Math of Big Data, Summer 2018

Prof: Gu

Name: **Forest Kobayashi**

HW #: **6**

Day: **Mon.** **Tue.** **Wed.** **Thu.** **Fri.**

Date: **05/22/2018**

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 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.

Problem 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

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

Solution:

(a) In 5.4.2.1, we have that for $X \sim \mathrm{Ber}(\theta)$, the log likelihood for a single sample is

$$\log\left(p(X \mid \theta)\right) = X \log(\theta) + (1 - X)\log(1 - \theta).$$

hence, for a mixture of bernoullis, we have

$$\begin{aligned}
\ell(\boldsymbol{\mu}) &= \sum_i \sum_k r_{ik} \log\left(\mathbb{P}(\mathbf{x}_i \mid \boldsymbol{\theta}_k)\right) \\
&= \sum_i \sum_k r_{ik} \sum_j \mathbf{x}_{ij} \log\left(\boldsymbol{\mu}_{kj}\right) + (1 - \mathbf{x}_{ij}) \log\left(1 - \boldsymbol{\mu}_{kj}\right)
\end{aligned}$$

differentiating, the k, j sums go away (since we're looking at a particular μ_{kj}), and we have

$$\begin{aligned}
0 &= \frac{\mathrm{d}\ell(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}_{kj}} \\
&= \sum_i r_{ik} \left(\frac{\mathbf{x}_{ij}}{\mu_{kj}} - \frac{1 - \mathbf{x}_{ij}}{1 - \mu_{kj}} \right) \\
\sum_i r_{ik} \frac{\mathbf{x}_{ij}}{\mu_{kj}} + \frac{\mathbf{x}_{ij}}{1 - \mu_{kj}} &= \sum_i r_{ik} \frac{1}{1 - \mu_{kj}} \\
\sum_i r_{ik} \mathbf{x}_{ij} \left(\frac{1 - \mu_{kj}}{\mu_{kj}} + 1 \right) &= \sum_i r_{ik} \\
\sum_i r_{ik} \mathbf{x}_{ij} \left(\frac{1}{\mu_{kj}} \right) &= \sum_i r_{ik} \\
\mu_{kj} &= \frac{\sum_i r_{ik} \mathbf{x}_{ij}}{\sum_i r_{ik}}
\end{aligned}$$

(b) Adding in a beta prior, we have

$$\begin{aligned}
\ell(\boldsymbol{\mu}) &= \sum_i \sum_k \left[r_{ik} \log\left(\mathbb{P}(\mathbf{x}_i \mid \boldsymbol{\theta}_k)\right) \right] + \log\left(\boldsymbol{\mu}_k\right) \\
&= \sum_i \sum_k r_{ik} \sum_j \left[\mathbf{x}_{ij} \log\left(\boldsymbol{\mu}_{kj}\right) + (1 - \mathbf{x}_{ij}) \log\left(1 - \boldsymbol{\mu}_{kj}\right) \right] + (a - 1)\log\left(\boldsymbol{\mu}_{kj}\right) + (b - 1)\log\left(1 - \boldsymbol{\mu}_{kj}\right)
\end{aligned}$$

hence

$$0 = \frac{\mathrm{d}\ell(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}_{kj}}$$

$$
\begin{aligned}
&= \sum_i \left(r_{ik} \left[\frac{\mathbf{x}_{ij}}{\mu_{kj}} \right] - \frac{1 - \mathbf{x}_{ij}}{1 - \mu_{kj}} \right) + \frac{a-1}{\mu_{kj}} - \frac{b-1}{1-\mu_{kj}} \\
&= \frac{\sum_i (r_{ik}\mathbf{x}_{ij}) + a - 1}{\mu_{kj}} - \frac{\sum_i (r_{ik}(1-\mathbf{x}_{ij})) + b - 1}{1 - \mu_{kj}} \\
&= \frac{1}{\mu_{kj}(1-\mu_{kj})} \left(\sum_i (r_{ik}\mathbf{x}_{ij}) + a - 1 \right)(1-\mu_{kj}) - \left(\sum_i (r_{ik}(1-\mathbf{x}_{ij})) + b - 1 \right)\mu_{kj}
\end{aligned}
$$

to avoid clutter, let $C_0 = \sum_i r_{ik}\mathbf{x}_{ij}$, and $C_1 = \sum_i r_{ik}$

$$
\begin{aligned}
&= \frac{1}{\mu_{kj}(1-\mu_{kj})} \Big[(C_0 + a - 1)(1 - \mu_{kj}) + (C_0 - C_1 - b + 1)\mu_{kj} \Big] \\
&= \frac{1}{\mu_{kj}(1-\mu_{kj})} \Big[C_0 + a - 1 - \cancel{C_0\mu_{kj}} - a\mu_{kj} + \mu_{kj} + \cancel{C_0\mu_{kj}} - C_1\mu_{kj} - b\mu_{kj} + \mu_{kj} \Big] \\
&= \frac{1}{\mu_{kj}(1-\mu_{kj})} \Big[C_0 + a - 1 - a\mu_{kj} - C_1\mu_{kj} - b\mu_{kj} + 2\mu_{kj} \Big] \\
(C_1 + a + b - 2)\mu_{kj} &= C_0 + a - 1 \\
\mu_{kj} &= \frac{C_0 + a - 1}{C_1 + a + b - 2}
\end{aligned}
$$

as desired. ■

Problem 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).

Solution:

We have

$$\begin{aligned} \nabla \|\mathbf{x}\|_1 &= \nabla \sum_i |x_i| \\ &= \text{sign}(\mathbf{x}) \end{aligned}$$

hence, with respect to $\mathbf{x}$

$$\begin{aligned} \nabla\Big(\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1\Big) &= \nabla\big(\mathbf{x}^\top A^\top A\mathbf{x} - 2\mathbf{b}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{b}\big) + \lambda \ \text{sign}(\mathbf{x}) \\ &= 2A^\top A\mathbf{x} - 2\mathbf{b}^\top A + \lambda \ \text{sign}(\mathbf{x}) \end{aligned}$$

After fixing a pernicious off-by-one error, we see that the most important features are

1. `timedelta`
2. `weekday_is_wednesday`
3. `weekday_is_thursday`
4. `weekday_is_friday`
5. `weekday_is_saturday`

which we can see in the lasso plot below, because they have the largest coefficients associated with them.

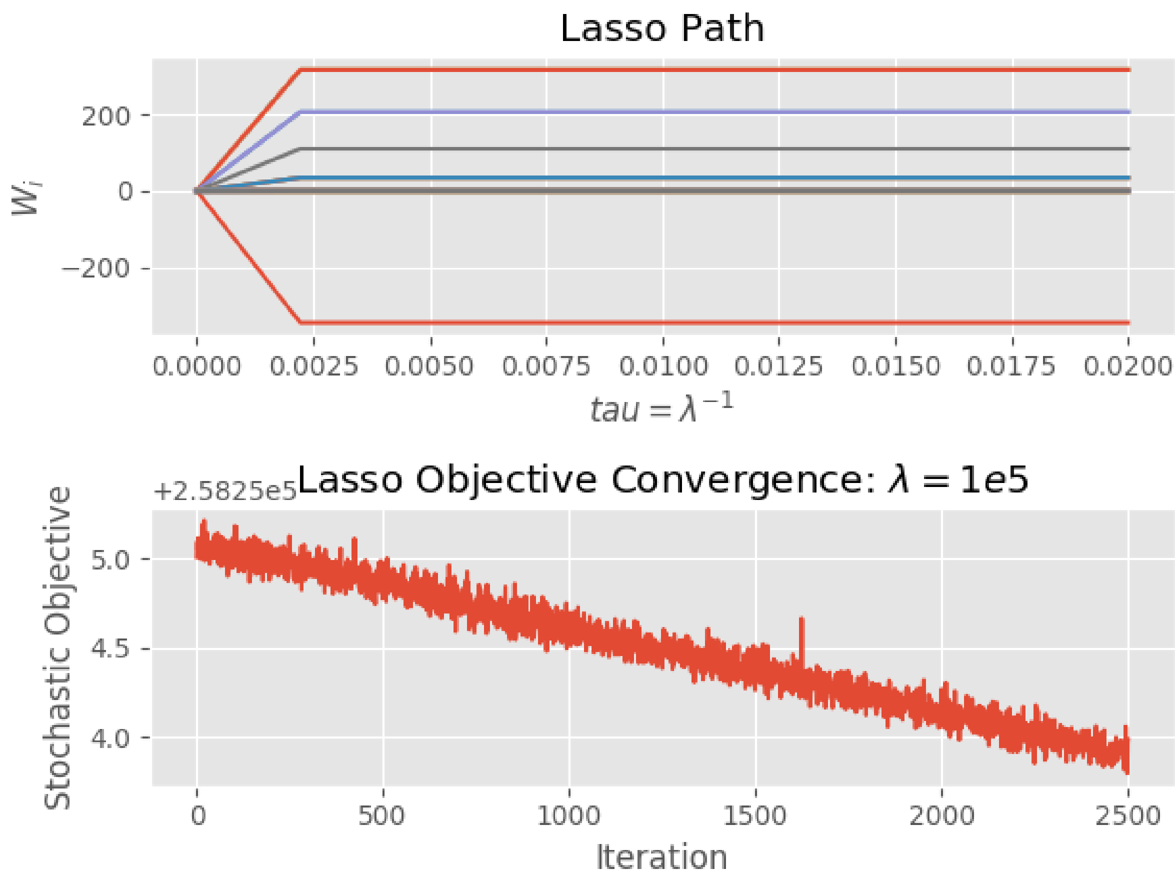

Figure 1: plot