Math189R SP19
Homework 7
Wednesday, Apr 8, 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 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

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

A mixture of Bernoullis is give by $p(x_i|z_i = k, \theta) = \Pi_{j=1}^{D} {x_{ij}}^{\mu_{jk}} (1 - x_{ij})^{(1-\mu_{jk})}$. We know that the complete data log likelihood is given by: (I took a look at the solutions to acquire this expression, I had the basic idea though)

$$l(\boldsymbol{\mu}) = \sum_i \sum_k r_{ik} \sum_j [x_{ij} \log(\mu_{jk}) + (1 - x_{ij}) \log((1 - \mu_{jk}))]$$

We will now take the derivative of this expression with respect to μ_{jk}

$$\frac{\partial l(\boldsymbol{\mu})}{\partial \mu_{jk}} = \sum_i r_{ik} [\frac{x_{ij}}{\mu_{jk}} - \frac{(1 - x_{ij})}{1 - \mu_{jk}}] = 0$$

Therefore,

$$\sum_i r_{ik} \frac{x_{ij}}{\mu_{jk}} = \sum_i r_{ik} \frac{1 - x_{ij}}{1 - \mu_{jk}}$$

So,

$$\sum_i r_{ik} x_{ij} = \mu_{jk} \sum_i r_{ik}$$

Which implies,

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

For the next step, we will proceed as above. Notice that a bernoulli distribution with a β prior looks like: $p(x_i|z_i = k, \boldsymbol{\mu}_k) = \Pi_{j=1}^{D} x_{ij}^{\mu_{jk}}(1 - x_{ij})^{(1-\mu_{jk})}\beta(\boldsymbol{\mu}_k|a, b)$ The log likelihood of this will be the log likelihood derived above + the log likelihood of the prior. This results in:

$$l(\boldsymbol{\mu}_k) = \sum_i \sum_k r_{ik} \sum_j [x_{ij} \log(\mu_{jk}) + (1 - x_{ij}) \log((1 - \mu_{jk}))]$$
$$+(a-1)\log(\mu_{jk}) + (b-1)\log(1-\mu_{jk}) - \log(B(a, b))$$

We proceed as above:

$$\frac{\partial l(\boldsymbol{\mu})}{\partial \mu_{jk}} = \sum_i r_{ik}[\frac{x_{ij}}{\mu_{jk}} - \frac{(1 - x_{ij})}{1 - \mu_{jk}}] + \frac{a-1}{\mu_{jk}} - \frac{b-1}{1-\mu_{jk}} = 0$$

$$\sum_i \frac{r_{ik}x_{ij} + a - 1}{\mu_{jk}} = \sum_i \frac{b + r_{ik} - r_{ik}x_{ij} - 1}{1 - \mu_{jk}}$$

So,

$$\sum_i r_{ik}x_{ij} + a - 1 - \mu_{jk}(r_{ik}x_{ij} + a - 1) = \sum_i \mu_{jk}(b + r_{ik} - r_{ik}x_{ij} - 1)$$

Therefore, $\mu_{jk} = \frac{\sum_i r_{ik}x_{ij} + a - 1}{\sum_i r_{ik} + a + b - 2}$, which is what we desire.

$\blacksquare$

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

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).

We know that $\frac{\partial |x|}{\partial x} = \text{sign}(x)$. Note that $||\mathbf{x}||_1 = \sum_i^n |x_i|^1$.
Therefore $\nabla \|\mathbf{x}\|_1 = \sum_i^n \frac{\partial}{\partial x_i} |x_i| = \sum_i^n \text{sign}(x_i) = \text{sign}(\mathbf{x})$

From this, we have:

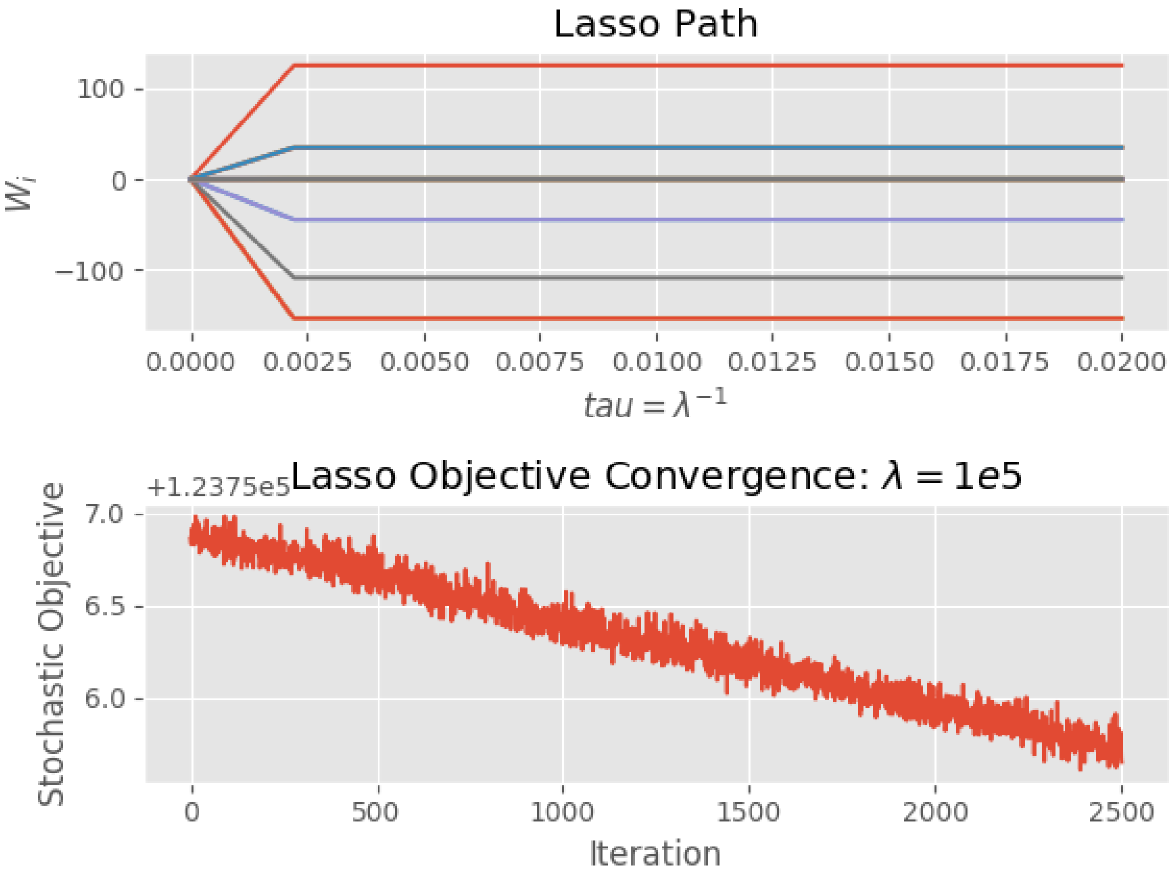
$$\nabla \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1 = \nabla (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - b) + \lambda \text{sign}(\mathbf{x})$$

$$= 2 \cdot A^T (A\mathbf{x} - b) + \lambda \text{sign}(\mathbf{x})$$

Upon implementing this, I found that the most important features are:

```
['timedelta' 'weekday_is_wednesday' 'weekday_is_thursday',
'weekday_is_friday' 'weekday_is_saturday']+
```

The convergence plot and regularization plots are attached below:


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: $\lambda = 1e5$
+1.2375e5
7.0
6.5
6.0
Stochastic Objective
0
500
1000
1500
2000
2500
Iteration

∎