Math of Big Data, Summer 2018

Prof: Gu

Name: Forest Kobayashi

HW #: 4

Day: Mon. Tue. Wed. (Thu.) Fri.

Date: 05/18/2018

No.	Points	Acknowledgments
1		
2		
Total		

This Assignment is (check one):

[X] On Time [ ] Late, without deduction [ ] Late, with deduction

Comments: 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 2 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.

Problem 1. (Conditioning a Gaussian)

(Conditioning a Gaussian) Note that from Murphy page 113. "Equation 4.69 is of such importance in this book that we have put a box around it, so you can easily find it." That equation is important. Read through the proof of the result. Suppose we have a distribution over random variables $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ that is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\mu}_2 = 5 \qquad \Sigma_{11} = \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} \qquad \boldsymbol{\Sigma}_{21}^\top = \begin{bmatrix} 5 \\ 11 \end{bmatrix} = \boldsymbol{\Sigma}_{12} \qquad \Sigma_{22} = [14]$$

Compute

(a) The marginal distribution $p(\mathbf{x}_1)$.

(b) The marginal distribution $p(\mathbf{x}_2)$.

(c) The conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$

(d) The conditional distribution $p(\mathbf{x}_2|\mathbf{x}_1)$

Solution:

(a) We have

$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 \mid \boldsymbol{\mu}_1, \Sigma_{11}) \\ &= \frac{1}{(2\pi)^{D/2}(\det(\Sigma_{11}))^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma_{11}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \\ &= \frac{1}{2\pi\sqrt{14}} \exp\left[-\frac{1}{2}\mathbf{x}^\top \frac{1}{14}\begin{bmatrix} 13 & -8 \\ -8 & 6 \end{bmatrix}\mathbf{x}\right] \\ &= \frac{1}{2\pi\sqrt{14}} \exp\left[-\frac{1}{28}\left(13x_1^2 - 16x_1x_2 + 6x_2^2\right)\right] \end{aligned}$$

(b) Looking at the answer key, I'm realizing that that last part was unnecessary and that I could have just said

$$\begin{aligned} p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 \mid \boldsymbol{\mu}_2, \Sigma_{22}) \\ &= \mathcal{N}(5, 14) \end{aligned}$$

(c)

$$\begin{aligned} p(\mathbf{x}_1 \mid \mathbf{x}_2) &= \mathcal{N}\left(\mathbf{x}_1 \mid \boldsymbol{\mu}_{1|2}, \Sigma_{1|2}\right) \\ &= \mathcal{N}\left(\mathbf{x}_1 \mid \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{1|2}\right) \\ &= \mathcal{N}\left(\mathbf{x}_1 \;\middle|\; \frac{1}{14}\begin{bmatrix} 5 \\ 11 \end{bmatrix}(\mathbf{x}_2 - 5), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) \\ &= \mathcal{N}\left(\begin{bmatrix} \frac{5}{14} \\ \frac{11}{14} \end{bmatrix}(\mathbf{x}_2 - 5), \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \frac{1}{14}\begin{bmatrix} 25 & 55 \\ 55 & 121 \end{bmatrix}\right) \end{aligned}$$

(d)

$$
\begin{aligned}
p(\mathbf{x}_2 \mid \mathbf{x}_1) &= \mathcal{N}\big(\mathbf{x}_2 \mid \boldsymbol{\mu}_{2|1}, \Sigma_{2|1}\big) \\
&= \mathcal{N}\big(\mathbf{x}_2 \mid 5 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \mathbf{0}), \Sigma_{2|1}\big) \\
&= \mathcal{N}\left(\mathbf{x}_2 \;\middle|\; 5 + \frac{1}{14}\begin{bmatrix}5 & 11\end{bmatrix}\begin{bmatrix}13 & -8 \\ -8 & 6\end{bmatrix}\mathbf{x}_1, [14] - \begin{bmatrix}\frac{171}{14}\end{bmatrix}\right) \\
&= \mathcal{N}\left(\mathbf{x}_2 \;\middle|\; 5 + \frac{1}{14}\begin{bmatrix}-23 & 26\end{bmatrix}\mathbf{x}_1, \begin{bmatrix}\frac{25}{14}\end{bmatrix}\right) \\
&= \mathcal{N}\left(\mathbf{x}_2 \;\middle|\; 5\begin{bmatrix}-\frac{23}{14} & \frac{13}{7}\end{bmatrix}\mathbf{x}_1, \begin{bmatrix}\frac{25}{14}\end{bmatrix}\right)
\end{aligned}
$$

Problem 2. (MNIST)

In this problem, we will use the MNIST dataset, a classic in the deep learning literature as a toy dataset to test algorithms on, to set up a model for logistic regression and softmax 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.

The problem is this: we have images of handwritten digits with 28×28 pixels in each image, as well as the label of which digit $0 \leq \texttt{label} \leq 9$ the written digit corresponds to. Given a new image of a handwritten digit, we want to be able to predict which digit it is. The format of the data is `label, pix-11, pix-12, pix-13, ...` where `pix-ij` is the pixel in the `ith` row and `jth` column.

(a) (**logistic**) Restrict the dataset to only the digits with a label of 0 or 1. Implement L2 regularized logistic regression as a model to compute $\mathbb{P}(y = 1|\mathbf{x})$ for a different value of the regularization parameter λ. Plot the learning curve (objective vs. iteration) when using Newton's Method *and* gradient descent. Plot the accuracy, precision ($p = \mathbb{P}(y = 1|\hat{y} = 1)$), recall ($r = \mathbb{P}(\hat{y} = 1|y = 1)$), and F1-score ($F1 = 2pr/(p+r)$) for different values of λ (try at least 10 different values including $\lambda = 0$) on the test set and report the value of λ which maximizes the accuracy on the test set. What is your accuracy on the test set for this model? Your accuracy should definitely be over 90%.

(b) (**softmax**) Now we will use the whole dataset and predict the label of each digit using L2 regularized softmax regression (multinomial logistic regression). Implement this using gradient descent, and plot the accuracy on the test set for different values of λ, the regularization parameter. Report the test accuracy for the optimal value of λ as well as it's learning curve. Your accuracy should be over 90%.

Solution:

(a) For the logistic model, we have the same equation as previously, just with a regularization term added in:

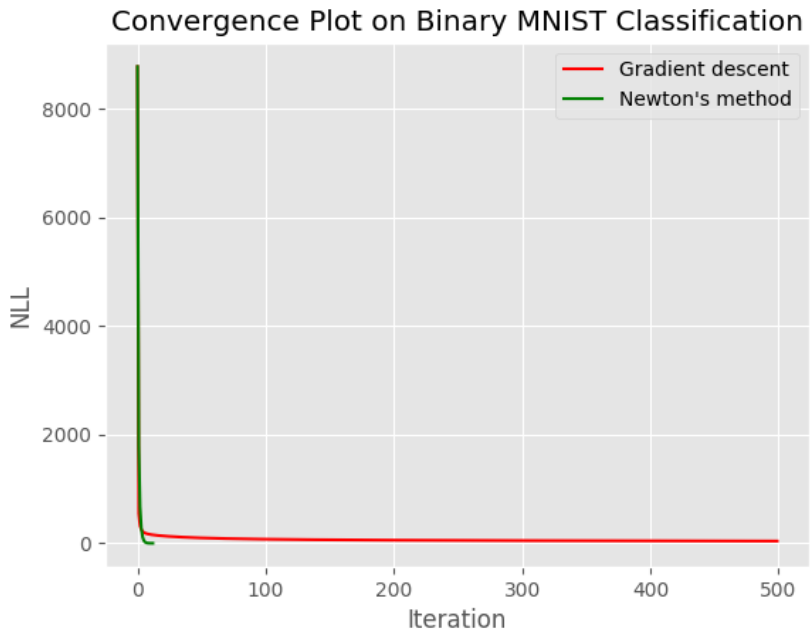
$$\text{NLL}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \Big[y_i \log\big(\sigma\big(\boldsymbol{\theta}^T \mathbf{x}_i\big)\big) + (1 - y^{(i)}) \log\big(1 - \sigma\big(\boldsymbol{\theta}^\top \mathbf{x}_i\big)\big) \Big] + \lambda \|\boldsymbol{\theta}\|_2^2$$

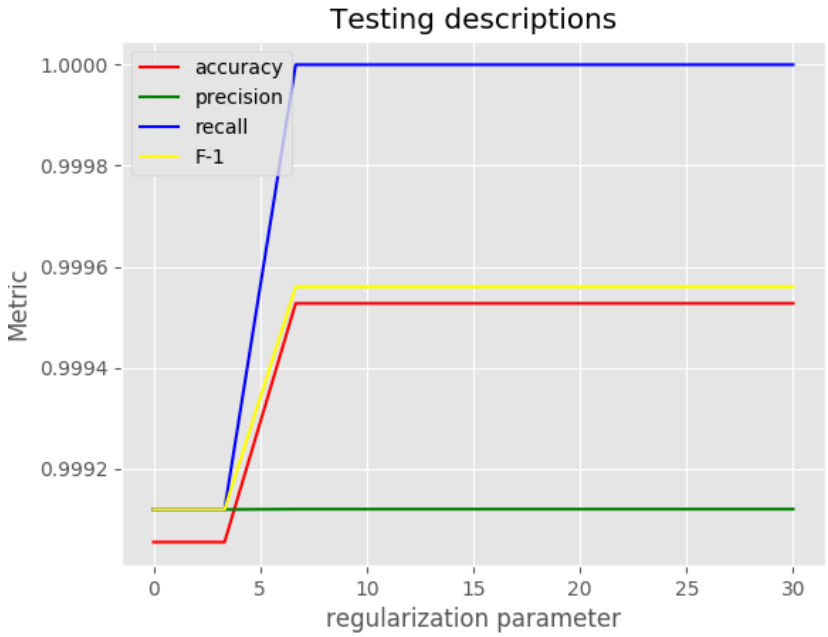
$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \text{NLL}(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \Big[y^{(i)} \big(1 - \sigma\big(\boldsymbol{\theta}^\top \mathbf{x}_i\big)\big) \mathbf{x}_i - (1 - y^{(i)}) \sigma\big(\boldsymbol{\theta}^\top \mathbf{x}_i\big) \mathbf{x}_i \Big] + 2\lambda \boldsymbol{\theta} \\
&= -\sum_{i=1}^{N} \Big[y^{(i)} - \sigma\big(\boldsymbol{\theta}^\top \mathbf{x}_i\big) \Big] \mathbf{x}_i + 2\lambda \boldsymbol{\theta} \\
&= -X^\top (\mathbf{y} - \sigma(X\boldsymbol{\theta})) + 2\lambda \boldsymbol{\theta} \\
&= X^\top (\sigma(X\boldsymbol{\theta}) - \mathbf{y}) + 2\lambda \boldsymbol{\theta}
\end{aligned}$$

the hessian is then

$$\begin{aligned}
H &= \nabla^2 \text{NLL}(\boldsymbol{\theta}) \\
&= \nabla \left(\sum_{i=1}^{N} \Big[y^{(i)} - \sigma\big(\boldsymbol{\theta}^\top \mathbf{x}_i\big) \Big] \mathbf{x}_i + 2\lambda \boldsymbol{\theta} \right)^\top \\
&= \sum_{i=1}^{N} \Big[\big(\sigma\big(\boldsymbol{\theta}^\top \mathbf{x}\big) \mathbf{x}_i\big)^\top \mathbf{x}_i \Big] + 2\lambda I \\
&= X^\top \operatorname{diag}\left(\mu_1(1 - \mu_1), \ldots, \mu_n(1 - \mu_n)\right) X + 2\lambda I
\end{aligned}$$

where $\mu_i = \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$





(b) We calculate the gradient of the softmax stuff. Our model is

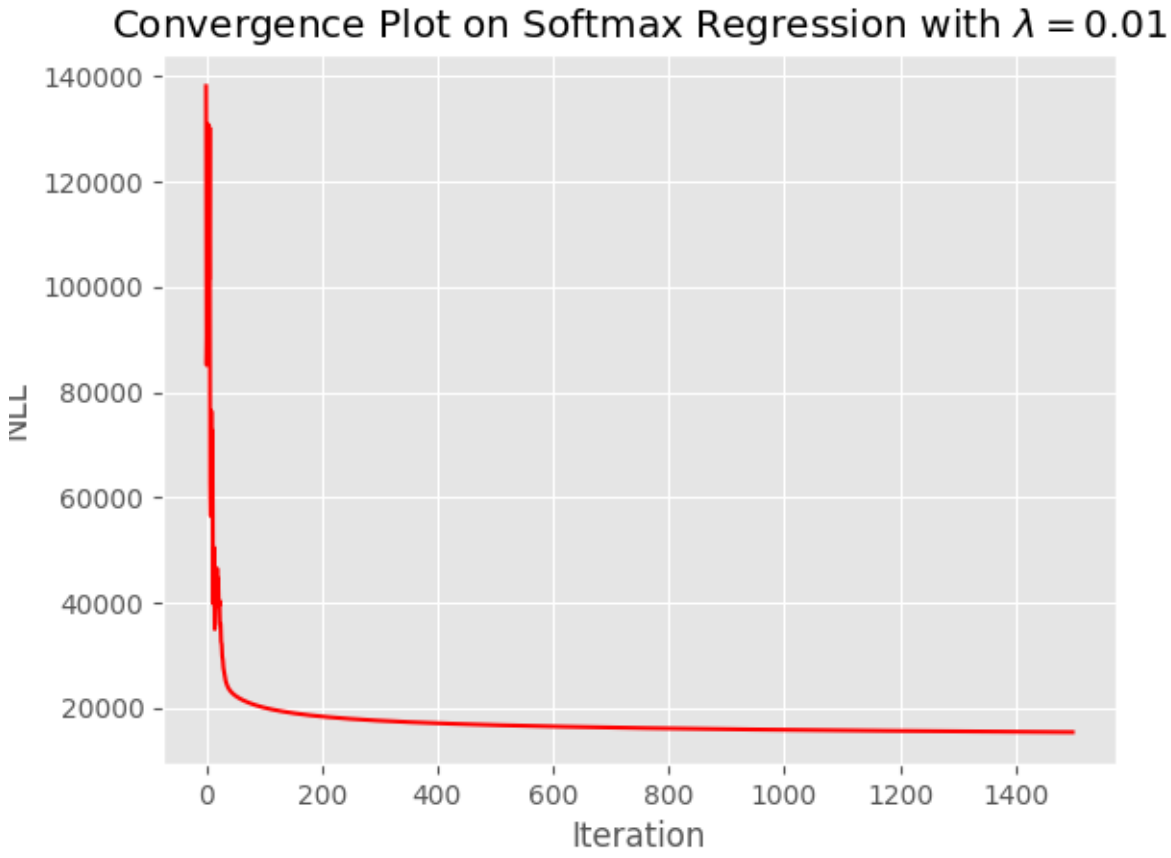
$$p(y = c \mid \mathbf{x}, \mathbf{W}) = \frac{\exp\left(\mathbf{W}_c^\top \mathbf{x}\right)}{\sum_{c'=1}^{C} \exp\left(\mathbf{W}_{c'}^\top \mathbf{x}\right)}$$

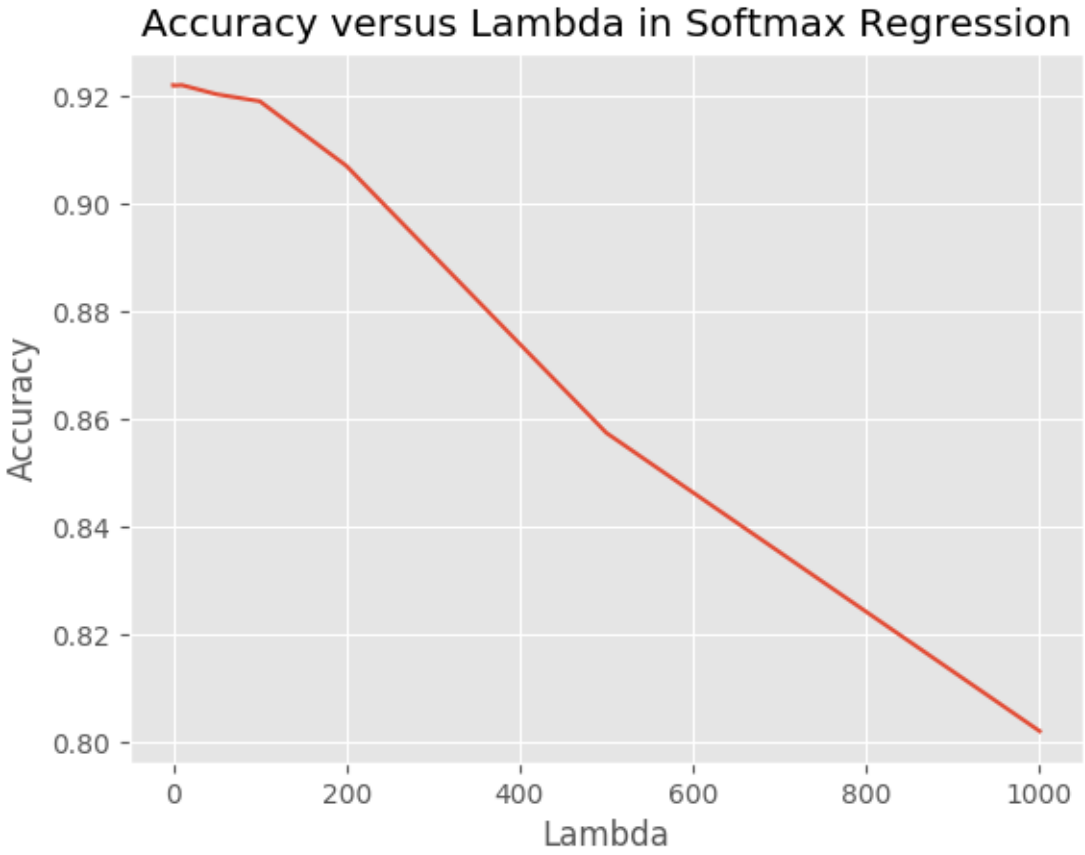
let $\mu_{ic} = p(y_i = c \mid \mathbf{x}_i, \mathbf{W}) = \mathcal{S}(\boldsymbol{\eta_i})_c$, where $\boldsymbol{\eta}_i = \mathbf{W}^\top \mathbf{x}_i$. Also, let $y_{ic} = \mathbb{I}(y_i = c)$. Let $\mathbf{w}_C = 0$. Hence our one-line log-likelihood equation is

$$\begin{aligned}\text{NLL}(\mathbf{W}) &= -\log\left(\prod_{i=1}^{N}\prod_{c=1}^{C}\mu_{ic}^{y_{ic}}\right) + \lambda\|\mathbf{W}\|_F \\ &= \sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log\left(\mu_{ic}\right) + \lambda\,\text{tr}\left(\mathbf{W}^\top\mathbf{W}\right)\end{aligned}$$

hence

$$\begin{aligned}\nabla\text{NLL}(\mathbf{W}) &= \sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}(1 - \mathcal{S}(\boldsymbol{\eta}_i)_c)\mathbf{x}_i^\top + \lambda\mathbf{W} \\ &= X^\top(\boldsymbol{\mu} - \boldsymbol{y}) + \lambda\mathbf{W}\end{aligned}$$




Accuracy versus Lambda in Softmax Regression
Accuracy
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0
200
400
600
800
1000
Lambda