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

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

1. For a mixture of Bernoullis, the probability distribution is given by:

$$p(\mathbf{x}_i | z_i = k, \boldsymbol{\theta}) = \prod_{j=1}^{D} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{1 - x_{ij}}, \tag{1}$$

where we suppose above that our data consists of D-dimensional bit vectors. Hence, the complete data log likelihood (omitting indeces in summations) is:

$$\ell(\boldsymbol{\mu}) = \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i, \boldsymbol{\theta}_k) = \sum_i \sum_k r_{ik} \left[\sum_j \mathbf{x}_{ij} \log \boldsymbol{\mu}_{kj} + (1 - \mathbf{x}_{ij}) \log(1 - \boldsymbol{\mu}_{kj}) \right] \tag{2}$$

which we get using the properties of the logarithm function. Taking the derivative with respect to $\boldsymbol{\mu}_{kj}$ and setting it equal to zero, we have

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_{kj}} = \sum_i r_{ik} \left[\frac{\mathbf{x}_{ij}}{\mu_{kj}} - \frac{1 - \mathbf{x}_{ij}}{1 - \mu_{kj}} \right] = 0. \tag{3}$$

We now wish to solve for μ_{kj}. Hence, the above implies

$$\frac{1}{\mu_{kj}(1-\mu_{kj})} \cdot \sum_i r_{ik} \left[\mathbf{x}_{ij} - \mu_{kj}\right] = 0, \tag{4}$$

which after rearrangement, gives us

$$\mu_{kj} = \frac{\sum_i r_{ik}\mathbf{x}_{ij}}{\sum r_{ik}} \tag{5}$$

the desired result.

2. The complete data log likelihood with prior is

$$\ell(\boldsymbol{\mu}) = \left[\sum_i \sum_k r_{ik} \log \pi_{ik} + \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\mu}_k)\right] + \log p(\pi) + \sum_k \log p(\boldsymbol{\mu}_k) \tag{6}$$

$$= \Bigg[\sum_i \sum_k r_{ik} \log \pi_{ik} + \sum_i \sum_k r_{ik} \left(\sum_j \mathbf{x}_{ij} \log \boldsymbol{\mu}_{kj} + (1-\mathbf{x}_{ij}) \log(1-\boldsymbol{\mu}_{kj})\right) + \log p(\pi) + (a-1)\log \boldsymbol{\mu}_{kj} + (b-1)\log(1-\boldsymbol{\mu}_{kj})\Bigg] \tag{7}$$

. Taking the derivative with respect to $\boldsymbol{\mu}_{kj}$ and setting that equal to zero we have:

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_{kj}} = \sum_i r_{ik} \left[\frac{\mathbf{x}_{ij}}{\boldsymbol{\mu}_{kj}} - \frac{1-\mathbf{x}_{ij}}{1-\boldsymbol{\mu}_{kj}} + \frac{a-1}{\boldsymbol{\mu}_{kj}} + \frac{b-1}{1-\boldsymbol{\mu}_{kj}}\right] = 0. \tag{8}$$

Solving for $\boldsymbol{\mu}_{kj}$ we have:

$$\sum_i r_{ik} \left[\frac{\mathbf{x}_{ij}+a-1}{\boldsymbol{\mu}_{kj}} - \frac{1-\mathbf{x}_{ij}+b-1}{1-\boldsymbol{\mu}_{kj}}\right] = \sum_i r_{ik} \left[\frac{(1-\boldsymbol{\mu}_{kj})(\mathbf{x}_{ij}+a-1)}{(1-\boldsymbol{\mu}_{kj})\boldsymbol{\mu}_{kj}} - \frac{\boldsymbol{\mu}_{kj}(1-\mathbf{x}_{ij}+b-1)}{\boldsymbol{\mu}_{kj}(1-\boldsymbol{\mu}_{kj})}\right] \tag{9}$$

$$= \frac{1}{\boldsymbol{\mu}_{kj}(1-\boldsymbol{\mu}_{kj})} \sum_i r_{ik} \left[\mathbf{x}_{ij} + a - 1 - \boldsymbol{\mu}_{kj}\mathbf{x}_{ij} - \boldsymbol{\mu}_{kj}a - \boldsymbol{\mu}_{kj} - \boldsymbol{\mu}_{kj} + \boldsymbol{\mu}_{kj}\mathbf{x}_{ij} - \boldsymbol{\mu}_{kj}b + \boldsymbol{\mu}_{kj}\right] \tag{10}$$

which, after rearranging, gives

$$= \frac{1}{\boldsymbol{\mu}_{kj}(1-\boldsymbol{\mu}_{kj})} \left[\sum_i r_{ik}\mathbf{x}_{ij} + a - 1 - \left(\sum_i r_{ik} + a + b - 2\right)\right] \tag{11}$$

and then:

$$\boldsymbol{\mu}_{kj} = \frac{\left(\sum_i r_{ik}\mathbf{x}_{ij}\right) + a - 1}{\left(\sum_i r_{ik}\right) + a + b - 2} \tag{12}$$

as desired.

■

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).

Recall that $||\mathbf{x}||_1 = \sum_i |\mathbf{x}_i|$. Hence

$$\nabla ||\mathbf{x}||_1[i] = \frac{\partial}{\partial \mathbf{x}_i} \sum_i |\mathbf{x}_i| = \text{sign}(\mathbf{x}_i) \tag{13}$$

(where we are choosing to abuse the computer science index notation on the i'th component of the gradient vector above for less writing. However, on closer examination, it seems we have written more in doing so than in not.) which shows that $\nabla ||\mathbf{x}||_1 = \text{sign}(\mathbf{x})$ (we have shown this component wise above).

Deriving the gradient can then be shown as follows:

$$\nabla \left[||A\mathbf{x} - \mathbf{b}||_2^2 + \lambda ||\mathbf{x}||_1 \right] = 2A^T A\mathbf{x} - 2\mathbf{b}^T A + \lambda \text{sign}(\mathbf{x}) \tag{14}$$

using the fact that $\nabla(\mathbf{x}^T A^T A\mathbf{x}) = 2A^T A\mathbf{x}$.

See github for plot. Also, the top five features are: 'timedelta' 'weekday_is_wednesday' 'weekday_is_thursday' 'weekday_is_friday' 'weekday_is_saturday'

■