Andy Liu
Math189R SU20
Homework 7
Thursday, July 9, 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

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

(a) Note that the log likelihood for a mixture of Bernoullis distribution, given a $\boldsymbol{\mu}$, is equal to $\sum\sum r_{ik} \sum \boldsymbol{\mu}_{kj} + (1-\mathbf{x}_{ij})\log(1-\boldsymbol{\mu}_{kj})$. To find the optimal value of $\boldsymbol{\mu}_{kj}$ given our inputs (which is the goal of the M step for ML estimation of a mixture of Bernoullis), we simply differentiate the log likelihood with respect to $\boldsymbol{\mu}$ and set the derivative equal to 0.

Taking the derivative of the log likelihood with respect to μ, we find that it is equal to $\sum_i r_{ik}(\frac{\mathbf{x}_{ij}}{\mu_{kj}} - \frac{1-\mathbf{x}_{ij}}{1-\mu_{kj}})$, which we wish to set to 0. Note that we can multiply the entire equation by $\mu_{kj}(1-\mu_{kj}$, which is not equal to 0, and set the resulting summation equal to 0 to find our optimal μ_{kj}. Note that multiplying by $\boldsymbol{\mu}_{kj}(1-\boldsymbol{\mu}_{kj})$ yields $\sum_i r_{ik}(\mathbf{x}_{ij})(1-\mu_{kj}) - (1-\mathbf{x}_{ij})(\mu_{kj}) = 0$. Note that the left-hand side of this equation simplifies to $\sum_i r_{ik}(\mathbf{x}_{ij} - \mu_{kj})$, so it suffices to set this equal to 0.

However, this is equal to 0 iff $\sum_i r_{ik}\mathbf{x}_{ij} = \sum_i r_{ik}\mu_{kj}$, which occurs when $\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{r_{ik}}$, as desired.

(b) To compute the log likelihood of the mixture of Bernoullis with $\beta(a,b)$ prior, we simply add the log of the prior to our log likelihood term, and similarly calculate the optimal value of μ_{kj}.

Thus, the expression that we seek to optimize is $\sum \sum r_{ik}(\sum \boldsymbol{\mu}_{kj} + (1 - \mathbf{x}_{ij}) \log(1 - \boldsymbol{\mu}_{kj})) + (a-1)\log(\mu_{kj}) + (b-1)\log(1-\mu_{kj}))$.

Similarly to the previous problem, we differentiate with respect to μ_{kj} and multiply by $\mu_{kj}(1-\mu_{kj})$ to get $\sum_i r_{ik}(\mathbf{x}_{ij} - \mu_{kj}) + \mu_{kj}(2-a-b) + (a-1)$, which we can set to 0 to compute our optimal μ_{kj}. Note that we can rearrange the equation $\sum_i r_{ik}(\mathbf{x}_{ij} - \mu_{kj}) + \mu_{kj}(2-a-b) + (a-1) = 0$ to isolate μ_{kj}, which yields $(\sum_i r_{ik}\mu_{kj}) + (\mu_{kj})(a+b-2) = \sum_i r_{ik}\mathbf{x}_{ij} + (a-1)$.

Dividing out by $\sum_i r_{ik} + a + b - 2$ on both sides yields the desired result: $\mu_{kj} = \frac{(\sum_i r_{ik} x_{ij} + a - 1}{(\sum_i r_{ik}) + a + b - 2}$.

■

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).

The top five features are ['timedelta', 'weekday-is-wednesday, 'weekday-is-thursday', 'weekday-is-friday' 'weekday-is-saturday'], as can be seen by running the code in the repository. The convergence and regularization path plots are also in the hw7 folder in the repository. ■