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

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

Given the complete data log likelihood

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

$$= \sum_i \sum_k r_{ik} \left[\sum_j x_{ij} \log \mu_{kj} + (1 - x_{ij}) \log(1 - \mu_{kj}) \right]$$

where i indexes the data points, k indexes the mixture components, and j indexes the dimensions of the D-dimensional binary vectors. We derive with respect to μ_{kj} as follows:

$$\frac{\partial \ell}{\partial \mu_{kj}} = \sum_i r_{ik} \left[\frac{x_{ij}}{\mu_{kj}} - \frac{1 - x_{ij}}{1 - \mu_{kj}} \right]$$

$$= \sum_i r_{ik} \left[\frac{x_{ij} - \mu_{kj}}{\mu_{kj}(1 - \mu_{kj})} \right]$$

$$= \frac{1}{\mu_{kj}(1 - \mu_{kj})} \sum_i r_{ik}(x_{ij} - \mu_{kj}) = 0.$$

This yields the optimality criterion

$$\sum_i r_{ik} x_{ij} = \mu_{kj} \sum_i r_{ik},$$

leading to the solution that we seek.

(b) Incorporating the Log Prior

Considering the complete data log likelihood with the log prior (omitting the π terms as we maximize regardless of them)

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

$$= \sum_i \sum_k r_{ik} \left[\sum_j x_{ij} \log \mu_{kj} + (1 - x_{ij}) \log(1 - \mu_{kj}) \right] + (a - 1) \log \mu_{kj} + (b - 1) \log(1 - \mu_{kj}).$$

Differentiating yields

$$\frac{\partial \ell}{\partial \mu_{kj}} = \sum_i \left[r_{ik} x_{ij} + a - 1 - r_{ik}(1 - x_{ij}) + b - 1 \right] \left[\frac{1}{\mu_{kj}} - \frac{1}{1 - \mu_{kj}} \right]$$

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

$$= \frac{1}{\mu_{kj}(1 - \mu_{kj})} \left[\sum_i r_{ik} x_{ij} - \left(\sum_i r_{ik} + a + b - 2 \right) \mu_{kj} + a - 1 \right] = 0.$$

This results in the optimality condition

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

providing the sought-after result. Note that setting $a = b = 1$ recovers the original maximum likelihood estimate, as expected since $\beta(1,1)$ is a uniform distribution over [0,1] implying the absence of a prior.

■

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).

A more nuanced discussion on the topic of proximal gradient methods for learning is elaborated upon in resources such as the Proximal Gradient Methods for Learning Wikipedia entry and Boyd et al.'s documentation on proximal algorithms. It is proposed to implement the gradient descent step through a thresholding operation defined by the proximal operator prox_γ, represented as:

$$\text{prox}_\gamma(x_i) = \begin{cases} x_i - \gamma & \text{if } x_i > \gamma \\ 0 & \text{if } |x_i| \leq \gamma \\ x_i + \gamma & \text{if } x_i < -\gamma \end{cases} \tag{1}$$

This operator is systematically employed in the iterative update given by:

$$x_{i+1} = \text{prox}_\gamma(x_i - \gamma \nabla f(x_i)), \tag{2}$$

where γ denotes the learning rate.

It is evident that:

$$\frac{\partial \|x\|_1}{\partial x_i} = \frac{\partial}{\partial x_i} \sum_j |x_j| = \text{sign}(x_i). \tag{3}$$

Subsequently, we can infer:

$$\nabla \|Ax - b\|_2^2 + \lambda \|x\|_1 = 2A^T Ax - 2A^T b + \lambda \text{sign}(x). \tag{4}$$

The images will be generated using Python Code. ∎