Math189R SP19
Homework 4
Monday, Feb 25, 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 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.

1 (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

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where

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

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 $p(\mathbf{x}_2|\mathbf{x}_1)$

(a) since we know the equation:

$$p(\mathbf{x}_1) = N(\mu_1, \Sigma_{11})$$

we can plug in the given values to find:

$$p(\mathbf{x}_1) = N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix})$$

(b) since we know the equation:

$$p(\mathbf{x}_2) = N(\mu_2, \Sigma_{22})$$

we can plug in the given values to find:

$$p(\mathbf{x}_1) = N(5, 14)$$

(c) The conditional distribution is given by:

$$p(\mathbf{x}_1|\mathbf{x}_2) = N(\mu_{1|2}, \Sigma_{1|2})$$

where $\mu_{1|2}$ is defined as:

$$\begin{aligned}
\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 5 \\ 11 \end{bmatrix} \begin{bmatrix} 14 \end{bmatrix}^{-1} (x_2 - 5) \\
&= \begin{bmatrix} \frac{5}{14} \\ \frac{11}{14} \end{bmatrix} (x_2 - 5)
\end{aligned}$$

and $\Sigma_{1|2}$ is defined as:

$$\begin{aligned}
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \begin{bmatrix} 5 \\ 11 \end{bmatrix} \begin{bmatrix} 14 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 11 \end{bmatrix}^\top \\
&= \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \begin{bmatrix} 14 \end{bmatrix}^{-1} \begin{bmatrix} 25 & 55 \\ 55 & 25 \end{bmatrix} \\
&= \begin{bmatrix} \frac{59}{14} & \frac{57}{14} \\ \frac{57}{14} & \frac{51}{14} \end{bmatrix}
\end{aligned}$$

(d) The conditional distribution is given by:

$$p(\mathbf{x}_2|\mathbf{x}_1) = N(\mu_{2|1}, \Sigma_{2|1})$$

where $\mu_{2|1}$ is defined as:

$$\begin{aligned}
\mu_{2|1} &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \\
&= 5 + \begin{bmatrix} 5 \\ 11 \end{bmatrix}^\top \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}^{-1} (x_1 - \begin{bmatrix} 0 \\ 0 \end{bmatrix}) \\
&= 5 + \begin{bmatrix} 5 & 11 \end{bmatrix} \begin{bmatrix} \frac{3}{14} & -\frac{4}{7} \\ -\frac{4}{7} & \frac{3}{7} \end{bmatrix} x_1 \\
&= 5 + \begin{bmatrix} -\frac{23}{14} & \frac{13}{7} \end{bmatrix} x_1
\end{aligned}$$

and $\Sigma_{2|1}$ is defined as:

$$
\begin{aligned}
\Sigma_{2|1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&= 14 - \begin{bmatrix} 5 \\ 11 \end{bmatrix}^{\top} \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 11 \end{bmatrix} \\
&= 14 - \begin{bmatrix} -\frac{23}{14} & \frac{13}{7} \end{bmatrix} \begin{bmatrix} 5 \\ 11 \end{bmatrix} \\
&= \frac{25}{14}
\end{aligned}
$$

■

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%.

(a) The equation for the negative log likelihood is:

$$\ell(\theta) = -\sum_i y_i log(h(x_i)) + (1 - y_i) log(1 - h(x_i)) + \frac{\lambda}{2}||\theta||_2^2$$

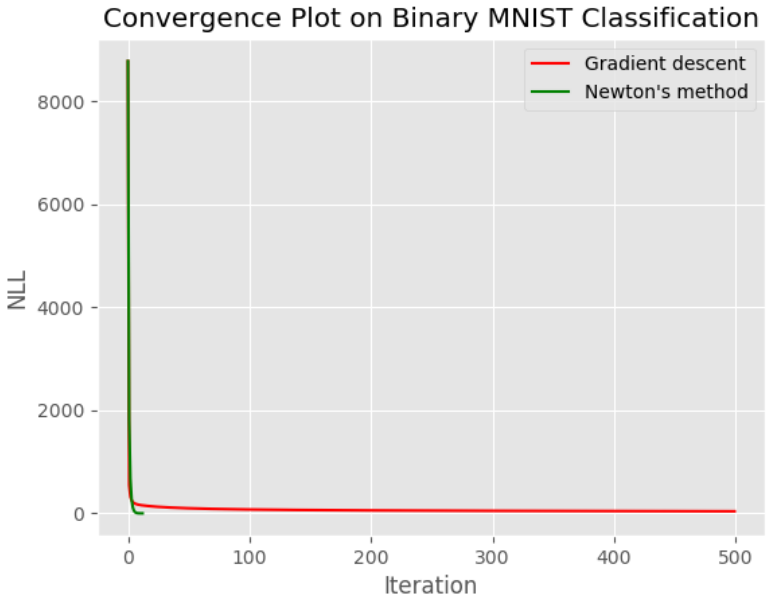
We can then take the gradient of the log likelihood equation:

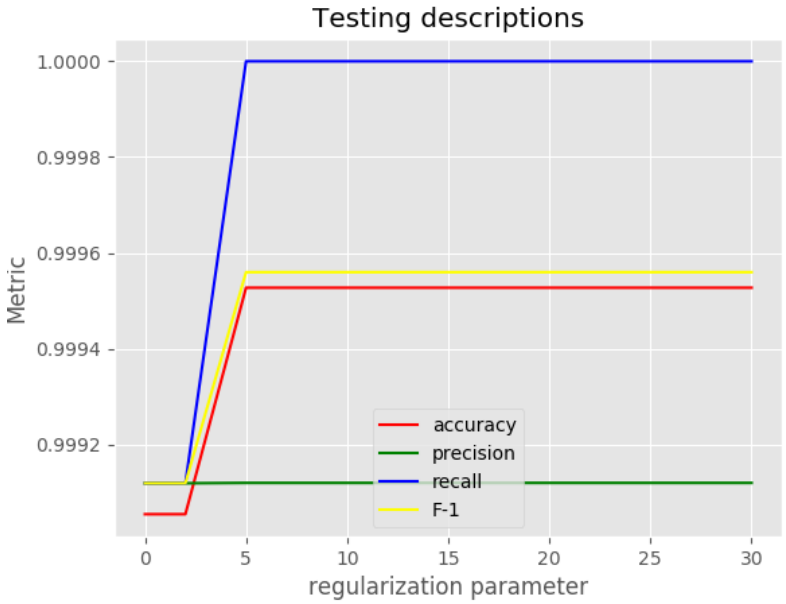
$$\begin{aligned}
\nabla_\theta \ell(\theta) &= -\nabla_\theta \sum_i y_i log(\sigma(\theta^T x_i)) + \nabla_\theta (1-y_i) log(1-\sigma(\theta^T x_i)) + \nabla_\theta \frac{\lambda}{2} ||\theta||_2 \\
&= \sum_i y_i \frac{1}{\sigma(\theta^T x_i)} (\sigma'(\theta^T x_i)) - (1-y_i) \frac{1}{\sigma(1-\theta^T x_i)} (-\sigma'(\theta^T x_i)) + \lambda\theta \\
&= \sum_i y_i \frac{\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))}{\sigma(\theta^T x_i)} x_i - (1-y_i) \frac{-\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))}{1-\sigma(\theta^T x_i)} x_i + \lambda\theta \\
&= \sum_i y_i (1-\sigma(\theta^T x_i)) x_i - (1-y_i)(-\sigma(\theta^T x_i)) x_i + \lambda\theta \\
&= \sum_i y_i x_i - y_i \sigma(\theta^T x_i) x_i - \sigma(\theta^T x_i) x_i + y_i \sigma(\theta^T x_i) x_i + \lambda\theta \\
&= \sum_i y_i x_i - \sigma(\theta^T x_i) x_i + \lambda\theta \\
&= X^\top (\sigma(\theta^T x_i) - y_i) + \lambda\theta
\end{aligned}$$

Now we find the Hessian Matrix :

$$\begin{aligned}
H_\theta &= \nabla_\theta (\nabla_\theta \ell(\theta))^T \\
&= \nabla_\theta (X^T (\sigma(\theta^T x_i) - y))^T + \lambda\theta \\
&= \nabla_\theta (\sigma(\theta^T x_i)^T X - y^T X) + \lambda\theta \\
&= \nabla_\theta (\sigma(\theta^T x_i))^T X + \lambda\theta \\
&= X^T diag(\sigma(X\theta)(1-\sigma(X\theta^T))) X + \lambda I
\end{aligned}$$

then we can produce our plots:





(b) We know the gradient of the negative log likelihood with a Gaussian prior on each column of W to be the following (some simplification below omitted since a detailed solution was done in part a):

$$\begin{aligned}\ell(\theta) &= -\sum_i \sum_c y_i log(\mu_{ic}) - \lambda tr(W^\top W) \\ \nabla_\theta \ell(\theta) &= \nabla_\theta \sum_i \sum_c y_{ic} log(\mu_{ic}) - \lambda tr(W^\top W) \\ &= X^\top(\mu - y) + \lambda W\end{aligned}$$

since we are predicting the label of each digit, we have $y \in 0, 1^{n \times c}$ where:

$$y_{ij} = \begin{cases} 1 & \text{if datum i is digit j} \\ 0 & \text{if otherwise} \end{cases} \qquad\qquad y1_c = 1_n$$

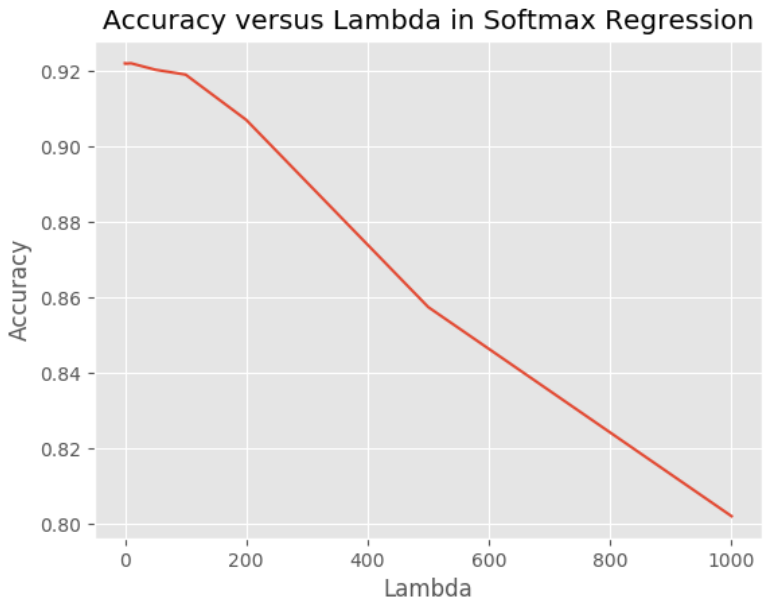
Since we are using softmax regression, we can define $\mu \in 0, 1^{n \times c}$ as $\mu_i = S(x_i)$. To express the softmax regression:

$$\begin{aligned}\mathbb{P}(y = c|x, W) &= \frac{1}{z} exp(w_c^\top x) \\ &= \frac{exp(w_c^\top x)}{z} \\ &= \frac{exp(w_c^\top x)}{\sum_i exp(w_i^\top x)}\end{aligned}$$

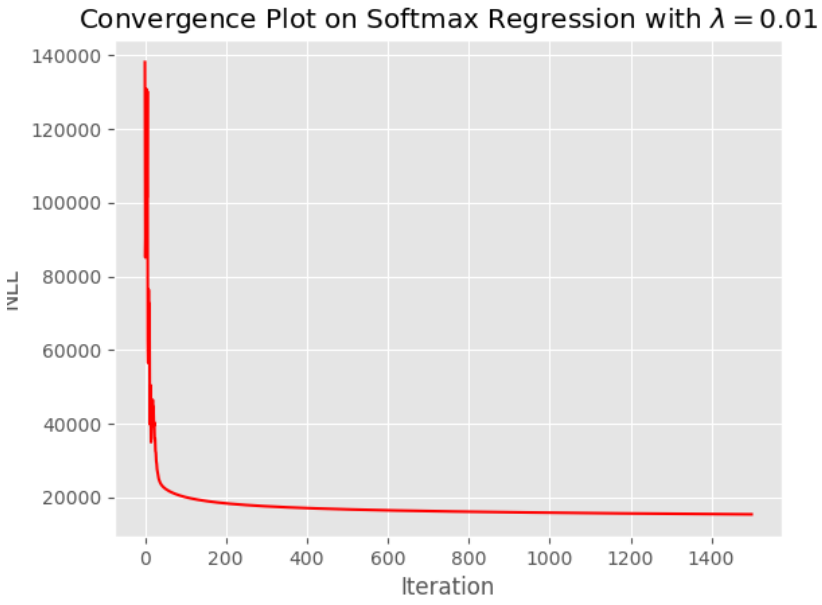
therefore we find:

$$S(x_i) = \frac{exp(w_c^\top x)}{1^\top exp(w_i^\top x)}$$

using gradient descent we get the following plot:



and we find the maximum test accuracy to be 0.922 with an optimal value of $\lambda = 0.01$. Using these parameters we have the following convergence plot:



■