Joseph Gardi
Big data
Homework 7
Due: Monday April 7, 2019

**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

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

I used the solution.
The likelihood is,

$$\sum_{-} \sum_{k} r_{ik} log P(x_i|\theta_k)$$
$$= sum_i \sum_k r_{ik} \sum_j (x_{ij} log \mu_{kj} + (1 - x_{ij}) log(1 - \mu_{kj}))$$

To optimize I will take the derivative with respect to $\mu_{kj}$, set it to zero and solve,

$$\frac{\partial L}{\partial \mu_{kj}} = \sum_i r_{ij} \left(\frac{x_{ij}}{\mu_{kj}} - \frac{1 - x_{ij}}{1 - \mu_{kj}}\right)$$
$$= \sum_i r_{ij} \left(\frac{x_{ij} - \mu kj}{\mu_{kj}(1 - \mu_{kj})}\right)$$
$$= \frac{1}{\mu_{kj}(1 - \mu_{kj})} \sum_i r_{ik}(x_{ij} - \mu_{kj}) = 0 \implies \sum_i r_{ik} x_{ij} = \mu_{kj} \sum_i r_{ik}$$

this gives the desired result.
Now for the MAP part. The likihood with prior is,

$$\sum_{-} \sum_{k} r_{ik} log P(x_i|\theta_k) + log P(\mu_k)$$
$$= sum_i \sum_k r_{ik} \sum_j (x_{ij} log \mu_{kj} + (1 - x_{ij}) log(1 - \mu_{kj})) + (a - 1) log \mu_{kj} + (b -$$

Take derivatives and set to zero to find optmial $\mu_{kj}$,

$$\begin{aligned}
\frac{\partial L}{\partial \mu} &= \sum_i \left(\frac{r_{ik}x_{ij} + a - 1}{\mu_{kj}} - \frac{r_{ik}(1 - x_{ij}) + b - 1}{1 - \mu_{kj}}\right) \\
&= \frac{1}{\mu_{kj}(1 - \mu_{kj})} \sum_i r_{ik}x_{ij} - r_{ik}\mu_{kj} + a - 1 - \mu_{kj}a + \mu_{kj} - \mu_{kj} - \mu_{kj}b + \mu_{kj} \\
&= \frac{1}{\mu_{kj}(1 - \mu_{kj})} \left(\sum_i r_{ik}x_{ij} - \left(\sum_i r_{ik} + a + b - 2\right)\mu_{kj} + a - 1\right) = 0 \\
\implies \sum_i r_{ik}x_{ij} + a - 1 &= \left(\sum_k r_{ik} + a + b - 2\right)\mu_{kj}
\end{aligned}$$

That gives the desired result. ∎

**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 $\ell_1$ 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 $\lambda$. 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 $\lambda$ 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).

$$\nabla \|\mathbf{x}\|_1 = \begin{Bmatrix} \partial \|\mathbf{x}\|_1 / \partial \mathbf{x}_1 \\ \vdots \end{Bmatrix} = \begin{Bmatrix} sign(\mathbf{x}_1) \\ \vdots \end{Bmatrix} = sign(\mathbf{x})$$

So now I can find the gradient of my objective. Recall that $\nabla \|A\mathbf{x} - \mathbf{b}\|_2^2 = 2A^T(A\mathbf{x} - \mathbf{b})$.

$$2A^T(A\mathbf{x} - \mathbf{b}) + \lambda sign(\mathbf{x})$$

I used the solution for the code part. The most important features were timedelta, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_saturday. ■