CSE 250B Winter 2020 - Take Home Quiz 1

Due: Mar 13, 2020 - No Time Limit **Name:** **PID:**

---

(1) This is an open book, take home quiz. **No collaboration is allowed.**

(2) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

(1) Sometimes machine learning is used on imperfect training data – for example, data collected via noisy sensors. In these cases, we might try to correct for noise while training the classifier.

Consider the following formulation for training a logistic regression classifier $w \in \mathbb{R}^d$ on a noisy training data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $y^{(i)} \in \{-1, +1\}$. For simplicity, we ignore the bias term b. Suppose we know that the noise magnitude is at most r. Then, instead of the standard logistic regression loss, we might want to minimize the following loss:

$$\tilde{L}(w) = \sum_{i=1}^{n} \max_{z^{(i)}:\|z^{(i)}-x^{(i)}\| \leq r} \log(1 + \exp(-y^{(i)} w^\top z^{(i)})),$$

where $\|v\|$ means the L_2-norm of vector v.

(a) (5 points) Prove that $\tilde{L}(w) = M(w)$, where $M(w) = \sum_{i=1}^{n} \log(1 + \exp(r\|w\| - y^{(i)} w^\top x^{(i)}))$. For full credit, show all the steps in your proof.

Let $z^{(i)} = x^{(i)} - \epsilon$ where ϵ is a vector of magnitude r. Then, for a specific i, the loss function reduces to:

$$\max_{\epsilon:\|\epsilon\| \leq r} \log(1 + \exp(-y^{(i)} w^\top x^{(i)} + y^{(i)} w^\top \epsilon)),$$

Since $\exp(-y^{(i)} w^\top x^{(i)}) > 0$ and log is an increasing function, for a specific $(x^{(i)}, y^{(i)})$, the maximum happens when $y^{(i)} w^\top \epsilon > 0$ and when $y^{(i)} w^\top \epsilon$ is the highest. Since ϵ can be in any direction, by the Cauchy Schwartz Inequality, this highest value occurs when ϵ is the vector of length r along $y^{(i)} w$, and the highest value is $r(y^{(i)})^2 w^\top \hat{w}$, where $\hat{w}$ is the unit vector along w, which is equal to $r\|w\|$. Plugging this in to the loss function completes the proof.

(b) (5 points) Write down the gradient descent update for minimizing $M(w)$.

Observe that for a specific $(x^{(i)}, y^{(i)})$, the gradient is:

$$\frac{1}{1 + \exp(-r\|w\| + y^{(i)} w^\top x^{(i)})} \cdot \left(-y^{(i)} x^{(i)} + r \frac{w}{\|w\|} \right)$$

Thus, the gradient descent update is:

$$w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} \frac{1}{1 + \exp(-r\|w_t\| + y^{(i)} w_t^\top x^{(i)})} \cdot \left(y^{(i)} x^{(i)} - r \frac{w_t}{\|w_t\|} \right),$$

where η_t is the learning rate at time t.

(c) (5 points) Suppose you already have code for a single stochastic gradient update for minimizing the logistic regression loss function. Specifically, you have code for a function `logistic-SGD`(w, x, y, η) that given the current w, a labeled example (x, y) and a learning rate η, returns you the updated weight vector:

$$w_{t+1} = w_t - \eta \nabla \log(1 + e^{-y w_t^\top x})$$

Show how you can use this function `logistic-SGD` to code up a stochastic gradient update on $\tilde{L}(w) = M(w)$. Specifically, given a w, a labeled example (x, y) and a learning rate η, your function should return the updated weight vector:

$$w_{t+1} = w_t - \eta \nabla M(w_t)$$

(Hint: You may need to give `logistic-SGD` inputs that are different from w, (x, y) and η.)

We use $\hat{v}$ to denote an unit vector along v. Observe from part (b) that the gradient of $M(w)$ for a specific $(x^{(i)}, y^{(i)})$ is:

$$\frac{1}{1 + \exp(-r\|w\| + y^{(i)} w^\top x^{(i)})} \cdot \left(-y^{(i)} x^{(i)} + r\frac{w}{\|w\|}\right)$$

Since $-r\|w\| + y^{(i)} w^\top x^{(i)} = w^\top (y^{(i)} x^{(i)} - r\hat{w})$, using `logistic-SGD`$(w_t, z^{(i)} = y^{(i)} x^{(i)} - r\hat{w}_t, 1, \eta_t)$ results in the update:

$$\begin{aligned} w_{t+1} &= w_t + \eta_t \cdot \frac{1 \cdot z^{(i)}}{1 + e^{1 \cdot w_t^\top z^{(i)}}} \\ &= w_t + \eta_t \cdot \frac{y^{(i)} x^{(i)} - r\hat{w}_t}{1 + \exp(y^{(i)} w_t^\top x^{(i)} - r\|w_t\|)} \end{aligned}$$

which is exactly the SGD update for $M(w)$.