Eli Weissler

Math189R SU17
Homework 7
Monday, June 12, 2017

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

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

In the M step, we are maximizing $Q(\theta, \theta^{(t-1)})$ w.r.t. θ_K.

Mixture coefficients

$$Q(\theta, \theta^{(t-1)}) = \underbrace{\sum_i}_{\text{datapoint}} \underbrace{\sum_K}_{\text{mixture component}} r_{iK} \log \pi_K + \sum_i \sum_K r_{iK} \log P(x_i \mid \theta_K)$$

$r_{iK} = P(z_i = K \mid x_i, \theta^{(t-1)})$ the resposibility of cluster K for the data

Only the second part depends on θ_K

$$\rightarrow \sum_i \sum_K r_{iK} \sum_j x_{ij} \log \mu_{Kj} + (1 - x_{ij}) \log(1 - \mu_{Kj})$$

Bernouli category index

■

Now to minimize take the derivatie w.r.t. μ_{kj}

$$\frac{\partial Q(\theta, \theta^{(t-1)})}{\partial \mu_{kj}} = \sum_i r_{ik} \left(\frac{x_{ij}}{\mu_{kj}} - \frac{1-x_{ij}}{1-\mu_{kj}} \right) = \sum_i r_{ik} \frac{x_{ij}(1-\mu_{kj}) - \mu_{kj}(1-x_{ij})}{\mu_{kj}(1-\mu_{kj})}$$

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

$$\rightarrow \sum_i r_{ik} x_{ij} = \sum_i r_{ik} \mu_{kj} = \mu_{kj} \sum_i r_{ik} \rightarrow \boxed{\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}}}$$

For the MAP estimation, we add a prior term to the likelihood function. The part that depends on μ_k is now

$$\sum_i \sum_k r_{ik} \log P(x_i | \theta_k) + P(\mu_k)$$

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

Take deiv w.r.t. μ_{kj}

$$\frac{\partial}{\partial \mu_{kj}} \rightarrow \frac{1}{\mu_{kj}(1-\mu_{kj})} \sum_i \left(r_{ik}(x_i - \mu_{kj}) \right) + (1-\mu_{kj})(a-1) - \mu_{kj}(b-1) = 0$$

$$\rightarrow \sum_i (r_{ik} x_i) + a - 1 = \sum_i (r_{ik} \mu_{kj}) + \mu_{kj} a - \mu_{kj} + \mu_{kj} b - \mu_{kj}$$

$$\rightarrow \boxed{\frac{\sum_i (r_{ik} x_i) + a - 1}{\sum_i (r_{ik} \mu_{kj}) + a + b - 2} = \mu_{kj}}$$

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).

To show $\frac{\partial |x|}{\partial x} = \text{sign}(x)$ recall that for $x < 0$ $|x| = -x$ and $x > 0$ $|x| = x$

$$\frac{\partial x}{\partial x} = 1 \qquad \frac{\partial -x}{\partial x} = -1$$

Therefore $\frac{\partial |x|}{\partial x} = \text{sign}(x)$.

Now consider a vector $\|\vec{x}\|_1 = \sum_i |x_i|$

$$\frac{\partial \|\vec{x}\|_1}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_i |x_i| = \text{sign}(x_j)$$

Therefore $\nabla \|\vec{x}\|_1 = \text{sign}(\vec{x})$ where sign is applied elementwise.

Now we can take the gradient of the objective

$$\nabla_x \left[\|Ax - b\|_2^2 + \lambda \|x\|_1\right] = 2A^T(Ax - b) + \lambda\,\text{sign}(x) = 2A^TAx - 2A^Tb + \lambda\,\text{sign}(x) = 0$$

$$\rightarrow x = (A^TA)^{-1}\left[A^Tb + \frac{\lambda}{2}\text{sign}(x)\right] = (A^TA)^{-1}\left[A^Tb + \lambda\,\text{sign}(x)\right], \text{ redefining } \lambda \text{ wlog.}$$

For the actual gradient descent instead of using this gradient, we must do proximate gradient descent, or so the answer key says.