{\nabla NLL = \vec{X}^T \cdot (\vec{\mu} - \vec{y})}$$

c) Now to get the hessian, consider

$$\frac{\partial}{\partial \theta_k}\left(\frac{\partial NLL}{\partial \theta_j}\right) = \sum x_j^{(i)} \frac{\partial \mu_i}{\partial \theta_k} = \boxed{\sum x_j^{(i)} x_k^{(i)} \mu_i(1-\mu_i) = X^T S X \doteq H}$$

H must be positive & semi-definite because it is diagonalizable, with $\lambda = \text{diag}(S) = (\mu_1(1-\mu_1) \ldots \mu_i(1-\mu_i) \ldots)$, which by definition are all between 0 and 1.

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.

Consider an x gausian multiplied by a y gaussian

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

$\downarrow$

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

Now switch to Polar coordinates $(r^2 = x^2+y^2,\ \tan\theta = \frac{y}{x})$

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

$$= -2\pi\sigma^2 \frac{d}{dr}\int_0^{\infty} \exp\left(\frac{-r^2}{2\sigma^2}\right)dr = -2\pi\sigma^2\left(\underbrace{\exp\left(\frac{-\infty^2}{2\sigma^2}\right)}_{0} - 1\right)$$

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

$\downarrow$

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

■

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

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

typo? ✓

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) First, recall the definition of Normal Dist:

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

Therefore, fully written out, the problem is

$$\arg\max \sum^{N \leftarrow \#\text{ of pts}} \log\left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{(y_i - w_0 - W^T x^{(i)})^2}{2\sigma^2}\right)}\right] + \sum^{D \leftarrow \#\text{ of dimensions}} \log \frac{1}{\sqrt{2\pi}\,\tau} e^{-\left(\frac{w_j^2}{2\tau^2}\right)}$$

$$= \arg\max \sum^{N} -(y_i - (w_0 + W^T x^{(i)}))^2 / 2\sigma^2 - \log(\sqrt{2\pi}\,\sigma) + \sum - \left(\frac{w_j^2}{2\tau^2}\right) - \log(\sqrt{2\pi}\,\tau)$$

We can get rid of additive & multiplicative constants without affecting the W that maximizes

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

$$= \boxed{\arg\min \sum_{i=1}^{N} (y_i - (w_0 + W^T x^{(i)}))^2 + \frac{\sigma^2}{\tau^2}|W|^2}$$

b)

$$\|A\vec{x} - \vec{b}\|^2 + \|\Gamma\vec{x}\|^2 = \underbrace{(A\vec{x}-\vec{b})^T}_{(\vec{x}^T A^T - \vec{b}^T)} \cdot (A\vec{x}-\vec{b}) + \underbrace{(\Gamma\vec{x})^T}_{(\vec{x}^T\Gamma^T)} \cdot \Gamma x = J(\vec{x})$$

A: feature coeffs; $\vec{x}$: coeffs; $\vec{b}$: y's; Γ: penalty matrix

Take gradient and set equal to 0

Transpose this scalar

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

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

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

c) See printout for graph

The best I got was 0.8628 = RMSE for $\lambda = 8.6431$

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 $\mathbf{x}^*$ 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 $(\mathbf{x}^*, b^*)$ 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) $||A\vec{x} + b\vec{1} - \vec{y}||^2 + ||\Gamma\vec{x}||^2 = (\vec{x}^T A^T + b\vec{1}^T - \vec{y}^T)(A\vec{x} + b\vec{1} - \vec{y}) + \vec{x}\Gamma^T\Gamma\vec{x}^T$

$= x^TA^TAx + 2b\vec{1}^TAx - 2x^TA^Ty - 2b\vec{1}^Ty + b^2\vec{1}^T\vec{1} + y^Ty + \vec{x}\Gamma^T\Gamma\vec{x}^T$

$\nabla_x \to 2A^TA\vec{x}^* + 2b^*A^T1 - 2A^Ty + 2\Gamma^T\Gamma\vec{x}^* = 0$

$\nabla_b \to 2\vec{1}^TA\vec{x}^* - 2\vec{1}^T\vec{y} + 2b^*n = 0$

* I had previously just done ∇_x and stuck w/ that x^*

Therefore the ideal b is given by

$$\boxed{b^* = \frac{\vec{1}^T\vec{y} - \vec{1}^TA\vec{x}^*}{n} = \frac{\vec{1}^T(\vec{y} - A\vec{x}^*)}{n}}$$

Now we can plug this back into our equation for $\nabla_x = 0$:

$$2A^TAx^* + 2\frac{\vec{1}^T(\vec{y} - A\vec{x}^*)\vec{1}A}{n} - 2A^Ty + 2\Gamma^T\Gamma x^* = 0$$

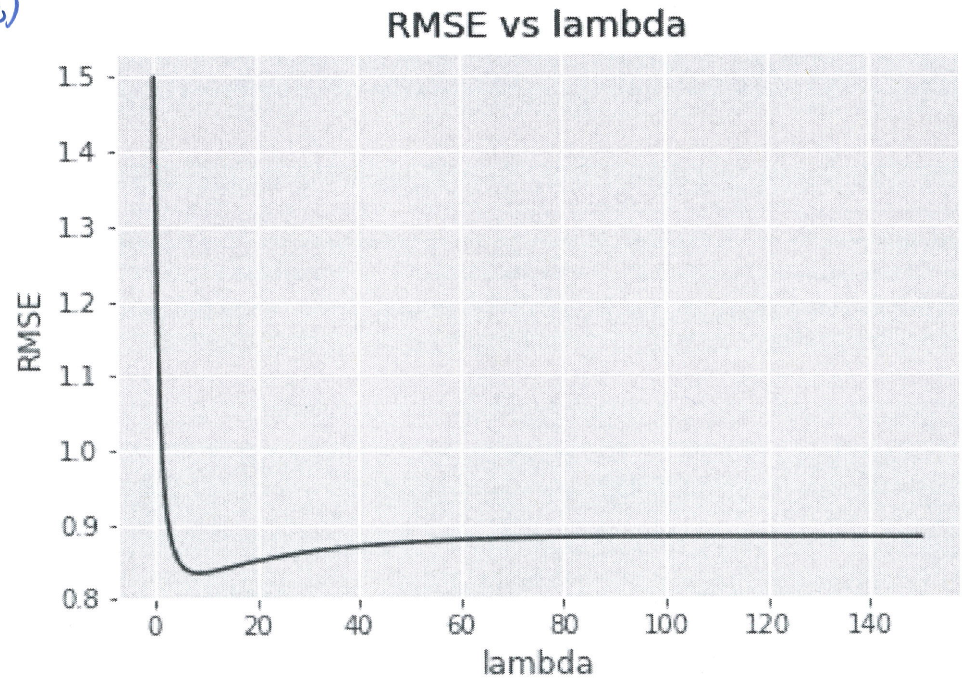
$$A^TAx^* - \frac{1}{n}A^T\vec{1}\vec{1}^TAx^* + \Gamma^T\Gamma x^* - Ay + \frac{1}{n}A^T\vec{1}\vec{1}^Ty = 0$$

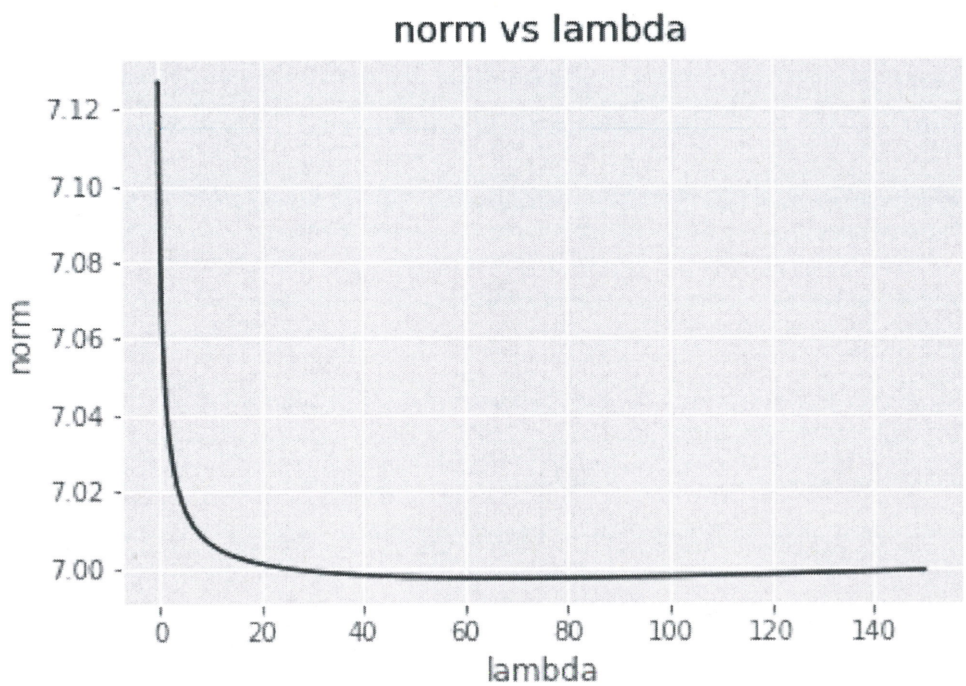
$$\boxed{x^* = \left(A^TA - \frac{1}{n}A^T\vec{1}\vec{1}^TA + \Gamma^T\Gamma\right)^{-1}\left(A^T + \frac{1}{n}A^T\vec{1}\vec{1}^T\right)y}$$ ■

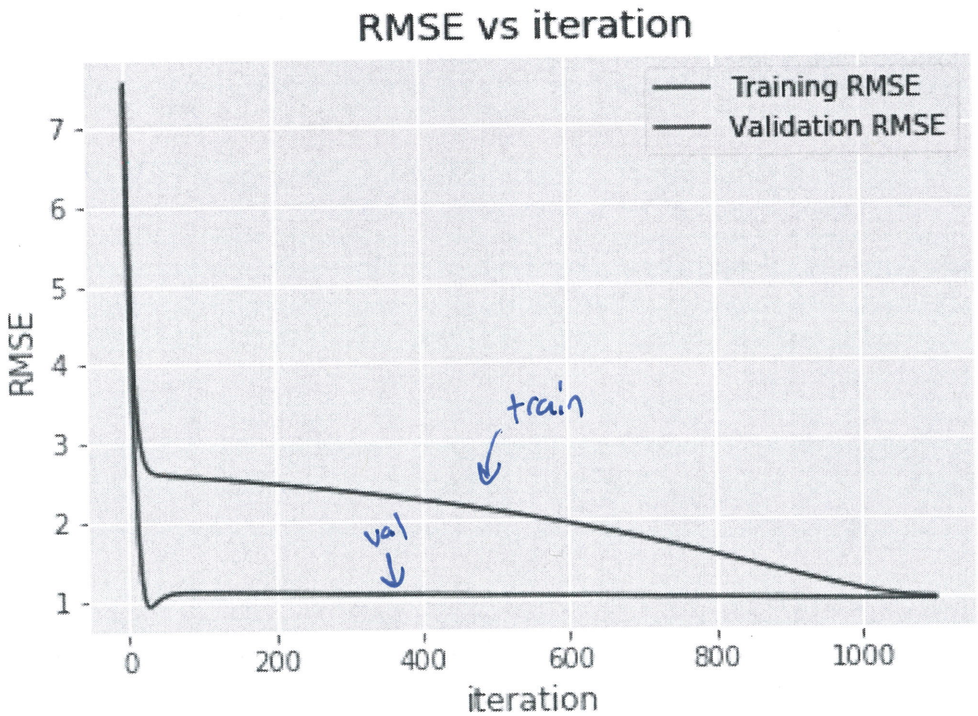
* used factored version for code in my key

This gives us the same bias & weights to within 10^{-9}

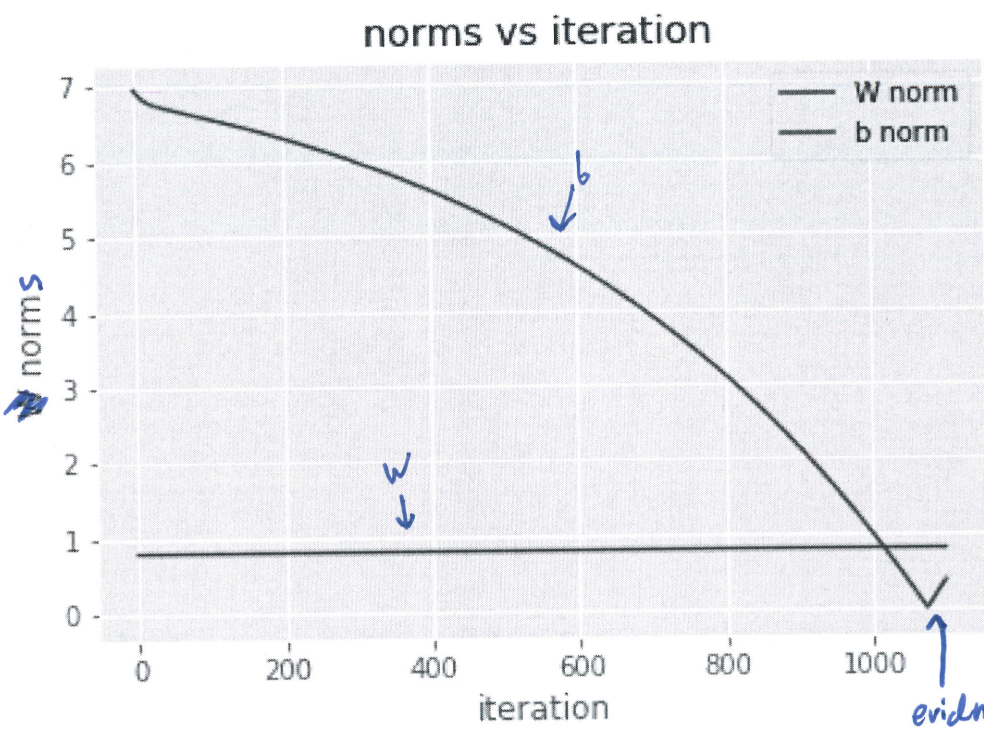
c)


RMSE vs lambda
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
RMSE
0
20
40
60
80
100
120
140
lambda


norm vs lambda
7.12
7.10
7.08
7.06
7.04
7.02
7.00
norm
0
20
40
60
80
100
120
140
lambda



I ended up getting pretty close to the optimal params



evidently I missed the ideal