CSE 251A Winter 2021 - Take Home Exam 2

1 hour **Name:** **PID:**

---

(1) This is an open book, open notes exam. You are free to consult any text book or notes. **You are not allowed to consult with any other person.**
(2) If you need any clarification, please post a private message to the instructors on Piazza.
(3) Remember that your work is graded on the *clarity* of your writing and explanation as well as the validity of what you write.
(4) This is a one-hour exam.

(1) Suppose we are given a training set $S = \{(x^{(i)}, y^{(i)}), i = 1 \dots n\}$ of labeled vectors where $x^{(i)} \in \mathbb{R}^d$ are d-dimensional feature vectors and $y^{(i)} \in \{-1, 1\}$ are labels. Let (w_{LR}, b_{LR}) be the solution to L_2-regularized logistic regression with regularization parameter λ on S.

Now suppose that we scale each feature vector by a positive constant $c > 1$ to get the training dataset S' – that is, $S' = \{(cx^{(i)}, y^{(i)}), i = 1, \dots, n\}$. Let (w'_{LR}, b'_{LR}) be the solution to L_2-regularized logistic regression with regularization parameter λ' on S'.

State whether the following statements are true or false. Justify your answer with either a proof or a counterexample.

(a) (5 points) Suppose that $\lambda' = \lambda = 0$, and that (w_{LR}, b_{LR}) and (w'_{LR}, b'_{LR}) are unique. Then $w'_{LR} \neq w_{LR}$ in general (for any arbitrary S).

Solution: True.

Proof. Let $L(w, b) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left(w^\top x^{(i)} + b\right)}\right)$ be the loss function of logistic regression on S. Now, on the training set $S' = \{(cx^{(i)}, y^{(i)}), i = 1, \dots, n\}$, let

$$\tilde{L}(w, b) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left(w^\top (cx^{(i)}) + b\right)}\right) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left((cw)^\top x^{(i)} + b\right)}\right)$$

be the loss function. Setting $\tilde{w} = cw$, we can rewrite the loss function as

$$\tilde{L}(w, b) = \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\left(\tilde{w}^\top x^{(i)} + b\right)}\right) = L(\tilde{w}, b) = L(cw, b).$$

We know that $(w_{LR}, b_{LR}) = \mathbf{argmin}_{(w,b)} L(w, b)$. Therefore, we have

$$(w'_{LR}, b'_{LR}) = \mathbf{argmin}_{(w,b)} \tilde{L}(w, b) = \mathbf{argmin}_{(w,b)} L(cw, b) = (\frac{w_{LR}}{c}, b_{LR}).$$

Hence, $w'_{LR} = \frac{w_{LR}}{c}$ and $w'_{LR} \neq w_{LR}$ unless $c = 1$. □

(b) (5 points) If $\lambda' = c^2\lambda > 0$, then $w'_{LR} = w_{LR}/c$ and $b'_{LR} = b_{LR}$.

Solution: True.

Proof. Let $G(w,b) = \sum_{i=1}^n \ln\left(1 + e^{-y^{(i)}\left(w^\top x^{(i)} + b\right)}\right) + \lambda\|w\|_2^2$ and $\tilde{G}(w,b) = \sum_{i=1}^n \ln\left(1 + e^{-y^{(i)}\left(w^\top (cx^{(i)}) + b\right)}\right) + \lambda'\|w\|_2^2$ be the loss functions.
Setting $\tilde{w} = cw$, we can rewrite $\tilde{G}(w,b)$ as

$$\begin{aligned}\tilde{G}(w,b) &= \sum_{i=1}^n \ln\left(1 + e^{-y^{(i)}\left((cw)^\top x^{(i)} + b\right)}\right) + c^2\lambda\|w\|_2^2 \\ &= \sum_{i=1}^n \ln\left(1 + e^{-y^{(i)}\left(\tilde{w}^\top x^{(i)} + b\right)}\right) + \lambda\|\tilde{w}\|_2^2 \\ &= G(\tilde{w}, b) = G(cw, b).\end{aligned}$$

Since $(w_{LR}, b_{LR}) = \mathbf{argmin}_{(w,b)} G(w,b)$, we have

$$(w'_{LR}, b'_{LR}) = \mathbf{argmin}_{(w,b)} \tilde{G}(w,b) = \mathbf{argmin}_{(w,b)} G(cw,b) = (\frac{w_{LR}}{c}, b_{LR}).$$

□

(2) Suppose we are given an input dataset $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ where each $x^{(i)} \in \mathbb{R}^d$. Let v be a parameter vector in $\mathbb{R}^d$. Consider the following loss function:

$$G(v) = \sum_{i=1}^n \frac{v^\top x^{(i)} (x^{(i)})^\top v}{v^\top v}$$

(a) (5 points) Write down the update step for stochastic gradient descent corresponding to a single data point $x^{(i)}$.

Solution: We first present two useful identities: Let y be a $d \times 1$ vector and A be a $d \times d$ symmetric, constant matrix. It is easy to show that $\nabla(y^\top A y) = 2Ay$ and $\nabla(y^\top y) = 2y$.
Now, Let

$$g(v, x^{(i)}) = \frac{v^\top x^{(i)} (x^{(i)})^\top v}{v^\top v}.$$

We have

$$\begin{aligned}(1) \qquad \nabla g(v, x^{(i)}) &= \left(\frac{d}{dv}\left(v^\top x^{(i)} (x^{(i)})^\top v\right)\right)\frac{1}{v^\top v} + (v^\top x^{(i)} (x^{(i)})^\top v)\left(\frac{d}{dv}\left(\frac{1}{v^\top v}\right)\right) \\ (2) \qquad &= \left(\frac{d}{dv}\left(v^\top x^{(i)} (x^{(i)})^\top v\right)\right)\frac{1}{v^\top v} + (v^\top x^{(i)} (x^{(i)})^\top v)\left(\frac{-1}{(v^\top v)^2} \cdot \frac{d(v^\top v)}{dv}\right) \\ (3) \qquad &= \frac{2x^{(i)}(x^{(i)})^\top v}{v^\top v} + (v^\top x^{(i)} (x^{(i)})^\top v)\left(\frac{-1}{(v^\top v)^2} \cdot (2v)\right) \\ &= \frac{2x^{(i)}(x^{(i)})^\top v}{v^\top v} - \frac{2(v^\top x^{(i)} (x^{(i)})^\top v)v}{(v^\top v)^2},\end{aligned}$$

where (1) follows from the product rule, (2) follows from the chain rule, and (3) uses the aforementioned identities.
Then, the update step for stochastic gradient descent corresponding to $x^{(i)}$ is:

$$\begin{aligned}v_{t+1} &= v_t - \eta \nabla g(v_t, x^{(i)}) \\ &= v_t - \eta\left(\frac{2x^{(i)}(x^{(i)})^\top v_t}{v_t^\top v_t} - \frac{2(v_t^\top x^{(i)} (x^{(i)})^\top v_t)v_t}{(v_t^\top v_t)^2}\right)\end{aligned}$$

(b) (5 points) Write down an algorithm that implements the stochastic gradient update in the first part in $O(d)$ time.

Solution: Let's consider the following algorithm.

(i) First, compute and store $a = v_t^\top x^{(i)} = (x^{(i)})^\top v_t$ and $b = v_t^\top v_t$; this can be done in $O(d)$ time. a and b are both scalars.

(ii) The stochastic gradient update can then be rewritten as

$$\begin{aligned} v_{t+1} &= v_t - \eta \nabla g(v_t, x^{(i)}) \\ &= v_t - \eta \left(\frac{2x^{(i)} \left((x^{(i)})^\top v_t \right)}{v_t^\top v_t} - \frac{2 \left(v_t^\top x^{(i)} \right) \left((x^{(i)})^\top v_t \right) v_t}{(v_t^\top v_t)^2} \right) \\ &= v_t - \frac{2\eta a}{b} x^{(i)} + \frac{2\eta a^2}{b^2} v_t, \end{aligned} \tag{4}$$

and computing (4) takes $O(d)$ time.

Overall, this algorithm takes $O(d) + O(d) = O(d)$ time in total.