Math 189R problem set 2

Adam Guo 2020-02-10

1. **(Murphy 8.3)** Gradient and Hessian of the log-likelihood for logistic regression.
   Hint: Use the **negative** log-likelihood of logistic regression for this problem.

   (a) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Show that

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

   Solution:

   $$\begin{aligned}
   \sigma'(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1}{1+e^{-x}}\right] \\
   &= -\frac{-e^{-x}}{(1+e^{-x})^2} \\
   &= \frac{e^{-x}}{(1+e^{-x})^2} \\
   \sigma(x)[1-\sigma(x)] &= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) \\
   &= \frac{1}{1+e^{-x}} - \left(\frac{1}{1+e^{-x}}\right)^2 \\
   &= \frac{e^{-x}}{(1+e^{-x})^2}
   \end{aligned}$$

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

   (b) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.

   Solution:
   By Murphy,

   $$\mathrm{NLL}(\mathbf{w}) = -\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$$

   where $\mu_i = \sigma(\mathbf{w}^T\mathbf{x}_i)$.

$$
\begin{aligned}
\mathrm{NLL}(\mathbf{w}) &= -\sum_{i=1}^{N}[y_i \log \sigma(\mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log\big(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\big)] \\
\nabla\mathrm{NLL}(\mathbf{w}) &= -\sum_{i=1}^{N}\left[y_i\frac{\sigma'(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i}{\sigma(\mathbf{w}^T\mathbf{x}_i)} + (1-y_i)\left(\frac{-\sigma'(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i}{1-\sigma(\mathbf{w}^T\mathbf{x}_i)}\right)\right] \\
&= -\sum_{i=1}^{N}\left[y_i\frac{\sigma(\mathbf{w}^T\mathbf{x}_i)[1-\sigma(\mathbf{w}^T\mathbf{x}_i)]}{\sigma(\mathbf{w}^T\mathbf{x}_i)}\mathbf{x}_i + (1-y_i)\left(\frac{-\sigma(\mathbf{w}^T\mathbf{x}_i)[1-\sigma(\mathbf{w}^T\mathbf{x}_i)]}{1-\sigma(\mathbf{w}^T\mathbf{x}_i)}\right)\mathbf{x}_i\right] \\
&= -\sum_{i=1}^{N}\big[y_i(1-\sigma(\mathbf{w}^T\mathbf{x}_i))\mathbf{x}_i + (1-y_i)(-\sigma(\mathbf{w}^T\mathbf{x}_i))\mathbf{x}_i\big] \\
&= -\sum_{i=1}^{N} y_i\mathbf{x}_i - \sigma(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i \\
&= \sum_{i=1}^{N}(\mu_i - y_i)\mathbf{x}_i \\
&= \mathbf{X}^T(\boldsymbol{\mu}-\mathbf{y})
\end{aligned}
$$

(c) The Hessian can be written as $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$ where $\mathbf{S} = \mathrm{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Derive this and show that $\mathbf{H} \succeq 0$ ($A \succeq 0$ means that A is positive semidefinite).

Solution:

$$
\begin{aligned}
\mathbf{H} = \nabla(\nabla\mathrm{NLL}(\mathbf{w}))^T &= \nabla(\mathbf{X}^T(\boldsymbol{\mu}-\mathbf{y}))^T \\
&= \nabla(\mathbf{X}^T\boldsymbol{\mu} - \mathbf{X}^T\mathbf{y})^T \\
&= \nabla(\boldsymbol{\mu}^T\mathbf{X}) \\
&= \nabla\left(\sum_i \mu_i\right)\mathbf{x}_i \quad \text{where } \mathbf{x}_i \text{ denotes the columns of } \mathbf{X} \\
&= \left(\sum_i \nabla\mu_i\right)\mathbf{x}_i \\
&= \left(\sum_i \nabla\sigma(\mathbf{x}_i^T\mathbf{w})\right)\mathbf{x}_i \\
&= \left(\sum_i \mu_i(1-\mu_i)\right)\mathbf{x}_i\mathbf{x}_i^T \\
&= \mathbf{X}^T\mathbf{S}\mathbf{X}
\end{aligned}
$$

The eigenvalues of $\mathbf{H}$ are given by $\mathbf{S}$. Since $\mathbf{H}$ is positive semidefinite iff its eigenvalues are non-negative, $\mathbf{H}$ is positive semidefinite iff $\mathbf{S}$ is positive semidefinite.
$\mu_i = \sigma(\mathbf{w}^T\mathbf{x}_i)$. By definition, $\sigma \in (0,1)$. Hence, $\mu_i \in (0,1)$ and $(1-\mu_i) \in (0,1)$. Hence, $\mu_i(1-\mu_i) \geq 0$. Thus, $\mathbf{S}$ and hence $\mathbf{H}$ are positive semidefinite.

2. **(Murphy 2.11)** Derive the normalisation 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.

Solution:

$$\int_{-\infty}^{\infty} \frac{1}{Z}\exp\left(\frac{-x^2}{2\sigma^2}\right)\mathrm{d}x = 1$$
$$\left(\int_{-\infty}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right)\mathrm{d}x\right)^2 = Z^2$$

We use this expression for Z^2 to integrate in polar coordinates.

$$\begin{aligned}
Z^2 &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\exp\left(\frac{-x^2}{2\sigma^2}\right)\mathrm{d}x\right)\exp\left(\frac{-y^2}{2\sigma^2}\right)\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(\frac{-x^2-y^2}{2\sigma^2}\right)\mathrm{d}x\,\mathrm{d}y \\
&= \int_{0}^{2\pi}\int_{0}^{\infty}\exp\left(\frac{-r^2}{2\sigma^2}\right)r\,\mathrm{d}r\,\mathrm{d}\theta
\end{aligned}$$

Let $u = r^2$. $\mathrm{d}u = 2r\,\mathrm{d}r$.

$$\begin{aligned}
Z^2 &= \frac{1}{2}\int_{0}^{2\pi}\int_{0}^{\infty}\exp\left(\frac{-u}{2\sigma^2}\right)\mathrm{d}u\,\mathrm{d}\theta \\
&= \frac{1}{2}\int_{0}^{2\pi} -2\sigma^2\,\mathrm{d}\theta \\
&= 2\pi\sigma^2 \\
Z &= \sqrt{2\pi}\sigma
\end{aligned}$$

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 optimise 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 $\mathbb{P}(\mathbf{w}) = \prod_j \mathcal{N}(w_j \mid 0, \tau^2)$ on the weights,

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

   is equivalent to the ridge regression problem

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

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

   Solution:

$$\begin{aligned}
&\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i \mid w_0 + \mathbf{w}^T \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j \mid 0, \tau^2) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2}{2\sigma^2}} \right) + \sum_{j=1}^{D} \log \left(\frac{1}{\sqrt{2\pi}\tau} e^{-\frac{w_j^2}{2\tau^2}} \right) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2}{2\sigma^2} \right) + \sum_{j=1}^{D} \left(\log \frac{1}{\sqrt{2\pi}\tau} - \frac{w_j^2}{2\tau^2} \right) \\
&= \arg\max_{\mathbf{w}} \log \frac{1}{(\sqrt{2\pi}\sigma)^N} + \log \frac{1}{(\sqrt{2\pi}\tau)^D} - \left(\sum_{i=1}^{N} \frac{(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2}{2\sigma^2} + \frac{||w||_2^2}{2\tau^2} \right)
\end{aligned}$$

   Since the first two constant log terms do not affect the optimisation problem, this is equivalent to minimising the summation term that is being subtracted. Hence, this is equivalent to

$$\begin{aligned}
&\arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2}{2\sigma^2} + \frac{||w||_2^2}{2\tau^2} \\
&= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \left(\sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \frac{\sigma^2 ||w||_w^2}{\tau^2} \right)
\end{aligned}$$

   Multiplying by a constant term $2\sigma^2$ and letting $\lambda = \frac{\sigma^2}{\tau^2}$, this is thus equivalent to the ridge regression problem

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - (w_o + \mathbf{w}^T \mathbf{x}_i))^2 + \lambda ||w||_2^2$$

(b) **(math)** Find a closed form solution $\mathbf{x}^*$ to the ridge regression problem:

$$\text{minimise: } ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2$$

Solution:

$$\begin{aligned}
\nabla(||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2) &= \nabla((A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x})) \\
&= \nabla((\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) + (\mathbf{x}^T\Gamma^T)(\Gamma\mathbf{x})) \\
&= \nabla(\mathbf{x}^T A^T A\mathbf{x} - \mathbf{x}^T A^T\mathbf{b} - \mathbf{b}^T A\mathbf{x} + \mathbf{b}^T\mathbf{b} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}) \\
&= 2A^T A\mathbf{x} - 2A^T\mathbf{b} + 2\Gamma^T\Gamma\mathbf{x}
\end{aligned}$$

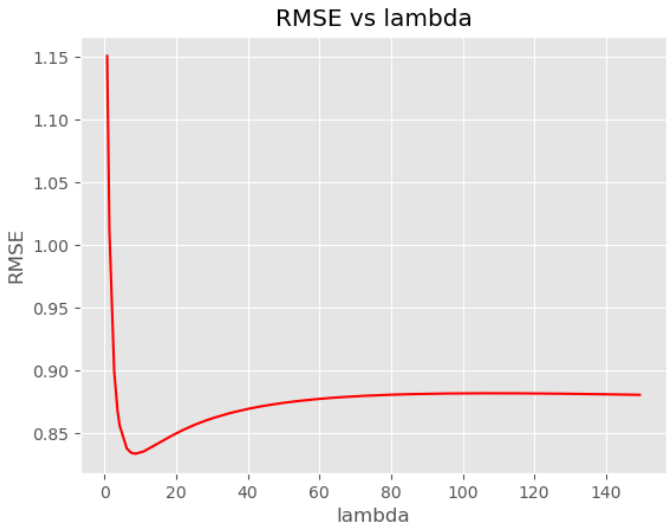
To optimise, let the gradient equal 0.

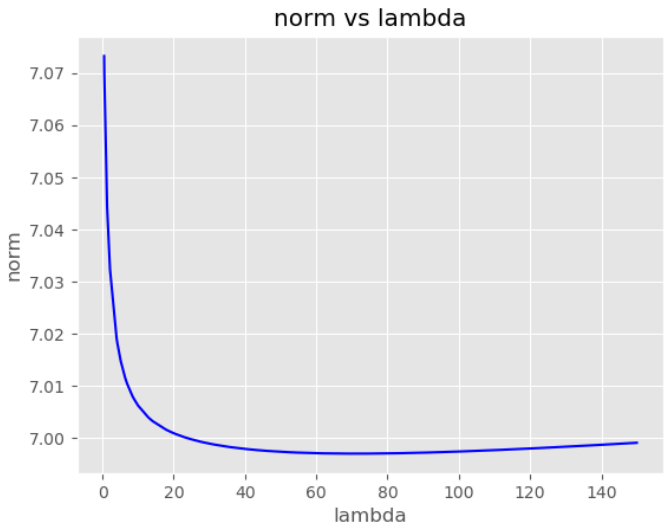
$$\begin{aligned}
2A^T\mathbf{b} &= 2A^T A\mathbf{x}^* + 2\Gamma^T\Gamma\mathbf{x}^* \\
A^T\mathbf{b} &= (A^T A + \Gamma^T\Gamma)\mathbf{x}^* \\
\mathbf{x}^* &= (A^T A + \Gamma^T\Gamma)^{-1}A^T\mathbf{b}
\end{aligned}$$

yielding the solution for $\mathbf{x}^*$.

(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 regularise the bias term.* Find the optimal regularisation 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 $||\theta^*||_2$ where θ is your weight vector. What is the final RMSE on the test set with the optimal λ^*?

Solution:





The optimal regularisation parameter is $\lambda = 9.1326$, with RMSE of 0.8341 on the validation set and 0.8628 on the test set.

(d) **(math)** Consider regularised linear regression where we pull the basis term out of the feature vectors. That is, instead of computing $\hat{\mathbf{y}} = \theta^T \mathbf{x}$ with $\mathbf{x}_0 = 1$, we compute $\hat{\mathbf{y}} = \theta^T \mathbf{x} + b$. This corresponds to solving the optimisation problem

$$\text{minimise: } ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2$$

Solve for the optimal $\mathbf{x}^*$ explicitly. Use this closed form to compute the bias term for the previous problem (with the same regularisation strategy). Make sure it is the same.

Solution:
Suppose $\mathbf{1}$ has length n.

$$\begin{aligned}
&||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2 \\
&= (A\mathbf{x} + b\mathbf{1} - \mathbf{y})^T(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x}) \\
&= (\mathbf{x}^T A^T + b\mathbf{1}^T - \mathbf{y}^T)(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} \\
&= \mathbf{x}^T A^T A\mathbf{x} + \mathbf{x}^T A^T b\mathbf{1} - \mathbf{x}^T A^T\mathbf{y} + b\mathbf{1}^T A\mathbf{x} + b^2\mathbf{1}^T\mathbf{1} - b\mathbf{1}^T\mathbf{y} - \mathbf{y}^T A\mathbf{x} - b\mathbf{y}^T\mathbf{1} + \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} \\
&= \mathbf{x}^T A^T A\mathbf{x} + 2b\mathbf{1}^T A\mathbf{x} - 2\mathbf{y}^T A\mathbf{x} - 2b\mathbf{1}^T\mathbf{y} + nb^2 + \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}
\end{aligned}$$

Need to optimise both $\mathbf{x}$ and b.

$$\nabla_b(||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2) = 2\mathbf{1}^T A\mathbf{x} - 2\mathbf{1}^T\mathbf{y} + 2bn$$

Setting $\nabla_b = 0$,

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

Using b^* to optimise x,

$$\begin{aligned}
\nabla_x(||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2) &= 2A^T A\mathbf{x} + 2bA^T\mathbf{1} - 2A^T\mathbf{y} + 2\Gamma^T\Gamma\mathbf{x} \\
&= 2A^T A\mathbf{x} + 2\frac{\mathbf{1}^T(\mathbf{y} - A\mathbf{x})}{n}A^T\mathbf{1} - 2A^T\mathbf{y} + 2\Gamma^T\Gamma\mathbf{x}
\end{aligned}$$

Setting $\nabla_x = 0$,

$$\begin{aligned}
0 &= (A^TA + \Gamma^T\Gamma)\mathbf{x}^* + \frac{1}{n}\mathbf{1}^T(\mathbf{y} - A\mathbf{x}^*)A^T\mathbf{1} - A^T\mathbf{y} \\
&= (A^TA + \Gamma^T\Gamma)\mathbf{x}^* + \frac{1}{n}\mathbf{1}^T\mathbf{y}A^T\mathbf{1} - \frac{1}{n}\mathbf{1}^TA\mathbf{x}^*A^T\mathbf{1} - A^T\mathbf{y} \\
&= (A^TA + \Gamma^T\Gamma)\mathbf{x}^* + \frac{1}{n}A^T\mathbf{1}\mathbf{1}^T\mathbf{y} - \frac{1}{n}A^T\mathbf{1}\mathbf{1}^TA\mathbf{x}^* - A^T\mathbf{y} \\
&= (A^TA + \Gamma^T\Gamma - \frac{1}{n}A^T\mathbf{1}\mathbf{1}^TA)\mathbf{x}^* + (\frac{1}{n}A^T\mathbf{1}\mathbf{1}^T - A^T)\mathbf{y}
\end{aligned}$$

$$\begin{aligned}
(A^TA + \Gamma^T\Gamma - \frac{1}{n}A^T\mathbf{1}\mathbf{1}^TA)\mathbf{x}^* &= (-\frac{1}{n}A^T\mathbf{1}\mathbf{1}^T + A^T)\mathbf{y} \\
&= A^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{y}
\end{aligned}$$

Hence,

$$\mathbf{x}^* = (A^TA + \Gamma^T\Gamma - \frac{1}{n}A^T\mathbf{1}\mathbf{1}^TA)^{-1}A^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{y}$$

is the optimal solution for $\mathbf{x}^*$.

(e) **(implementation)** We can also compute the solution to the least squares problem using gradient descent. Consider the same bias-relocated objective

$$\text{minimise: } f = ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2$$

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.

Solution: