Math189R SU17
Homework 2
Wednesday, May 16, 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

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

(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)

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

b) The negative log likelihood equation for logistic regression is:

$$nll(\theta) = -\sum_i y_i log\sigma(\theta^\top x_i) + (1-y_i)log(1-\sigma(\theta^\top x_i))$$

The gradient of the negative log likelihood equation is:

$$\nabla nll(\theta) = -\sum_i y_i \frac{1}{\sigma(\theta^\top x_i)}\sigma'(\theta^\top x_i) + (1-y_i)\frac{1}{1-\sigma(\theta^\top x_i)}(-\sigma'(\theta\top x_i))$$

$$= -\sum_i y_i(1-\sigma(\theta^\top x_i))x_i - (1-y_i)\sigma(\theta^\top x_i)x_i$$

$$= \sum_i (\sigma(\theta^\top x_i) - y_i)x_i$$
$$= \sum_i (\mu_i - y_i)x_i$$
$$= X^T(\mu - y)$$

c)

$$H_\theta = \nabla_\theta(\nabla nll(\theta))^\top = \nabla[X^\top(\mu - y)]\top$$
$$= \nabla(\mu^\top X - y^\top X)$$
$$= X^\top diag(\mu(1-\mu)X)$$

As a result, we can see that $H_\theta = X^\top SXwhereS = diag(\mu(1-\mu)$. Since

$$\mu_i(1-\mu_i) = \sigma(\theta^\top x_i)(1-\sigma(\theta^\top x_i)) >= 0,$$

then we know that S is positive semi-definite and H_θ is also positve 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.

If $\mathbb{P}(x;\sigma^2)$ is a valid density, then the integral of it should be 1. Thus take the integration of it we get:

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

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

Square the expression, we get that:

$$\begin{aligned} Z^2 &= \int_{\mathbb{R}} exp(-\frac{x^2}{2\sigma^2})dx \int_{\mathbb{R}} exp(-\frac{y^2}{2\sigma^2})dy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} exp(-\frac{x^2+y^2}{2\sigma^2})dxdy \\ &= \int_0^{\infty}\int_0^{2\pi} exp(-\frac{r^2}{2\sigma^2})rdrd\theta \\ &= 2\pi\theta^2 \end{aligned}$$

Then we take the square rood of it, we get:

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

■

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) We plug in the Gaussian probability distribution into:

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

and we get:

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(y_i - w_0 - \mathbf{w}^\top x_i)^2}{2\sigma^2}) + \sum_{j=1}^{D} \log \frac{1}{sqrt2\pi\sigma exp(-\frac{w_j^2}{2\tau^2})}$$

$$= \sum_{i=1}^{N} (-\frac{(y_i - w_0 - \mathbf{w}^\top x_i)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma}) + \sum_{j=1}^{D} (-\frac{w_j^2}{2\tau^2}) - \log(sqrt2\pi\sigma)$$

Since the constant values do not influence the optimal value of w, then we can take them out and simplify the problem

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

As a result, we get want we want:

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

with $\lambda = \sigma^2/\tau^2$.

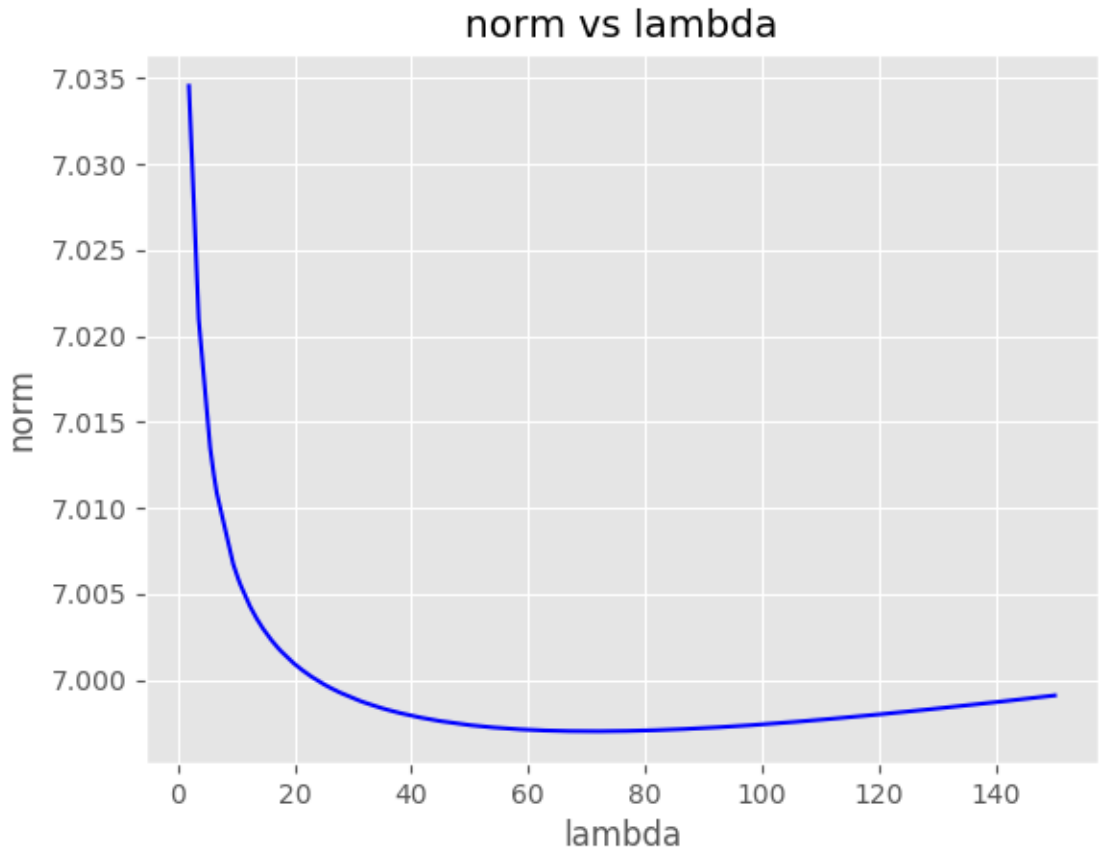
b) To find the value of x for the ridge regression problem, we take the gradient of f with respect to x and set it to 0.

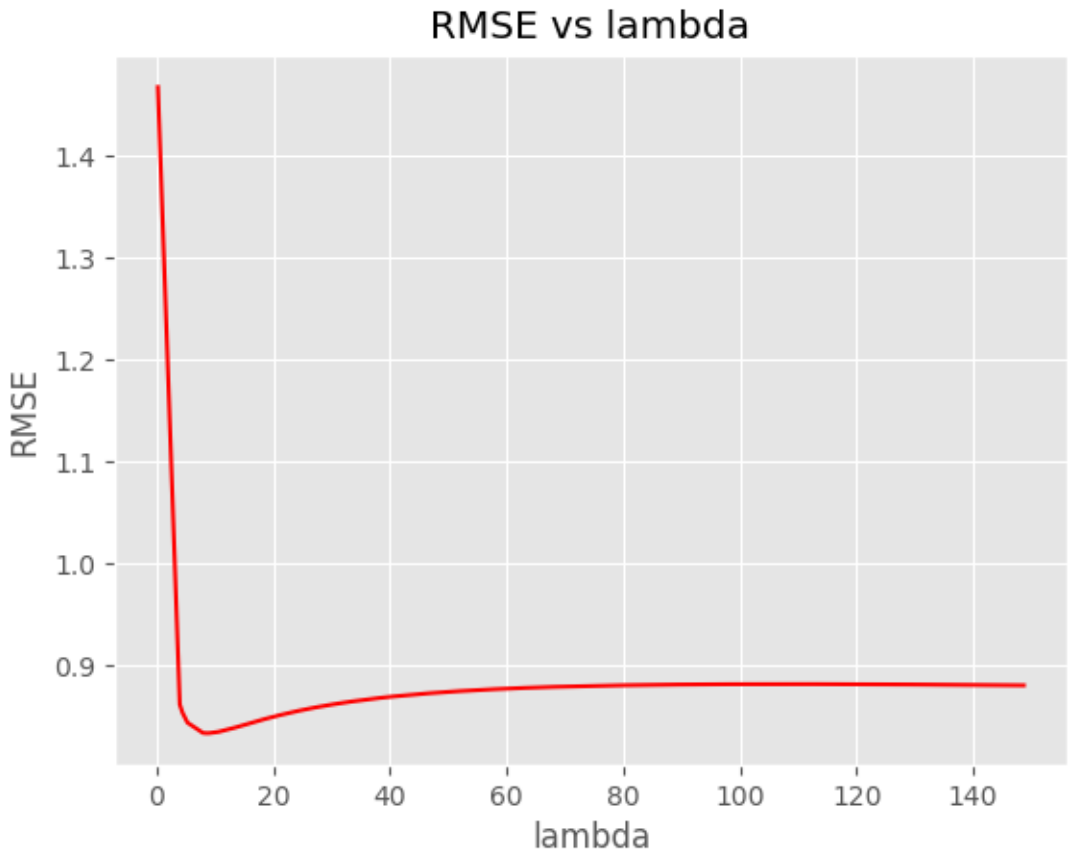
$$\nabla_x f = \nabla[(Ax-b)^\top(Ax-b) + (\tau x)^\top(\tau x)]$$

$$= 2A^\top Ax - 2A^T b + 2\tau^\top \tau x = 0$$

$$x = (A^T A + \tau^T \tau)x = A^T b$$

c) The graphing results are attached below:





When $\lambda = 9.03$ we get the optimal parameter. d) To find the minimal value of f, we take the derivative with respect to x and b, respectively.

$$\nabla_x f = 2A^T Ax + 2b\mathbf{1}1A^T - 2ATy + 2\tau^\tau x = 0$$

$$\nabla_b f = 2\mathbf{1}1^T Ax - 2\mathbf{1}1^T y + 2bn = 0$$

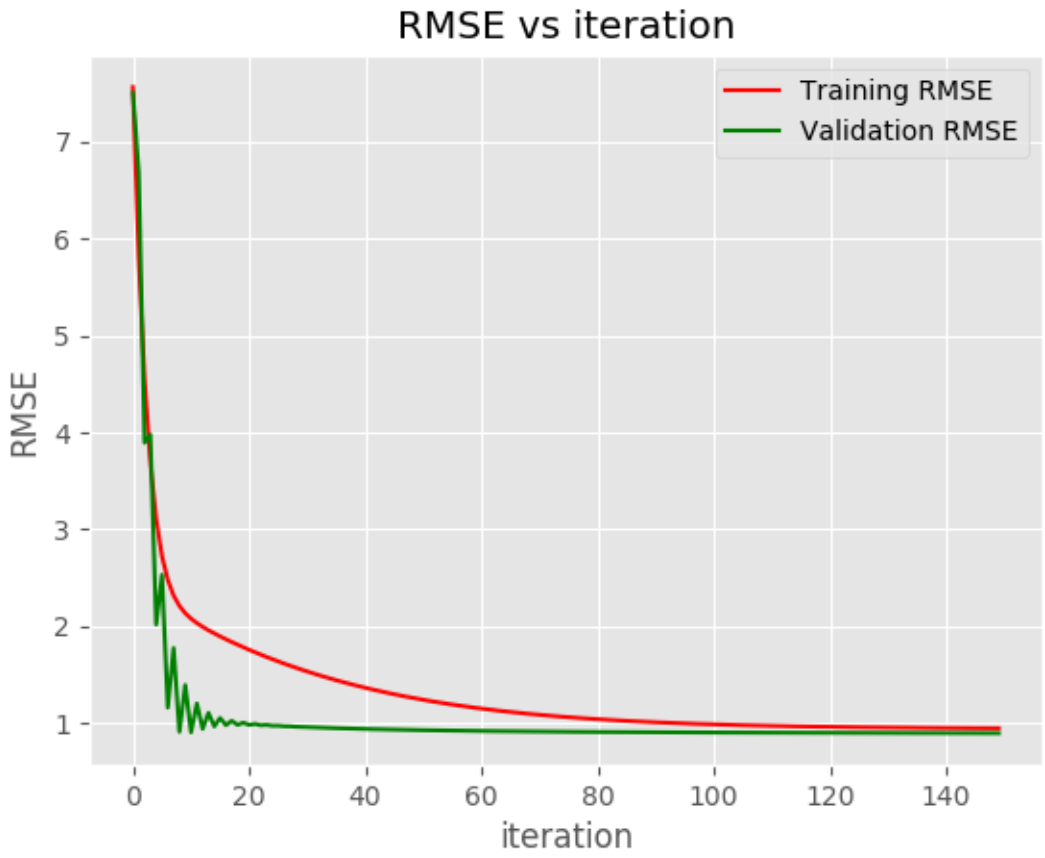
Solving for b:

$$b = \mathbf{1}1^T(y - Ax)$$

/n Plug the result of b back in x:

$$x = [A^T(I - \frac{1}{n}I)A + \tau^T\tau]^{-1}A^T(I - \frac{1}{n}\mathbf{1}1\mathbf{1}1^T)y$$

e) The convergence plot is attached below:


RMSE vs iteration
Training RMSE
Validation RMSE
RMSE
iteration
1
2
3
4
5
6
7
0
20
40
60
80
100
120
140

■