Joshua Hill

Professor Vikalo

ECE 381V: Statistical Machine Learning

16 February 2024

Homework #1

1)

a. If w* is an optimal vector of weights, it implies that the PLA will **not** make any incorrect classifications. Hence, (yn w∗Txn) will always be positive. Therefore, the minimum value (p) will also always be positive.

b.

Base Case ($t = 1$):

- **Claim:** The inequality $w^T(1)w^* \geq \rho$ holds.
- **Rationale:** At the initial step ($t = 1$), the expression $w^T(1)w^* = y_i x_i^T w^* = y_i w^{*T} x_i$ exceeds or equals ρ due to the definition of ρ.

Inductive Step:

- **Assumption:** Assuming the validity of $w^T(t)w^* \geq w^T(t-1)w^* + \rho$ for some t.
- **Objective:** To demonstrate $w^T(t+1)w^* \geq w^T(t)w^* + \rho$ for $t+1$.

Proof for Inductive Step:

$$
\begin{aligned}
&\mathrm{w}^T(t+1)w^* = w^T(t)w^* + y_i x_i^T w^* \\
&\geq w^T(t-1)w^* + \rho + y_i x_i^T w^* \\
&= \left(w^T(t-1) + y_i x_i^T\right) w^* + \rho \\
&= w^T(t)w^* + \rho
\end{aligned}
$$

c.

1c. Let's use the update rule for PLA

$w(t) = w(t-1) + y(t-1)x(t+1)$

$\|w(t)\|^2 = \|w(t-1) + y(t-1)x(t-1)\|^2$

$= (w(t-1) + y(t+1)x(t-1))^T (w(t-1) + y(t+1)x(t-1))$

$= w(t-1)^T w(t-1) + 2y(t-1)w(t-1)^T x(t-1) + y(t-1)^2 x(t-1)^T x(t-1)$

$= w(t-1)^T x(t-1) + \|x(t-1)\|^2$

– considering $x(t-1)$ is misclassified, $y(t-1)(w(t-1)^T x(t-1))$

d.

1d. Base case: $t = 0$

$\circ\ \|w(0)\|^2 = 0$

$\circ\ 0 \leq (0) R^2 = 0$

Induction:

at $t+1$, $\|w(t)\|^2 \leq \|w(t-1)\|^2 + \|x(t-1)\|^2$

$" \leq (t-1) R + R^2 = (t) R^2$

Hence, $\|w(t)\|^2 \leq tR^2$

e.

$$\frac{w^T(t)}{\|w(t)\|} w^* \geq \frac{w^T(t-1)\, w^* + \rho}{\|w(t)\|} \geq \frac{t\rho}{\|w(t)\|}$$

$$\frac{t\rho}{\|w(t)\|} \geq \frac{t\rho}{\sqrt{t}\, P} = \frac{\sqrt{t}\, \rho}{R}$$

Therefore, $\dfrac{w^T(t)}{\|w(t)\|} w^* \geq \sqrt{t} \cdot \dfrac{\rho}{R}$

2.

2a. $w + 2\gamma \sum_{i=1}^{N} (w^T x_i - y_i) x_i = 0$

$w + 2\gamma X^T X w - 2\gamma X^T y = 0$

$(I + 2\gamma X^T X) w = 2\gamma X^T y$

$w^* = (I + 2\gamma X^T X)^{-1} 2\gamma X^T y$

2b. N: number of records

d: number of dimensions

the most computationally expensive segment of the algorithm is matrix multiplication

therefore, $O(d^2 (N + d))$