Kevin Loun
Math179 SP24
Homework 2
Monday, Jan 29, 2024

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) We can show this as follows:

$$\begin{aligned}
\sigma(x) = \frac{1}{1+e^{-x}} = \\
\sigma(x) = (1+e^{-x})^{-1} \therefore \\
\sigma'(x) = e^{-x}(1+e^{-x})^{-2} = \\
(\frac{1}{1+e^{-x}})(\frac{e^{-x}}{1+e^{-x}}) = \\
(\sigma(x))(\frac{1+e^{-x}-1}{1+e^{-x}}) = \\
(\sigma(x))(1-\frac{1}{1+e^{-x}}) = \\
\sigma(x)[1-\sigma(x)] \quad \blacksquare
\end{aligned}$$

b) From lecture we know that the log likelihood for logisitic regression is written as

$l(\theta) := \sum_{i=1}^{m} y_i \log h(x_i) + (1 - y_i) \log(1 - h(x_i))$ there fore we seek to find $\frac{\partial}{\partial \theta} l(\theta)$

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \theta_j} &= \left(y \frac{1}{\sigma(\theta^T x)} - (1-y) \frac{1}{1 - \sigma(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y \frac{1}{\sigma(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) \sigma(\theta^T x)(1 - \sigma(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \\
&= \Big(y(1 - \sigma(\theta^T x)) - (1-y)\sigma(\theta^T x) \Big) x_j \\
&= (y - h_\theta(x)) x_j
\end{aligned}$$

Which can be represented as the normal equation:

$$X^T (y - h_\theta)$$

c) We can find the hessian Matrix by doing the following:

$$\begin{aligned}
H_\theta &= \nabla_\theta (\nabla_\theta \ell(\theta))^T = \nabla_\theta [\boldsymbol{X}^T (\mu - \boldsymbol{y})]^T \\
&= \nabla_\theta (\mu^T \boldsymbol{X} - \boldsymbol{y}^T \boldsymbol{X}) \\
&= \nabla_\theta \mu^T \boldsymbol{X} = \nabla_\theta \sigma(\boldsymbol{X}\theta)^T \boldsymbol{X} \\
&= \boldsymbol{X}^T \mathrm{diag}(\mu(1 - \mu)) \boldsymbol{X} \\
&= \boldsymbol{X}^T \boldsymbol{S} \boldsymbol{X}
\end{aligned}$$

We need to show that the following is true:

$$\mu_i(1 - \mu_i) = \sigma(\theta^T \boldsymbol{x}_i)(1 - \sigma(\theta^T \boldsymbol{x}_i)) \geq 0$$

Which becomes simple once we realize that $0 < \sigma < 1$, which implies we must have $\sigma(1 - \sigma) \geq 0$. Hence, the Hessian matrix H is indeed 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.

As we know, a probability density function will integrate to one therefore:

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

Then we can rewrite Z as:

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

Now consider the situation where we sqaure Z:

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

and since this result is Z^2 we find that our final result is:

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

■

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 are tasked with solving a maximum a posteriori estimation problem, which can be formulated in logarithmic terms as follows:

$$\arg\max_{w} \left\{ \sum_{i=1}^{N} \log \mathcal{N}(y_i|\mu_0 + w^T x_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j|0,\tau^2) \right\} \tag{1}$$

Applying the Gaussian distribution $\mathcal{N}(x|\mu,\sigma)$, which has the probability density function $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, we can reformulate our problem as:

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

Leveraging the properties of logarithms, this is equivalent to:

$$\arg\max_{w}\left\{\sum_{i=1}^{N}\left[-\frac{(y_i-w_0-w^Tx_i)^2}{2\sigma^2}-\log\sqrt{2\pi}\sigma\right]+\sum_{j=1}^{D}\left[-\frac{w_j^2}{2\tau^2}-\log\sqrt{2\pi}\tau\right]\right\} \tag{3}$$

Further simplifying, we can express our objective as:

$$\arg\max_{w}\left\{-(N+D)\log\sqrt{2\pi}\sigma+\sum_{i=1}^{N}\left[-\frac{(y_i-w_0-w^Tx_i)^2}{2\sigma^2}\right]+\sum_{j=1}^{D}\left[-\frac{w_j^2}{2\tau^2}\right]\right\} \tag{4}$$

Note that the constant term $-(N+D)\log\sqrt{2\pi}\sigma$ does not influence the value of w^* that maximizes our expression. Consequently, we can omit this constant and scale the problem by $2\sigma^2$ without altering the optimal solution w^*. Thus, we can reformulate our problem as:

$$\arg\min_{w}\left\{\sum_{i=1}^{N}(y_i-w_0-w^Tx_i)^2+\frac{\sigma^2}{\tau^2}\sum_{j=1}^{D}w_j^2\right\} \tag{5}$$

By defining $\lambda=\frac{\sigma^2}{\tau^2}$, the problem becomes:

$$\arg\min_{w}\left\{\sum_{i=1}^{N}(y_i-w_0-w^Tx_i)^2+\lambda\sum_{j=1}^{D}w_j^2\right\} \tag{6}$$

This leads us to the final equivalent form of the optimization problem:

$$\arg\min_{w}\left\{\sum_{i=1}^{N}(y_i-w_0-w^Tx_i)^2+\lambda||w||^2\right\} \tag{7}$$

(b) The closed-form solution for the ridge regression problem can be derived by minimizing the function:

$$f=\|Ax-b\|^2+\lambda\|x\|^2.$$

To obtain this solution, we calculate the gradient of f with respect to x and set it to zero:

$$\nabla_x f=\nabla_x\left((Ax-b)^T(Ax-b)+(\Gamma x)^T(\Gamma x)\right)$$

$$= \nabla_x \left((x^T A^T - b^T)(Ax - b) + x^T \Gamma^T \Gamma x \right)$$
$$= \nabla_x \left(x^T A^T Ax - 2x^T A^T b + b^T b + x^T \Gamma^T \Gamma x \right)$$
$$= 2A^T Ax - 2A^T b + 2\Gamma^T \Gamma x.$$

Setting $\nabla_x f = 0$, we obtain the equation:

$$(A^T A + \Gamma^T \Gamma)x = A^T b.$$

Thus, the closed-form solution is:

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

By setting $\Gamma = \sqrt{\lambda} I$, we align our objective with the ridge regression formulation:

$$f = \|Ax - b\|^2 + \lambda x^T x.$$

This yields the closed-form optimal solution:

$$x^* = (A^T A + \lambda I)^{-1} A^T b.$$

c) Coding Question ■

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) We address the optimization problem with the objective function defined as follows:

$$f = \|Ax + b1 - y\|_2^2 + \|\Gamma x\|_2^2.$$

Expanding the objective function, we get:

$$\begin{aligned} f &= \|Ax + b1 - y\|_2^2 + \|\Gamma x\|_2^2 \\ &= (Ax + b1 - y)^T(Ax + b1 - y) + (\Gamma x)^T(\Gamma x) \\ &= (x^T A^T + b1^T - y^T)(Ax + b1 - y) + x^T\Gamma^T\Gamma x \\ &= x^T A^T Ax + 2b1^T Ax - 2y^T Ax - 2b1^T y + b1^T b1 + y^T y + x^T\Gamma^T\Gamma x. \end{aligned}$$

To find the gradient of f with respect to x and b and set it to zero, we perform the following computations:

$$\nabla_x f = 2A^T Ax + 2bA^T 1 - 2A^T y + 2\Gamma^T\Gamma x = 0 \quad —(*)$$

$$\nabla_b f = 21^T Ax - 21^T y + 2bn = 0$$

Solving for b^*, we find:

$$b^* = \frac{1^T(y - Ax)}{n}$$

This result is intuitive: if the model predicts a flat line (i.e., $x = 0$), the optimal bias term b^* is the average of the outputs y, as expected.

Substituting b^* back into equation (*) to solve for x^*, we obtain:

$$(A^TA + \Gamma^T\Gamma)x + \left(\frac{1^T(y - Ax)}{n}\right) A^T1 - A^Ty = 0$$

$$(A^TA + \Gamma^T\Gamma)x + \frac{1}{n}A^T11^Ty - \frac{1}{n}A^T11^TAx - A^Ty = 0$$

$$A^TA + \Gamma^T\Gamma - \frac{1}{n}A^T11^TAx = A^Ty - \frac{1}{n}A^T11^Ty$$

$$A^T\left(I - \frac{1}{n}11^T\right)A + \Gamma^T\Gamma x = A^T\left(I - \frac{1}{n}11^T\right)y$$

The closed-form solution for x^* becomes:

$$x^* = \left[A^T\left(I - \frac{1}{n}11^T\right)A + \Gamma^T\Gamma\right]^{-1} A^T\left(I - \frac{1}{n}11^T\right)y$$

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

(e) Code Question ■