Georgia Pope

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

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

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

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

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

$= \frac{1}{(1+e^{-x})} \cdot \frac{e^{-x}}{(1+e^{-x})}$

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

Now, using the method of partial fractions we will separate $\frac{e^{-x}}{1+e^{-x}}$ into a sum:

$\frac{e^{-x}}{(1+e^{-x})} = \frac{A}{1} + \frac{B}{(1+e^{-x})}$

$e^{-x} = A(1+e^{-x}) + B$

$e^{-x} = Ae^{-x} + A + B$

$A = 1$
$A + B = 0$ $\Rightarrow$ So $A = 1$
$B = -1$

So $\frac{e^{-x}}{1+e^{-x}} = 1 - \frac{1}{1+e^{-x}}$

Plugging this in we get

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

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

■

b) In class, we derived the negative log likelihood for logistic regression to be

$$\ell(\theta) = -\sum y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)})) \quad \text{where}$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta x}}$$

Now taking the partial with respect to $\theta_j$ we get

$$\frac{\partial}{\partial\theta_j}\ell(\theta) = -\sum y^{(i)} \cdot \frac{-x_j(i)}{h_\theta(x)} \cdot h'_\theta(x) + (1-y^i) \cdot \frac{-x_j(i)}{1-h_\theta(x)} \cdot h'_\theta(x)$$

From part a, we know that $h'_\theta(x) = h_\theta(x)(1-h_\theta(x))$. Plugging this in, we get:

$$\frac{\partial}{\partial\theta_j}\ell(\theta) = -\sum y^{(i)} \frac{x_j(i)(\cancel{h_\theta(x)})(1-h_\theta(x))}{\cancel{h_\theta(x)}} + (1-y^i) \cdot \frac{-x_j(i)(h_\theta(x))(\cancel{1-h_\theta(x)})}{(\cancel{1-h_\theta(x)})}$$

$$= -\sum\left[y^{(i)}(1-h_\theta(x^{(i)})) - (1-y^i)(h_\theta(x^{(i)}))\right]x_j(i)$$

$$= -\sum_i\left[y^{(i)} - y^{(i)}h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)}h_\theta(x^{(i)})\right]x_j(i)$$

$$= \boxed{-\sum\left[y^{(i)} - h_\theta(\vec{x}^{(i)})\right]x_j(i)}$$

c) From the textbook we know that $-\sum\left[y^{(i)} - h_\theta(x^{(i)})\right]x_j(i)$ is $X^T(\mu - y)$

So now we need to take the derivative of the gradient with respect to the weights.

We have

$$\text{gradient} = \sum_i\left[h_\theta(\vec{x}(i)) - y(i)\right]x_j(i) = \vec{x}_j^T \sum\left[h_\theta(\vec{x}(i)) - y(i)\right]$$

Now, we will take the partial with respect to $\theta_k$, one of the weights.

$$\frac{\partial}{\theta_k} \frac{\partial}{\theta_j}\ell(\theta) = \vec{x}_j^T \sum_i\left[h'_\theta(\vec{x}(i))(x_k(i))\right]$$

$$= \vec{x}_j^T \sum_i\left[h_\theta(\vec{x}(i))\left[1-h_\theta(\vec{x}(i))\right]\right]\vec{x}_k$$

For all collumns k and j, this can be generalized to be

$$= X^T S X \quad \text{where } S = \operatorname{diag}(\mu_1(1-\mu_1) \ldots \mu_n(1-\mu_n)).$$

We know that $X^TSX$ is a positive semidefinite matrix because all eigen values of $S$ are $\geq 0$ because $0 \leq \mu_i \leq 1$ and $X$ is squared so, since $X^TSX \succeq 0$ if $A^T \succeq 0$, then $H \succeq 0$.

Looked at solution:

**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 then $\int \mathbb{P}(x;\sigma^2)dx = 1$.
So we have

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

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

We also know $Z = \int_{\mathbb{R}} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy$

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

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

We can rewrite this in terms of polar coordinates
So we have

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

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

Let $u = \frac{-r^2}{2\sigma^2}$

$du = \frac{-2r}{2\sigma^2}dr$

$\frac{-\sigma^2 du}{r} = dr$

$$Z^2 = -2\pi \int_0^{-\infty} \sigma^2 e^u du$$

$$Z^2 = -2\pi\sigma^2(0-1)$$

$$Z^2 = 2\pi\sigma^2 \quad \text{So } \boxed{Z = \sqrt{2\pi}\,\sigma}$$

■

**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

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

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 $\lambda$ from the validation set. Plot both $\lambda$ 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 $\lambda$ 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 & b on the back of this page

■

a) We want to show that $\underset{w}{\arg\max} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + \vec{w}^T \vec{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$

equals $-\arg\min \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + \vec{w}^T \vec{x}_i))^2 + \lambda \|\vec{w}\|_2^2$ where $\lambda = \sigma^2 / \tau^2$.

So, we have

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

Plugging in the Gaussian Distribution we get

$$\arg\max \sum_{i=1}^{N} \log \left[ \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(y_i - (w_0 + \vec{w}^T \vec{x}_i))^2} \right] + \sum_{j=1}^{D} \log \left[ \frac{1}{\sqrt{2\pi}\,\tau} e^{-\frac{1}{2\tau^2}(w_j)^2} \right]$$

$$= \arg\max \sum_{i=1}^{N} \left[ \underbrace{\log \cancel{\tfrac{1}{\sqrt{2\pi}\sigma}}}_{\text{not important}} - \frac{1}{2\sigma^2}(y_i - (w_0 + \vec{w}^T \vec{x}_i)) \right] + \sum_{j=1}^{D} \left[ \underbrace{\log \cancel{\tfrac{1}{\sqrt{2\pi}\tau}}}_{\text{not important}} + \left(-\frac{1}{2\tau^2}(w_j)^2\right) \right]$$

Multiply through by $-2\sigma^2$

$$= \arg\max \sum_{i=1}^{N} [y_i - (w_0 + \vec{w}^T \vec{x}_i)] + \frac{\sigma^2}{\tau^2} \sum_{i=1}^{D} (w_j)^2$$

We know $\|\vec{w}\|_2^2 = \sum_j w_j^2$ and $\lambda = \frac{\sigma^2}{\tau^2}$ so we get

$$= \arg\max \sum_{i=1}^{N} (y_i - (w_0 + \vec{w}^T \vec{x}_i))^2 + \lambda \|\vec{w}\|_2^2 \text{ as desired.}$$

b) Looked at solution.

We want to find the closed form solution $x^*$ to minimize $f = \|A\vec{x} + \vec{b}\|_2^2 + \|\Gamma \vec{x}\|_2^2$

To do this, we want to set the gradient equal to zero.

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

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

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

$$= \nabla_x \left[ \vec{x}^T A^T A \vec{x} - 2\vec{x}^T A^T \vec{b} + \vec{x}^T \Gamma^T \Gamma \vec{x} \right]$$

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

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

$$A^T \vec{b} = (A^T A + \Gamma^T \Gamma)\vec{x}$$

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

**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 $\ell_2$ 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) Now we want to minimize $f = ||A\vec{x} + b\vec{1} - \vec{y}||_2^2 + ||\Gamma\vec{x}||_2^2$.

To do is, we want to set the gradient equal to zero.

$$f = [(A\vec{x} + b\vec{1} - \vec{y})^T(A\vec{x} + b\vec{1} - \vec{y}) + (\vec{x}^T\Gamma^T\Gamma\vec{x})]$$

$$f = [(\vec{x}^TA^T + b\vec{1}^T - \vec{y}^T)(A\vec{x} + b\vec{1} - \vec{y}) + (\vec{x}^T\Gamma^T\Gamma\vec{x})]$$

$$f = [(\vec{x}^TA^TA\vec{x} + \vec{x}^TA^Tb\vec{1} - \vec{x}^TA^T\vec{y} + b\vec{1}^TA\vec{x} + b\vec{1}^Tb\vec{1} - b\vec{1}^T\vec{y} - \vec{y}^TA\vec{x} - \vec{y}^Tb\vec{1} - \vec{y}^T\vec{y}) + \vec{x}^T\Gamma^T\Gamma\vec{x}]$$

$$f_x = \nabla_x[\vec{x}^TA^TA\vec{x} + 2\vec{x}^TA^Tb\vec{1} - 2\vec{x}^TA^T\vec{y} - 2b\vec{1}^T\vec{y} + b^2N - \vec{y}^T\vec{y} + \vec{x}^T\Gamma^T\Gamma\vec{x}]$$

$$= 2A^TA\vec{x} + 2A^Tb\vec{1} - 2A^T\vec{y} + 2\Gamma^T\Gamma\vec{x}$$

$$= 2A^TA\vec{x} + 2A^Tb\vec{1} - 2A^T\vec{y} + 2\lambda I\vec{x}$$

$$f_b = \nabla_b[\vec{x}^TA^TA\vec{x} + 2\vec{x}^TA^Tb\vec{1} - 2\vec{x}^TA^T\vec{y} - 2b\vec{1}^T\vec{y} + b^2N - \vec{y}^T\vec{y} + \vec{x}^T\Gamma^T\Gamma\vec{x}]$$

$$= [2\vec{x}^TA^T\vec{1} - 2\vec{1}^T\vec{y} + 2bN]$$

Now setting both $f_x$ * $f_b$ equal to zero:

① $0 = 2A^TA\vec{x} + 2A^T\vec{1}b - 2A^T\vec{y} + 2\lambda I\vec{x}$

② $0 = 2\vec{1}^TA\vec{x} - 2\vec{1}^T\vec{y} + 2bN$

Solving for b in equation ②:

$$bN = \vec{1}^T\vec{y} - \vec{1}^TA\vec{x}$$

$$b = \frac{\vec{1}^T\vec{y}}{N} - \frac{\vec{1}^TA\vec{x}}{N}$$

$$b = \frac{\vec{1}^T(\vec{y} - A\vec{x})}{N}$$

Plugging b into equation ①:

$$0 = A^TA\vec{x} + A^T\vec{1}\left(\frac{\vec{1}^T(\vec{y} - A\vec{x})}{N}\right) - A^T\vec{y} + \lambda I\vec{x}$$

$$0 = A^TA\vec{x} + \frac{A^T\vec{1}\vec{1}^T\vec{y}}{N} - \frac{A^T\vec{1}\vec{1}^TA\vec{x}}{N} - A^T\vec{y} + \lambda I\vec{x}$$

$$\left(A^T\vec{y} - \frac{A^T\vec{1}\vec{1}^T\vec{y}}{N}\right)\left(A^TA - \frac{A^T\vec{1}\vec{1}^TA}{N} + \lambda I\right)^{-1} = \vec{x}^*$$

$$\boxed{\left(A^T\left(I - \frac{\vec{1}\vec{1}^T}{N}\right)A + \lambda I\right)^{-1} A^T\left(I - \frac{\vec{1}\vec{1}^T}{N}\right)\vec{y} = \vec{x}^*}$$

They are the same!