Colin Adams
Math189R SU20
Homework 7
July 2020

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

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

Okay, so we are maximizing a weighted sum of $p(\mathbf{x}_i|\boldsymbol{\theta}_k)$ which is the same thing as maximizing its logarithm. To do that, we take the logarithm of it

$$\sum_{ij} r_{ij} \log p(\mathbf{x}_i|\boldsymbol{\theta}_j) = \sum_{ij} r_{ij} \sum_k \mathbf{x}_{ij} \log \boldsymbol{\mu}_{jk} + (1-\mathbf{x}_{ij}) \log\left(1-\boldsymbol{\mu}_{jk}\right)$$

which we take the gradient of with respect to $\boldsymbol{\mu}_{jk}$ which gives us

$$\nabla_{\boldsymbol{\mu}_{jk}} \sum_{ij} r_{ij} \log p(\mathbf{x}_i|\boldsymbol{\theta}_j) = \sum_i r_{ij} \left(\frac{\mathbf{x}_{ij} - \boldsymbol{\mu}_{jk}}{\boldsymbol{\mu}_{jk}(1-\boldsymbol{\mu}_{jk})} \right).$$

Setting this to zero, we find that

$$\sum_i r_{ij} \mathbf{x}_{ij} = \sum_i r_{ij} \boldsymbol{\mu}_{jk} \qquad \Rightarrow \boldsymbol{\mu}_{jk} = \frac{1}{\sum_i r_{ij}} \sum_i r_{ij} \mathbf{x}_{ij}$$

which we wished to show.

Now if we do a similar thing, except we have a prior belief (i.e. the MAP maximization), we hope to maximize a mixture of $p(\mathbf{x}_i|\boldsymbol{\mu}_k)p(\boldsymbol{\mu}_k)$ where $p(\boldsymbol{\mu}_k) \sim \beta(a,b)$. Again, this is the same thing as maximizing the log of that, so writing that out we have

$$\sum_{ij} r_{ij} \log p(\mathbf{x}_i|\boldsymbol{\mu}_k) p(\boldsymbol{\mu}_k) = \sum_{ij} r_{ij} \log p(\mathbf{x}_i|\boldsymbol{\mu}_k) + \log p(\boldsymbol{\mu}_k).$$

The first term is the same as above and the second term is very similar, so we have

$$\sum_{ij} r_{ij} \log p(\mathbf{x}_i|\boldsymbol{\mu}_k) + \log p(\boldsymbol{\mu}_k) = \sum_{ij} r_{ij} \sum_k \mathbf{x}_{ij} \log \boldsymbol{\mu}_{jk} + (1-\mathbf{x}_{ij}) \log\Big(1-\boldsymbol{\mu}_{jk}\Big) \\ + (a-1)\log \boldsymbol{\mu}_{jk} + (b-1)\log\Big(1-\log \boldsymbol{\mu}_{jk}\Big).$$

Taking the derivative of this monstrosity with respect to $\boldsymbol{\mu}$ gives us, after a few pages of algebra and some help from Mathematica,

$$\nabla_{\boldsymbol{\mu}}(\text{above}) = \frac{1}{\boldsymbol{\mu}_{jk}(1-\boldsymbol{\mu} jk)} \left[\sum_i r_{ij} \mathbf{x}_{ij} + (a-1) - \boldsymbol{\mu}_{jk} \left((a-1) + (b-1) + \sum_i r_{ij} \right) \right].$$

Set $[\cdot]$'s to zero, and we can solve for $\boldsymbol{\mu}_{jk}$ which gives us

$$\boldsymbol{\mu}_{jk} = \frac{\sum_i r_{ij} \mathbf{x}_{ij} + (a-1)}{(a-1) + (b-1) + \sum_i r_{ij}}$$

which is what we wished to show.

■

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

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

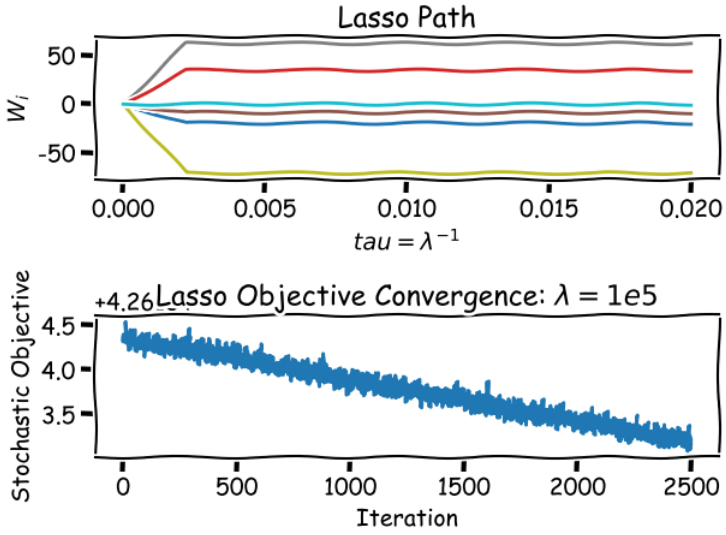
Then, implement a gradient descent based solution of the above optimization problem for this data. Produce the convergence plot (objective vs. iterations) for a non-trivial value of λ. In the same figure (and different axes) produce a 'regularization path' plot. Detailed more in section 13.3.4 of Murphy, 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).

Doing the math bit, we have a given data vector $\mathbf{x}$, so

$$\nabla \|\mathbf{x}\|_1 = \nabla \sum_i |x_i| = \sum_i \text{sign}(x_i)\hat{\mathbf{e}} = \text{sign}(\mathbf{x}).$$

Okay, if we hope to minimize $\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$ then we just take the gradient and set to zero and all that so, using our matrix calculus rules we get

$$0 = \nabla\left(\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1\right) = \lambda \text{sign}(\mathbf{x}) + 2(\mathbf{A}^T\mathbf{Ax} - \mathbf{b}^T\mathbf{A}) \qquad \Rightarrow \frac{\lambda}{2}\text{sign}(\mathbf{x}) = \mathbf{A}^T\mathbf{Ax} - \mathbf{b}^T\mathbf{A}.$$



Apparently the most important features are `['timedelta', 'weekday_is_wednesday', 'weekday_is_thu` `'weekday_is_friday', 'weekday_is_saturday']` and our plot is given above.