Bennett Mountain
Math189R SP19
Homework 2
Monday, Feb 11, 2019

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.

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

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

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

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

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

$$\boxed{\sigma'(x) = \sigma(x)(1-\sigma(x))}$$

■

B) The negative log likelihood for logistic regression is:

$$NLL(\theta) = -\sum_i y_i \log \sigma(\theta^T x_i) + (1-y_i)\log(1-\sigma(\theta^T x_i))$$

Taking the gradient w.r.t. θ, we find

$$\nabla_\theta NLL(\theta) = -\sum_i y_i \frac{1}{\sigma(\theta^T x_i)} \sigma'(\theta^T x_i) + (1-y_i)\frac{1}{1-\sigma(\theta^T x_i)}(-\sigma'(\theta^T x_i))$$

$$= -\sum_i y_i \frac{1}{\cancel{\sigma(\theta^T x_i)}} \cancel{\sigma(\theta^T x_i)}(1-\sigma(\theta^T x_i)) + (1-y_i)\frac{1}{\cancel{1-\sigma(\theta^T x_i)}}(-\sigma(\theta^T x_i)\cancel{(1-\sigma(\theta^T x_i))})$$ as $\sigma' = \sigma(1-\sigma)$ from part A.

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

$$= -\sum_i y_i x_i - \cancel{y_i\sigma(\theta^T x_i)x_i} - \sigma(\theta^T x_i)x_i + \cancel{y_i\sigma(\theta^T x_i)x_i}$$

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

$$= \sum_i (\mu_i - y_i)x_i$$

$\nabla_\theta NLL(\theta) = X^T(\mu - y)$, where we let $\mu_i = \sigma(\theta^T x_i)$ and x_i is the i^{th} column of X^T

C) The Hessian is

$$\nabla_\theta(\nabla_\theta NLL(\theta))^T = \nabla_\theta(X^T(\mu - y))^T$$

$$= \nabla_\theta(X^{TT}(\mu^T - y^T))$$

$$= \nabla_\theta(\mu^T X - y^T X)$$

$$= \nabla_\theta(\mu^T X) - \nabla_\theta(\cancel{y^T X})$$ $\partial(y^T X)$ w.r.t $= 0$ b/c there is no θ in either y^T or X

$$= \nabla_\theta(\sigma(\theta^T x_i)^T X) = \nabla_\theta \sigma(X\theta)^T X$$

$$= X^T \cdot \text{diag}(\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i)))X$$ $\sigma(\theta^T x_i) = \mu$

$$= X^T \cdot \text{diag}(\mu(1-\mu))X$$

$$H = X^T S X$$

To show that H is positive semi-definite, consider S, as S is $\text{diag}(\mu_1(1-\mu_1), \dots, \mu_n(1-\mu_n))$. For H to be positive semi-definite, S must be positive-semidefinite. Since S is a diagonal matrix, the eigenvalues of S are its entries (diagonal entries). Thus, we only need to consider $\mu_i(1-\mu_i)$, where $\mu_i(1-\mu_i) = \sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))$.

Since $0 < \sigma(\theta^T x_i) < 1$, we know that $\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i)) \geq 0$, thus H is positive semidefinite.

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.

We wish to find a Z such that $\int \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right) = 0$.

Thus, $Z = \int \exp\left(\frac{-x^2}{2\sigma^2}\right)$.

For Z^2, we have

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

$$= \int_0^{2\pi}\int_0^{\infty} r\exp\left(-\frac{r^2}{2\sigma^2}\right) dr\,d\theta \quad \text{by switching to polar coordinates}$$

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

$$= 2\pi \int_0^{\infty} \underbrace{\exp\left(\frac{-r^2}{2\sigma^2}\right) dr}_{-\sigma^2 du} \qquad u = \exp\left(-\frac{r^2}{2\sigma^2}\right),\ du = \frac{-4}{\sigma^2} r\exp\left(-\frac{r^2}{2\sigma^2}\right) dr$$

$$=$$

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

$$= 2\pi \cdot -\sigma^2 \exp\left(\frac{-r^2}{2\sigma^2}\right)\Big|_0^{\infty}$$

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

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

Thus, we find $Z = \sqrt{2\pi}\,\sigma$

■

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 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 start with $\underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + w^T x_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$. We can sub in for $\mathcal{N}$, as $\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{w_j^2}{2\tau^2}\right)$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \left(-\frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right) + \sum_{j=1}^{D} \left(-\frac{w_j^2}{2\tau^2} - \log \sqrt{2\pi}\sigma \right)$$

$$= \underset{w}{\operatorname{argmax}} -\left((N+D) \log \sqrt{2\pi}\sigma + \sum_{i=1}^{N} \frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2} \right)$$

$(N+D)\log\sqrt{2\pi}\sigma$ doesn't actually affect our optimal solution, w^*. Further, we can multiply through by σ^2 and have it not affect our optimal solution.

Thus, our problem becomes

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2$$ as maximizing the negative is equivalent to minimizing

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda \|w\|_2^2$$ if we let $\frac{\sigma^2}{\tau^2} = \lambda$, we obtain the desired result

B) We wish to find a closed form solution to minimize: $\|Ax-b\|_2^2 + \|\Gamma x\|_2^2$, meaning, we need to find th gradient w.r.t x and set it to 0.

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

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

$$= A^T A x - 2A^T b + x \Gamma^T \Gamma$$

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

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

so $x^* = (A^T A + \Gamma^T \Gamma)^{-1} A^T b$ is our closed form solution.

Calling $\Gamma = \sqrt{\lambda} I$, the minimization function becomes $\|Ax-b\|_2^2 + \lambda x^T x$, which gives the closed form optimal solution of

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

C) The optimal regularization parameter is 8.4375

The RMSE on the validation set w/ the optimal regularization parameter is 0.8340

The RMSE on the test set is 0.8628.