HW 4 Solutions

Please hand in this Homework as follows:
• Upload to Gradescope `HW4`: A pdf of the theoretical homework **combined with** the pdf of your jupyter notebook for problem 3.
• Upload to Gradescope `HW4 programming`: A `ipynb` file for problem 3.
You must upload both the pdf of your jupyter notebook to `HW4` and the code `ipynb` file to `HW4 programming`.

1. *Logistic regression* (15 points)

   (a) Recall the in class we wrote the logistic regression loss as

$$L(\theta) = \sum_{i=1}^{n} \log(1 + e^{-Y_i X_i^t \theta}),\ Y_i \in \{-1, 1\}$$

(I've removed the factor 2 in the exponent as it can be absorbed in θ.)
Give a detailed derivation of the Newton algorithm for logistic regression and show that each iteration corresponds to solving a weighted least square problem, i.e. starting with θ^{old}, θ^{new} solves:

$$\arg\min_{\theta} (Z - X\theta)^t W (Z - X\theta), \text{or } X^t W X \theta = X^t W Z,$$

where $X \in R^{n \times d}$ is the training data matrix, $W \in R^{n \times n}$ is a diagonal matrix and $Z \in R^n$. Write W and Z in terms of $X, Y, \theta^{(old)}$. Hint: Note that in this model $p_i = P(Y_i = 1|X_i) = \frac{1}{1+e^{-\eta_i}} = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, where $\eta_i = X_i^t \theta$, and $\frac{e^{Y_i \eta_i}}{[1+e^{Y_i \eta_i}]^2} = p_i(1 - p_i)$.
We write the model for logistic regressionas $y_i \sim Bern(p_i)$, where $p_i = \frac{1}{1+e^{-X_i^t \theta}}$, the negative log-likelihood function is:

$$L(\theta|X, y) = \sum_{i=1}^{n} \log\left(1 + e^{-Y_i X_i^t \theta}\right)$$

So to minimize we take the gradient:

$$\nabla L = -\sum_{i=1}^{n} \left[Y_i X_i \frac{e^{-Y_i X_i^T \theta}}{1 + e^{-Y_i X_i^T \theta}} \right] = -\sum_{i=1}^{n} \left[Y_i X_i \frac{1}{1 + e^{Y_i X_i^T \theta}} \right].$$

$$H = \nabla^2 L = \sum_{i=1}^{n} \frac{e^{Y_i X_i^T \theta}}{(1 + e^{Y_i X_i^T \theta})^2} Y_i^2 X_i X_i^T = \sum_{i=1}^{n} p_i(1 - p_i) X_i X_i^t$$

The Newton iteration is $-H(\theta^{new} - \theta) = \nabla L$, or $-H\theta^{new} = -H\theta + \nabla L$. Using the fact that $X_i^t \theta = \eta_i$, and

$$Y_i/(1+e^{Y_iX_i^t\theta}) = \begin{cases} 1-p_i & Y_i = 1 \\ -p_i & Y_i = -1 \end{cases} = (Y_i+1)/2 - p_i.$$

$$\begin{aligned} \sum_{i=1}^n X_i X_i^t p_i(1-p_i)\theta^{new} &= \sum_{i=1}^n X_i p_i(1-p_i)\eta_i + \sum_{i=1}^n Y_i X_i \frac{1}{1+e^{Y_iX_i^T\theta}} \\ &= \sum_{i=1}^n X_i p_i(1-p_i)\left[\eta_i + \frac{1}{p_i(1-p_i)} Y_i \frac{1}{1+e^{Y_iX_i^T\theta}}\right] \end{aligned}$$

$$W_{ii} = p_i(1-p_i) \qquad Z_i = \eta_i + \frac{(Y_i+1)/2 - p_i}{p_i(1-p_i)},$$

Thus, each iteration is equivalent to solving a weighted LS problem.

(b) Assume the data are perfectly linearly separable, i.e. there exist θ such that $x_i^t\theta < 0$ if $y_i = 0$ and $x_i^t\theta > 0$ if $y_i = 1$. Show that the maximum likelihood estimator for the logistic regression model does not exist. Comment on the behavior of the iteratively reweighted least squares algorithm. Hint: If θ is a perfect separator then $\alpha\theta$ is also for any $\alpha > 0$. It may be easier to work with the likelihood instead of the log-likelihood.

Consider the likelihood function:

$$f(\theta) = \prod_{y_i=1} \frac{e^{x_i^T\theta}}{e^{x_i^T\theta}+1} \prod_{y_i=0} \frac{1}{e^{x_i^T\theta}+1}.$$

If θ can separate data well, then consider $\alpha\theta$

$$f(\alpha\theta) = \prod_{y_i=1} \frac{e^{\alpha x_i^T\theta}}{e^{\alpha x_i^T\theta}+1} \prod_{y_i=0} \frac{1}{1+e^{\alpha x_i^T\theta}}.$$

This function is continuous increasing to 1 as $\alpha \to \infty$. However, $f(\theta) < 1$ for all θ. Thus the MLE does not exist in this case.

For solving the iteratively reweighted least squares algorithm, as the loss function is strictly convex, the function value will increase to 1 after iterations. When $f(\theta) \geq 1-\epsilon$, $x_i^T\theta > \log\frac{1-\epsilon}{\epsilon}$ for $y_i = 1$ and $x_i^T\theta < -\log\frac{1-\epsilon}{\epsilon}$ for $y_i = 0$. Then $w_i(1-w_i) \to 0$ for all i and $\|H\| \to 0$. The Newton iteration that $\theta^{new} - \theta = H^{-1}L$ is diverging.

(c) What happens with the hinge and the quadratic losses in the perfectly separable setting. In both cases discuss whether there is a minimizer, and explain your conclusions.

Hinge: $L(\theta) = \sum_{i=1}^n [1 - Y_iX_i^t\theta]_+$.

Quadratic: $L(\theta) = \sum_{i=1}^n [1 - Y_iX_i^t\theta]^2$.

For hinge loss, if $Y_iX_i^t\theta > 0$ for $i = 1, ..., n$, then any large enough α can make $L(\alpha\theta) = \sum_{i=1}^n [1 - \alpha Y_iX_i^t\theta]_+ = 0$, so there is a minimizer but it's not unique.

For quadratic loss, $\nabla^2 L(\theta) = 2\sum_{i=1}^n X_iX_i^t$ is positive semi-definite, so $L(\theta)$ is convex and there is a minimizer.

2. *Lasso minimization* (15 points) We are given data $X_1, Y_1, \ldots, X_n, Y_n$, with $X_i \in R^d, Y_i \in R$. We assume each coordinate of the X_i's has mean 0 and variance 1. In class we discussed coordinatewise minimization of the Lasso loss function:

$$L(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^t\theta)^2 + \lambda\sum_{j=1}^{d}|\theta_j|,$$

with $\lambda > 0$.

   (a) Fixing all coordinates except for the k'th coordinate we minimize:
   $f(\theta_k) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - c_i - X_{ik}\theta_k|)^2 + C + \lambda|\theta_k|$.
   Write out the expressions for c_i and C.
   $c_i = \sum_{j=1,j\neq k}^{d} X_{ij}\theta_j$, $C = \sum_{j=1,j\neq k}^{d}\lambda|\theta_j|$

   (b) Show that minimizing $f(\theta_k)$ is equivalent to minimizing
   $g(\theta_k) = \frac{1}{2}\theta_k^2 - \frac{1}{n}\sum_{i=1}(Y_i - c_i)X_{ik}\theta_k + \lambda|\theta_k|$.
   $f(\theta_k) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - c_i - X_{ik}\theta_k|)^2 + C + \lambda|\theta_k| = \frac{1}{2n}(\sum_{i=1}^{n} X_{ik}^2)\theta_k^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - c_i)X_{ik}\theta_k + \frac{1}{2n}(\sum_{i=1}^{n}(Y_i - c_i)^2) + C + \lambda|\theta_k|$.
   Because each coordinate of the X_i's has mean 0 and variance 1, $\sum_{i=1}^{n} X_{ik}^2 = n$, so we can minimize $g(\theta_k) = \frac{1}{2}\theta_k^2 - \frac{1}{n}\sum_{i=1}(Y_i - c_i)X_{ik}\theta_k + \lambda|\theta_k|$.

   (c) Define the function $h(x) = \frac{1}{2}x^2 - tx + \lambda|x|,\ \lambda > 0$, show that it is strictly convex, and thus has a unique minimum.
   $\frac{1}{2}x^2 - tx = \frac{1}{2}(x-t)^2 - \frac{1}{2}t^2$ is strictly convex, $\lambda|x|$ is convex, so $h(x)$ is strictly convex, and thus has a unique minimum.

   (d) Show that the minimum is given by $x^* = sign(t)[|t| - \lambda]_+$. (Hint: If a strictly convex function h is smoothly differentiable at a point x and $h'(x) = 0$ then it is minimized at x.)
   If $|t| - \lambda > 0$, $h^{'}(x^*) = (x^* - t) + \lambda sign(t) = sign(t)|t| - sign(t)\lambda - t + \lambda sign(t) = 0$, so $h(x)$ is minimized when $x = x^*$. If $|t| - \lambda <= 0$, for any x between 0 and t, $|(\frac{1}{2}(x-t)^2)'| = |x - t| <= |t| <= \lambda$, so $h(x)$ is minimized when $x = 0 = sign(t)[|t| - \lambda]_+ = x^*$.

3. *Multinomial gradient* (10 points) You have C classes and labeled data $X_1, Y_1, \ldots, X_n, Y_n$, with $X_i \in R^d$ and $Y_i \in \{1, \ldots, C\}$. Let Z_i be the 'one-hot' vector corresponding to Y_i, i.e. $Z_{ij} = 1_{j=Y_i}, j = 1, \ldots, C$. Let $\mathbf{X}$ be the $n \times d$ data matrix. Let $\mathbf{Z}$ be the $n \times C$ label matrix.

We model

$$P(Y = c|X = x) = \frac{\exp\theta_c^t x}{\sum_{k=1}^{C}\exp\theta^k x},$$

for $\theta_c, c = 1, \ldots, C$ unknown parameters in R^d.

In class we wrote the likelihood $\theta = (\theta_1, \ldots, \theta_C)$ as

$$L(X, Y, \theta) = \prod_{i=1}^{n} \prod_{c=1}^{C} \left[\frac{\exp \theta_c^t X}{\sum_{k=1}^{C} \exp \theta^k X} \right]^{Z_{ic}}.$$

Denote by $\pi_{ic} = P(Y = c | X_i, \theta)$, let $\pi_c = (\pi_{1c}, \ldots, \pi_{nc})$ and let π be the $n \times C$ matrix with columns $\pi_c, c = 1 \ldots, C$.

(a) Write the log-likelihood $\log L(X, Y, \theta)$.

$$\begin{aligned} \log L(X, Y, \theta) &= \sum_{i=1}^{n} \sum_{c=1}^{C} \left[Z_{ic} \left(\theta_c^T X_i - \log \left(\sum_{k=1}^{C} \exp \theta_k^T X_i \right) \right) \right] \\ &= \sum_{i=1}^{n} \sum_{c=1}^{C} Z_{ic} \theta_c^T X_i - \sum_{i=1}^{n} \sum_{c=1}^{C} Z_{ic} \log \left(\sum_{k=1}^{C} \exp \theta_k^T X_i \right) \\ &= \sum_{i=1}^{n} \sum_{c=1}^{C} Z_{ic} \theta_c^T X_i - \sum_{i=1}^{n} \log \left(\sum_{k=1}^{C} \exp \theta_k^T X_i \right). \end{aligned}$$

using the fact that $\sum_{c=1}^{C} Z_{ic} = 1$.

(b) Write the gradient of $\nabla_{\theta_c} \log L(\theta)$ w.r.t to θ_c in terms of $\mathbf{Z}, \pi_c$ and $\mathbf{X}$.

$$\begin{aligned} \nabla_{\theta_c} \log L(\theta) &= \sum_{i=1}^{n} Z_{ic} X_i - \sum_{i=1}^{n} \frac{\exp \theta_c^T X_i}{\sum_{k=1}^{C} \exp \theta_k^T X_i} X_i \\ &= \sum_{i=1}^{n} Z_{ic} X_i - \sum_{i=1}^{n} \pi_{ic} X_i \\ &= X^T (Z_c - \pi_c) \end{aligned}$$

where Z_c is the c-th column of $Z, c = 1, \ldots, C$.

(c) Write the $C \times d$ matrix of the C gradients $\nabla_{\theta_c} \log L, c = 1, \ldots, C$ as a matrix product in terms of $\mathbf{Z}, \pi$ and X, yielding a $d \times C$ matrix.

$$\nabla_\theta \log L(\theta) = X^T (Z - \pi).$$