Thitaree Tanprasert (Mint)
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

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

- **ML estimation:** Based on the textbook, equation 11.30, we know the log likelihood for the M step is

  $$\ell(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i | \boldsymbol{\theta}_k)$$

  In this case, we do not care about $\boldsymbol{\Sigma}_k$ and we want to sum the result over all k with an additional dimension to the variable x call j, so we can modify the equation as

  $$\ell(\boldsymbol{\mu}) = \sum_i \sum_k r_{ik} \sum_j x_{ij} \log \boldsymbol{\mu}_{kj} + (1 - x_{ij}) \log(1 - \boldsymbol{\mu}_{kj})$$

  Then, we can take a derivative of this equation with respect to $\boldsymbol{\mu}_{kj}$ and set the deriva-

tive equal to 0.

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

- **MAP estimation:** Similarly, we will start with equation 11.40 from the textbook. For this problem, we have two additional terms $\log p(\pi)$ and $\log p(\boldsymbol{\theta}_k)$. The first term is irrelevant to our maximization.

  We can therefore write the log likelihood for the M step with the given parameter a, b from the mixture of Bernoulli as follow:

$$\ell(\boldsymbol{\mu}) = \sum_i \sum_k r_{ik} \sum_j [x_{ij}\log\boldsymbol{\mu}_{kj} + (1-x_{ij})\log(1-\boldsymbol{\mu}_{kj}) + (a-1)\log\boldsymbol{\mu}_{kj} + (b-1)\log(1-\boldsymbol{\mu}_{kj})]$$

  Then, we will take a derivative of this in the same way that we did for the ML estimation.

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\mu}_{kj}} &= \sum_i r_{ik}\left(\frac{\mathbf{x}_{ij}}{\boldsymbol{\mu}_{kj}} + \frac{(1-\mathbf{x}_{ij})}{-(1-\boldsymbol{\mu}_{kj})} + \frac{a-1}{\boldsymbol{\mu}_{kj}} + \frac{b-1}{-(1-\boldsymbol{\mu}_{kj})}\right) \\
&= \sum_i r_{ik}\left(\frac{\mathbf{x}_{ij}-a-1}{\boldsymbol{\mu}_{kj}} + \frac{(1-\mathbf{x}_{ij})+b-1}{1-\boldsymbol{\mu}_{kj}}\right) \\
&= \sum_i \frac{r_{ik}(\mathbf{x}_{ij}-\boldsymbol{\mu}_{kj}) + (a-1)(1-\boldsymbol{\mu}_{kj}) + (b-1)\boldsymbol{\mu}_{kj}}{\boldsymbol{\mu}_{kj}(1-\boldsymbol{\mu}_{kj})} \\
0 &= \sum_i r_{ik}(\mathbf{x}_{ij}-\boldsymbol{\mu}_{kj}) + a - 1 + -a\boldsymbol{\mu}_{kj} + \boldsymbol{\mu}_{kj} + -b\boldsymbol{\mu}_{kj} + \boldsymbol{\mu}_{kj} \\
&= (\sum_i r_{ik}\mathbf{x}_{ij}) - (\sum_i r_{ik} + a + b + 2)\boldsymbol{\mu}_{kj} + a - 1 \\
(\sum_i r_{ik} + a + b + 2)\boldsymbol{\mu}_{kj} &= \sum_i r_{ik}\mathbf{x}_{ij} + a - 1 \\
\boldsymbol{\mu}_{kj} &= \frac{\sum_i r_{ik}\mathbf{x}_{ij} + a - 1}{\sum_i r_{ik} + a + b + 2}
\end{aligned}$$

■

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).

Based on section 13.3.2 in the textbook, we will use proximal gradient with hard thresholding. As mentioned in the problem statement, this will cause undifferentiability at 0. However, at each x, we can get a formula for iteration:

$$\mathbf{x}_{i+1} = \text{prox}_l(\mathbf{x}_i - l\nabla f(\mathbf{x}))$$

where f defined by equation 13.43 and l is the learning rate of the optimizer. Elementwise, we take $\frac{\partial ||\mathbf{x}||_1}{\partial \mathbf{x}_i} = \text{sign}(\mathbf{x}_i)$ as given in the problem statement. This is equivalent to saying $\nabla ||\mathbf{x}||_1 = \text{sign}(x)$ as desired.

Then, combine with the equation for our iteration, we see that

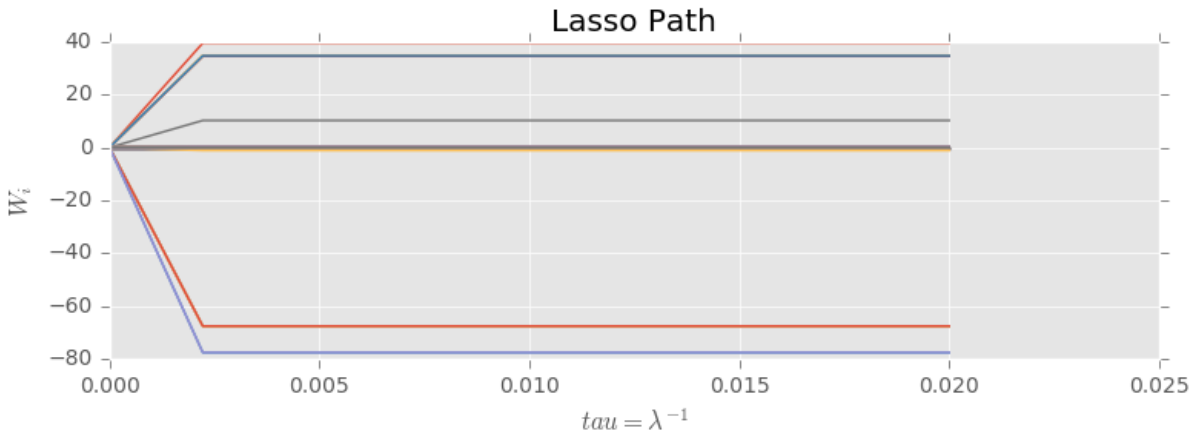
$$\begin{aligned} \nabla ||A\mathbf{x} - \mathbf{b}||_2^2 + \lambda ||\mathbf{x}||_1 &= (\nabla \mathbf{x}^\top A^\top A\mathbf{x} - 2\mathbf{b}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{b}) + \lambda ||\mathbf{x}||_1 \\ &= 2A^\top A\mathbf{x} - 2\mathbf{b}^\top A + \lambda \text{sign}(x) \end{aligned}$$

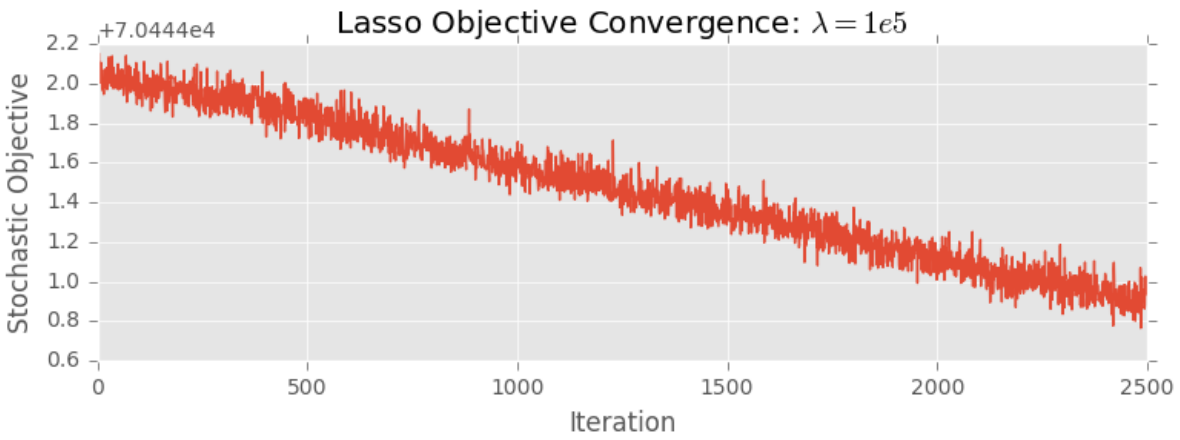
After running the completed code on the news article data, the top five features in predicting the log-shares are

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

.

The generate plots are shown below (on the next page):


Lasso Path
40
20
0
−20
−40
−60
−80
W_i
0.000
0.005
0.010
0.015
0.020
0.025
$tau = \lambda^{-1}$


Lasso Objective Convergence: $\lambda = 1e5$
+7.0444e4
Stochastic Objective
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0
500
1000
1500
2000
2500
Iteration

■