Aimee Co
MA189 SP24
Homework 7
Mar 25, 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 for problem 2 can be found under the Resource tab on course website. The plot for problem 2 generated by the sample solution has been included in the starter files for reference. 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 11.3 - EM for Mixtures of Bernoullis) Show that the M step for ML estimation of a mixture of Bernoullis is given by

$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}}.$$

Show that the M step for MAP estimation of a mixture of Bernoullis with a $\beta(a, b)$ prior is given by

$$\mu_{kj} = \frac{\left(\sum_i r_{ik} x_{ij}\right) + a - 1}{\left(\sum_i r_{ik}\right) + a + b - 2}.$$

Given the log-likelihood function:

$$L = \sum_{i,k} r_{ik} \log(\mu_{kj}^{x_{ij}} (1 - \mu_{kj})^{1 - x_{ij}})$$

taking the derivative and setting it to zero gives:

$$0 = \sum_i r_{ik} \left(\frac{x_{ij}}{\mu_{kj}} - \frac{1 - x_{ij}}{1 - \mu_{kj}} \right)$$
$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}}.$$

Including the Beta prior in the log-likelihood produces:

$$L = \sum_{i,k} r_{ik} \log(\mu_{kj}^{x_{ij}} (1 - \mu_{kj})^{1 - x_{ij}}) + \log(\mu_{kj}^{a-1} (1 - \mu_{kj})^{b-1})$$

and we set the derivative to 0 to obtain:

$$
\begin{aligned}
0 &= \sum_i r_{ik} \left(\frac{x_{ij}}{\mu_{kj}} - \frac{1 - x_{ij}}{1 - \mu_{kj}} \right) + \frac{a-1}{\mu_{kj}} - \frac{b-1}{1-\mu_{kj}} \\
\mu_{kj} &= \frac{\left(\sum_i r_{ik} x_{ij}\right) + a - 1}{\left(\sum_i r_{ik}\right) + a + b - 2}.
\end{aligned}
$$

$\blacksquare$

2 (Lasso Feature Selection) In this problem, we will use the online news popularity dataset we used in hw2pr3. 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.

First, ignoring undifferentiability at $x = 0$, take $\frac{\partial |x|}{\partial x} = \text{sign}(x)$. Using this, show that $\nabla \|\mathbf{x}\|_1 = \text{sign}(\mathbf{x})$ where sign is applied elementwise. Derive the gradient of the ℓ_1 regularized linear regression objective

$$\text{minimize: } \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

Then, implement a gradient descent based solution of the above optimization problem for this data. Produce the convergence plot (objective vs. iterations) for a non-trivial value of λ. In the same figure (and different axes) produce a 'regularization path' plot. Detailed more in section 13.3.4 of Murphy, a regularization path is a plot of the optimal weight on the y axis at a given regularization strength λ on the x axis. Armed with this plot, provide an ordered list of the top five features in predicting the log-shares of a news article from this dataset (with justification).

First, consider the ℓ_1 norm of a vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ in $\mathbb{R}^n$, the ℓ_1 norm is defined as $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$. The gradient of $\|\mathbf{x}\|_1$ is a vector of the partial derivatives with respect to each component of $\mathbf{x}$.

$$\nabla \|\mathbf{x}\|_1 = \left[\frac{\partial |x_1|}{\partial x_1}, \frac{\partial |x_2|}{\partial x_2}, \ldots, \frac{\partial |x_n|}{\partial x_n} \right] = \text{sign}(\mathbf{x})$$

The objective function is given by:

$$\text{minimize: } f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

Now we take the gradient with respect to $\mathbf{x}$. The first term, $\|A\mathbf{x} - \mathbf{b}\|_2^2$, is differentiable, and its gradient can be computed:

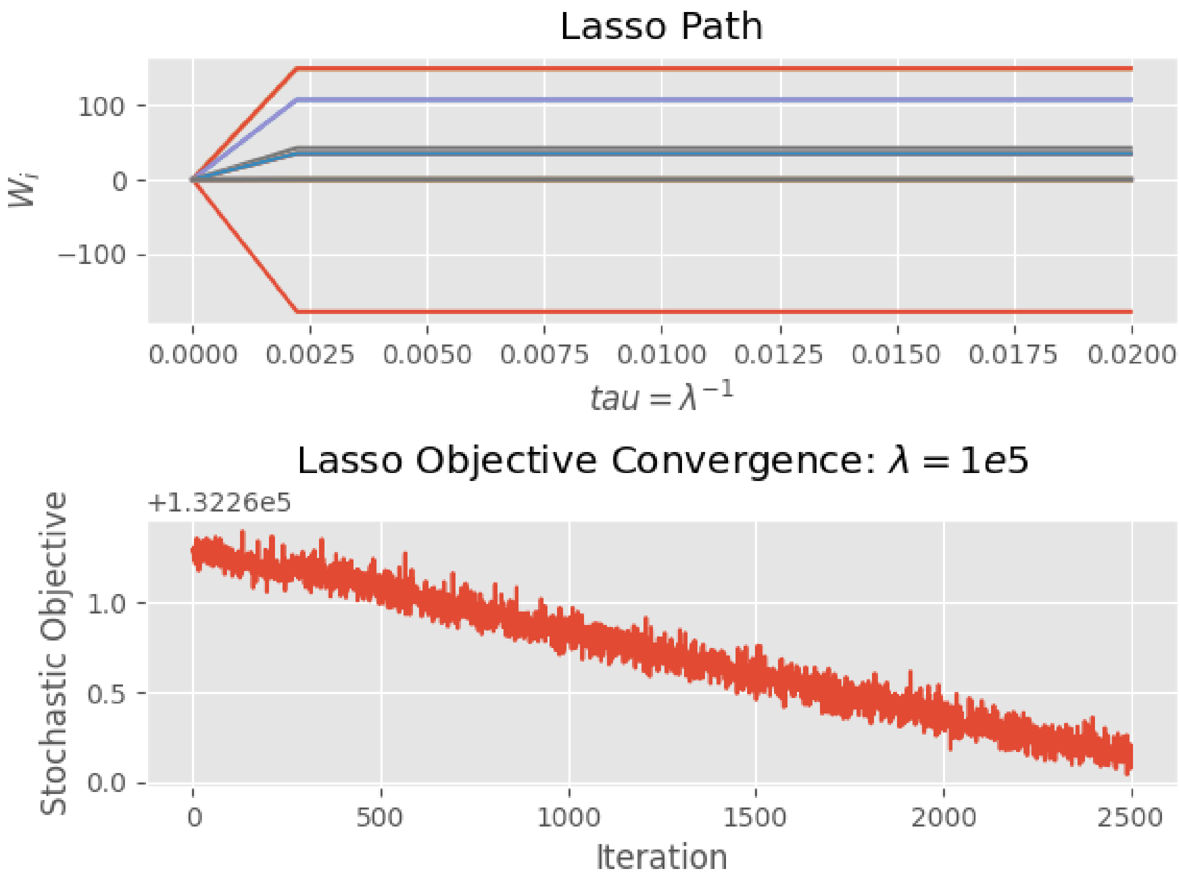
$$\nabla_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 = 2A^\top (A\mathbf{x} - \mathbf{b})$$

The second term, $\lambda \|\mathbf{x}\|_1$, is not differentiable everywhere due to the absolute value. However, outside of the non-differentiable points (at $x_i = 0$), we can represent the derivative in terms of the sign function.

Thus, the gradient of the objective function is:

$$\nabla f(\mathbf{x}) = 2A^\top (A\mathbf{x} - \mathbf{b}) + \lambda \text{sign}(\mathbf{x})$$

■



Lasso Path
100
0
−100
W_i
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
$tau = \lambda^{-1}$
Lasso Objective Convergence: $\lambda = 1e5$
+1.3226e5
1.0
0.5
0.0
Stochastic Objective
0
500
1000
1500
2000
2500
Iteration