CMSC 25025 / STAT 37601

HW3 - DUE: April 25, 2019

Student: Dimitrios Tanoglidis

Problem 2: Lasso

First, note that $[|t|-\lambda]_+ = max\{0,[|t|-\lambda]\}$.
Take cases:
• Case 1: $\beta > 0$.

$$f(\beta) = -t\beta + \frac{1}{2}\beta^2 + \lambda\beta \Rightarrow f'(\beta) = -t + \beta + \lambda = 0 \Longrightarrow \boxed{\beta = t - \lambda} \tag{1}$$

only when $t - \lambda > 0$. (In order our assumption, $\beta > 0$ to be true).

• Case 2: $\beta < 0$.

$$f(\beta) = -t\beta + \frac{1}{2}\beta^2 - \lambda\beta \Rightarrow f'(\beta) = -t + \beta - \lambda = 0 \Longrightarrow \boxed{\beta = t + \lambda} \tag{2}$$

only when $t + \lambda > 0$. (In order our assumption, $\beta < 0$ to be true).

So, the above solutions work for $t > \lambda$ or $t < -\lambda$, or - in compact form - $|t| - \lambda < 0$. Note also that $t + \lambda = -(|t| - \lambda)$ for $t < 0$.

So, we can write all the above in the compact form:

$$\boxed{\hat{\beta} = [|t| - \lambda]_+} \tag{3}$$

Problem 3: Logistic regression

(a) I will follow the notation of two classes $Y \in \{0, 1\}$.

In the logistic regression model, the probability to get class $Y_i = 1$ is :

$$P(Y_i = 1|X_i) = \pi_1(X_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \quad \eta_i = X_i^T\theta \tag{4}$$

The probability to get class $Y_i = 0$ is then:

$$P(Y_i = 0|X_i) = 1 - \pi(X_i). \tag{5}$$

Together, I can write the probability to get class Y_i in the compact form:

$$P(Y_i|X_i) = \pi_1(X_i)^{Y_i}\,(1 - \pi_1(X_i)))^{1-Y_i} \tag{6}$$

Furthermore, for convenience, I set $\pi_1(X_i) \equiv p_i$. Then, I can write the likelihood, L as (for n observation):

$$L(\theta) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i} \tag{7}$$

And the negative log-likelihood, is:

$$\boxed{\ell(\theta) = -\log L(\theta) = \sum_{i=1}^{n}[(Y_i - 1)\log(1 - p_i) - Y_i \log p_i]} \tag{8}$$

Now, let's calculate the derivative $\frac{\partial p_i}{\partial \theta_j}$, it will be useful in a while. We have:

$$
\begin{aligned}
\frac{\partial p_i}{\partial \theta_j} &= \frac{dp_i}{d\eta_i}\frac{\partial \eta_i}{\partial \theta_j} \\
&= \frac{d}{d\eta_i}\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)\frac{\partial (X_i^T\theta)}{\partial \theta_j} \qquad (9)
\end{aligned}
$$

It is easy to show that:

$$
\frac{d}{d\eta_i}\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right) = \frac{e^{\eta_i}}{1+e^{\eta_i}}\left(1-\frac{e^{\eta_i}}{1+e^{\eta_i}}\right) = p_i(1-p_i), \qquad (10)
$$

and:

$$
\frac{\partial (X_i^T\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_k}\sum_k X_{ik}\theta_k = \sum_k X_{ik}\delta_{jk} = X_{ij} \qquad (11)
$$

So, finally:

$$
\boxed{\frac{\partial p_i}{\partial \theta_j} = p_i(1-p_i)X_{ij}} \qquad (12)
$$

Now, newton iteration for current θ is given by:

$$
\boxed{\theta^{new} = \theta - H^{-1}(\theta)\nabla\ell(\theta),} \qquad (13)
$$

where H the Hessian (matrix of second derivatives).

Taking the fist derivative of the negative log-likelihood, we have:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta_j} &= \sum_{i=1}^n \left[(Y_i-1)\frac{1}{1-p_i}\left(-\frac{\partial p_i}{\partial \theta_j}\right) - y_i\frac{1}{p_i}\frac{\partial p_i}{\partial \theta_j}\right] \\
&= \sum_{i=1}^n \left[(1-Y_i)\frac{1}{1-p_i}p_i(1-p_i) - Y_i\frac{1}{p_i}p_i(1-p_i)\right] X_{ij} \\
&= \sum_{i=1}^n \left[(1-Y_i)p_i - Y_i(1-p_i)\right] X_{ij} \\
&= \sum_{i=1}^n X_{ij}\left[p_i - Y_i\right] \qquad (14)
\end{aligned}
$$

Or, in vector notation:

$$
\boxed{\nabla\ell(\theta) = X^T(P-Y),} \qquad (15)
$$

where P and V vectors of p_i and Y_i.

Now, easily, we can get the second derivative of the (negative) log-likelihood:

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} &= \sum_{i=1}^n X_{ij}\frac{\partial p_i}{\partial \theta_k} \\
&= \sum_{i=1}^n X_{ij}X_{ik}p_i(1-p_i) \\
&= [X^TWX]_{jk} \\
&= H_{jk} \qquad (16)
\end{aligned}
$$

Where in the third line, I defined:

$$
W = \mathrm{diag}[p_i(1-p_i)] \qquad (17)
$$

So, summarizing, we have:

$$\begin{aligned}\theta^{new} &= \theta - (X^TWX)^{-1}X^T(P-Y) \\ &= \theta + (X^TWX)^{-1}X^T(Y-P) \\ &= \theta + (X^TWX)^{-1}X^TW\tilde{Z}\end{aligned} \tag{18}$$

where I defined $\tilde{Z} = W^{-1}(Y-P)$, with elements:

$$\tilde{z}_i = \frac{Y_i - p_i}{p_i(1-p_i)} \tag{19}$$

Now, notice that:

$$(X^TWX)^{-1}X^TW\eta = (X^TWX)^{-1}X^TW(X\theta) = (X^TWX)^{-1}(X^TWX)\theta = \theta \tag{20}$$

So, the θ can be absorbed if we define:

$$z_i = \eta_i + \tilde{z} = \eta_i + \frac{Y_i - p_i}{p_i(1-p_i)}$$

.

So, finally we have:

$$\boxed{\theta^{new} = (X^TWX)^{-1}X^TWZ} \tag{21}$$

But this is exactly the solution of weighted least squares, so we can equivalently write:

$$\boxed{\theta^{new} = \arg\min_{\theta} \left(\mathbf{Z} - \mathbf{X}\theta\right)^T W \left(\mathbf{Z} - \mathbf{X}\theta\right)} \tag{22}$$

QED!!

(b) Let's rewrite the likelihood, by separating explicitly the two classes.

$$L = \prod_i I[Y_i = 1]p_i \prod_j I[Y_j = 0](1 - p_j) \tag{23}$$

Let's write the p_i as:

$$p_i = \frac{1}{e^{-\alpha x^T\theta} + 1} \tag{24}$$

and see if there is a value of α that maximizes L.

If $Y_i = 1$ then $x^T\theta > 0$. Then, the terms p_i associated with this $Y_i = 1$ increase with increasing values of α:

$$\lim_{\alpha\to+\infty} p_i = \lim_{\alpha\to+\infty} \frac{1}{e^{-\alpha x^T\theta} + 1} = 1 \tag{25}$$

(increasing with increasing α).

If $Y_j = 0$ then $x^T\theta < 0$. Then, the terms $1 - p_j$ associated with this $Y_j = 0$ increase with increasing values of α:

$$\lim_{\alpha\to+\infty} (1 - p_j) = \lim_{\alpha\to+\infty} \left(1 - \frac{1}{e^{-\alpha x^T\theta} + 1}\right) = 1 \tag{26}$$

So, the likelihood L keeps increasing with higher and higher values of α; there is no finite value of α that can maximize it.

Problem 4: Bernoulli mixtures

(a) Let me introduce a notation where each vector $X_i = [x_{i1}, \ldots, x_{id}]$, $i = 1, \ldots, n$. In other words any vector of the sample has d components. The first index denotes the number of the vector/sample while the second index denotes the element of the vector in the R^d space.

The total likelihood is:

$$L(p) = \prod_{i=1}^{n} P(X_i) = \prod_{i=1}^{n} \prod_{j=1}^{d} p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})} \tag{27}$$

The log-likelihood is then:

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{d} [x_{ij} \log p_j + (1 - x_{ij}) \log(1 - p_j)] \tag{28}$$

Cost equations:

$$\frac{\partial \log L}{\partial p_k} = 0 \Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{d} \left[x_{ij} \frac{1}{p_j} \delta_{jk} + (1 - x_{ij}) \frac{1}{1 - p_j} (-1) \delta_{jk} \right] = 0 \tag{29}$$

Which becomes:

$$\begin{aligned} \frac{\partial \log L}{\partial p_k} &= \frac{1}{p_k} \sum_{i=1}^{n} x_{ik} - \frac{1}{1 - p_k} \sum_{i=1} (1 - x_{ik}) = 0 \Rightarrow \\ &\Rightarrow \frac{1}{p_k} \sum_{i=1}^{n} x_{ik} - \frac{1}{1 - p_k} \left(n - \sum_{i=1}^{n} x_{ik} \right) = 0 \Rightarrow \end{aligned}$$

And solving for p_k (rename to p_j), we get the ML estimator:

$$\boxed{\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}} \tag{30}$$

(b) i. For n samples X_i, the likelihood is:

$$L = \prod_{i=1}^{n} \sum_{m=1}^{M} \pi_m f_m(X_i; \theta_m) \tag{31}$$

And the log-likelihood:

$$\ell = \log L = \sum_{i=1}^{n} \log \sum_{m=1}^{M} \pi_m f_m(X_i; \theta_m) \tag{32}$$

Let's take the derivative with respect to π_k:

$$\frac{\partial \ell}{\partial \pi_k} = \sum_{i=1}^{n} \frac{\pi_k f_k(X_i; \theta_k)}{\sum_{m=1}^{M} \pi_m f_m(X_i; \theta_m)}. \tag{33}$$

Adding a lagrange multiplier:

$$\frac{1}{\pi_k} \sum_{i=1}^{n} \frac{\pi_k f_k(X_i; \theta_k)}{\sum_{m=1}^{M} \pi_m f_m(X_i; \theta_m)} - \lambda = 0. \tag{34}$$

Define now the responsibilities:

$$\boxed{w_{ki} \equiv \frac{\hat{\pi}_k f_k(X_i; \theta_k)}{\sum_{m=1}^{M} \hat{\pi}_m f_m(X_i; \theta_m)}}. \tag{35}$$

where $\hat{\pi}_k, \theta_m$ current estimates.
ii. With the above definition, we get:

$$\frac{1}{\pi_k}\sum_{i=1}^{n} w_{ki} - \lambda = 0 \tag{36}$$

Then (with $\lambda = n$), we have the new estimates for $\hat{\pi}$:

$$\boxed{\hat{\pi}_k^{new} = \frac{1}{n}\sum_{i=1}^{n} w_{ki}} \tag{37}$$

Let's now calculate the new estimates for θ_m (the $p_{j,m}$s):

$$\frac{\partial \ell}{\partial p_{k,\eta}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial f_k(X_i;\theta_k)}{\partial p_{k,\eta}}}{\sum_{m=1}^{M} \hat{\pi}_m f_m(X_i;\theta_m)} = \sum_{i=1}^{n} w_{ki} \frac{\partial \log f_k(X_i;\theta_k)}{\partial p_{k,\eta}} = 0 \tag{38}$$

Let's calculate:

$$\begin{aligned}
\frac{\partial \log f_k(X_i;\theta_k)}{\partial p_{k,\eta}} &= \frac{\partial}{\partial p_{k,\eta}} \log \prod_{j=1}^{d} p_{j,k}^{x_{ij}} (1-p_{j,k})^{(1-x_{ij})} \\
&= \frac{\partial}{\partial p_{k,\eta}} \sum_{j=1}^{n} x_{ij} \log p_{j,k} + (1-x_{ij}) \log(1-p_{j,k}) \\
&= \frac{1}{p_{\eta,k}} x_{i\eta} - \frac{1}{1-p_{\eta,k}}(1-x_{i\eta})
\end{aligned} \tag{39}$$

Thus:

$$\frac{\partial \ell}{\partial p_{k,\eta}} = \frac{1}{p_{\eta,k}} \sum_{i=1}^{n} w_{ki} x_{i\eta} - \frac{1}{1-p_{\eta,k}} \left(\sum_{i=1}^{n} w_{ik} - \sum_{i=1}^{n} w_{ki} x_{i\eta} \right) = 0 \tag{40}$$

And, rearranging we finally get:

$$\boxed{\hat{p}_{\eta,k} = \frac{\sum_{i=1}^{n} w_{ik} x_{i\eta}}{\sum_{i=1}^{n} w_{ik}}} \tag{41}$$

which looks like a weighted mean!