HW3

Guy Oren Oren Renard

January 23, 2017

Theory Questions

1 Solution

Define $e_i =< 0, \ldots, 1, 0, \ldots, 0 >$, so it has value 1 only at the index i, and $S = \{e_i | 1 \leq i \leq d\}$ as set of inputs.

We will proof that S can be shattered by C.

For any $y_1, ..., y_n$ possible labeling for the given inputs, we define the weights as follows:

$$W^{(1)} = \begin{pmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_n \end{pmatrix}$$

$$\forall 1 < t < L : W^{(t)} = I_{d \times d}$$

$$\boldsymbol{w}^{(L)} =< 1, 1, \ldots, 1 >$$

$$\forall 1 \leq t < L : \boldsymbol{b}^{(t)} =< 0, 0, \ldots, 0 >$$

$$b^{(L)} = d - \frac{1}{2}$$

Therefore, for input $z_0 = e_i$, the following always holds:

$$\boldsymbol{z}_1 = h(W^{(1)} z_0 - \boldsymbol{b}^{(1)}) = h(y_i e_i) =< 1, 1, \ldots, \underbrace{y_i}_{i}, \ldots, 1 >$$

For $t + 1 < L$:

$$\boldsymbol{z}_{t+1} = h(W^{(t+1)}z_t - \boldsymbol{b}^{(t+1)}) = h(Iz_t) = h(z_t) = h(< 1, 1, \ldots, \underbrace{y_i}_{i}, \ldots, 1 >)$$

$$=< 1, 1, \ldots, \underbrace{y_i}_{i}, \ldots, 1 >$$

For z_L:

$$z_L = h(\boldsymbol{w}^{(L)}z_{L-1} - b^{(L)}) = h(< 1, 1, \ldots, \underbrace{y_i}_{i}, \ldots, 1 > \cdot \boldsymbol{z}_{L-1} - (d - \frac{1}{2}))$$

$$= \begin{cases} h(\frac{1}{2}), & y_i = 1 \\ h(-\frac{1}{2}), & y_i = -1 \end{cases} = y_i$$

That is, we found a set S of size d that shattered by C, and therefore $VC - dim(C) \geq d$.

2 Solution

2.a

Define $F \triangleq H^d = \{h_1 \times h_2 \times \cdots \times h_d \mid 1 \leq i \leq d \; h_i \in H\}$.

We saw in class that for H and $m \geq d + 1$:

$$\pi_H(m) \leq \big(\frac{em}{d+1}\big)^{d+1}$$

We also saw, that for $G = G_1 \times G_2$:

$$\pi_G(m) \leq \pi_{G_1}(m) \times \pi_{G_2}(m)$$

Therefore, for F we get:

$$\pi_F(m) \leq \big(\frac{em}{d+1}\big)^{d(d+1)}$$

2.b

Define $C \triangleq H \circ \underbrace{F \circ F \circ \cdots \circ F}_{\text{L-1 times}} = H \circ \underbrace{H^d \circ H^d \circ \cdots \circ H^d}_{\text{L-1 times}}$.

We saw in class that for $G = G_1 \circ G_2$:

$$\pi_G(m) \leq \pi_{G_1}(m) \times \pi_{G_2}(m)$$

Therefore, for C we get:

$$\begin{aligned}\pi_C(m) &\leq \left(\frac{em}{d+1}\right)^{(L-1)d(d+1)+d+1} \\ &= \left(\frac{em}{d+1}\right)^{(L-1)d^2+Ld+1}\end{aligned}$$

2.c

$$\begin{aligned}N &= \underbrace{d^2}_{\text{Size of W for L-1 layers}} \times \underbrace{L-1}_{\text{L-1 Layers}} + \underbrace{d}_{\text{Weight for last layer}} + \underbrace{d(L-1)+1}_{\text{Biases}} \\ &= (L-1)d^2 + Ld + 1\end{aligned}$$

2.d

Bonus

2.e

$$\begin{aligned}\pi_C(m) &\leq \left(\frac{em}{d+1}\right)^{(L-1)d^2+Ld+1} \\ &= \left(\frac{em}{d+1}\right)^N \\ &\underbrace{\leq}_{d+1\geq 1} (em)^N\end{aligned}$$

Define $dim \triangleq VC - dim(C)$, then $2^{dim} = \pi_C(dim) \leq (e \times dim)^N$

$$\Rightarrow dim \leq 2Nlog_2(eN)$$

3 Solution

3.a

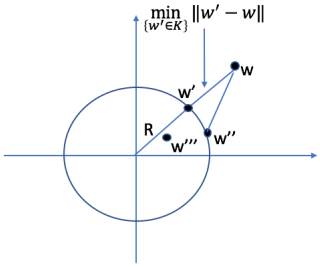
SGD with projection:

1. Initialize $w^{(0)} = 0$
2. For t in $1, \ldots, T$:
   (a) Choose i uniformly at random (u.a.r) from $[m]$.
   (b) Calculate: $w_{mid}^{(t+1)} = (1 - \eta_t)w^{(t)} + \eta_t C y_i x_i$
   (c) Project: $w^{(t+1)} = \pi_K(w_{mid}^{(t+1)})$
3. return $w^{(T+1)}$

When:

$$\pi_K(w) = \begin{cases} w, & \|w\| \leq R \\ R \cdot \frac{w}{\|w\|}, & otherwise \end{cases}$$

Thats because, the closest w' to w with $\|w'\| \leq R$ is with norm R and same direction as w. We can see it clearly for the 2 dimensional case:



Clearly, every w'' with norm R, and every w''' with norm $< R$ has greater distance from w. The above holds for every dimension.

3.b

Bonus

3.c

The outline of the proof we saw in class, is the same for the projection case, but instead of the equation:

$$\left\|W^{(t+1)} - W^*\right\|^2 = \left\|W^{(t)} - W^*\right\| + \eta^2 \left\|V_t\right\|^2 - 2\eta * (W^{(t)} - W^*)V_t$$

We have:

$$\left\|W_{mid}^{(t+1)} - W^*\right\|^2 = \left\|W^{(t)} - W^*\right\| + \eta^2 \left\|V_t\right\|^2 - 2\eta * (W^{(t)} - W^*)V_t$$

Therefore:

$$V_t(W_t - W^*) = \frac{\left\|W^{(t)} - W^*\right\| - \left\|W_{mid}^{(t+1)} - W^*\right\|^2}{2\eta} + \frac{1}{2}\eta \left\|V_t\right\|^2$$

Using the result of (b) we get:

$$\begin{aligned}
&\left\|W^{(t)} - W^*\right\| - \left\|W_{mid}^{(t+1)} - W^*\right\|^2 \\
&= \left\|W^{(t)} - W^*\right\| \underbrace{- \left\|W_{mid}^{(t+1)} - W^*\right\|^2 + \left\|W^{(t+1)} - W^*\right\|^2}_{\text{from (b), its} \leq 0} - \left\|W^{(t+1)} - W^*\right\|^2 \\
&\leq \left\|W^{(t)} - W^*\right\| - \left\|W^{(t+1)} - W^*\right\|^2
\end{aligned}$$

$$\Rightarrow V_t(W_t - W^*) \leq \frac{\left\|W^{(t)} - W^*\right\| - \left\|W^{(t+1)} - W^*\right\|^2}{2\eta} + \frac{1}{2}\eta \left\|V_t\right\|^2$$

And the proof continue as we saw in class.

4 Solution

4.a

If we take $W_1^*(C) = \frac{1}{2}W^*(C')$ and $W_2^*(C) = -\frac{1}{2}W^*(C')$, then we get binary classifier:

$$argmax_{j\in\{1,2\}} w_j x = \begin{cases} 1, & w \cdot x > 0 \\ 2, & otherwise \end{cases}$$

and f become:

$$f = \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m} max\{0, 1 - \underbrace{(2y_i - 3)}_{\in\{-1,1\}} wx_i\}$$

$$= \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m} max\{0, 1 - \underbrace{y_i'}_{\in\{-1,1\}} wx_i\}$$

and That is exactly the form for the binary classifier we saw. Therefore, $C = C'$.

4.b

We saw in class that for $l(w) = max_{j\in[K]} l_j(w)$, if $l_i(w) = l(w)$, then $\nabla l_i(w) \in \partial l(w)$. In our case:

$$s \triangleq argmax_{j\in[K]}(w_j x_i - w_{y_i} x_i + \mathbb{1}_{j\neq y_i})$$

$$\nabla_{w_j}(w_s x_i - w_{y_i} x_i + \mathbb{1}_{j\neq y_i}) = \begin{cases} x_i, & s = j \text{ and } j \neq y_i \\ -x_i, & s \neq j \text{ and } j = y_i \\ 0, & otherwise \end{cases}$$

SGD for multiclass:

1. $\forall j \in [K]: \; w_j^{(0)} = 0$
2. For t in $1, \ldots, T$:
   (a) Choose u.a.r $i \in [m]$
   (b) find $s = argmax_{j\in[K]}(w_j x_i - w_{y_i} x_i + \mathbb{1}_{j\neq y_i})$
   (c) If $s \neq y_i$:

$$w_s^{(t+1)} = (1-\eta) w_s^{(t)} - \eta C x_i$$
$$w_{y_i}^{(t+1)} = (1-\eta) w_{y_i}^{(t)} + \eta C x_i$$

$$\forall j \in [K] \setminus \{s, y_i\} : \; w_j^{(t+1)} = (1-\eta) w_j^{(t)}$$

else:

$$\forall j \in [K] : \; w_j^{(t+1)} = (1-\eta) w_j^{(t)}$$

3. return $w_1, \ldots, w_K$

Predict using $argmax_{j \in [K]} w_j \cdot x$.

4.c

From the previous SGD algorithm, we can see that:

$$w_l^{(T+1)} = \sum_{(*)} (-1)^{\mathbb{1}_{\{l \neq y_{i_J}\}}} (1-\eta)^{(T-J)} \eta C x_{i_J}$$

when (*): (l is argmax in J iteration, and $l \neq y_{i_J}$) or (l is not argmax in J iteration, and $l = y_{i_J}$)

and i_J: is the index we chose u.a.r at iteration J.

To use the algorithm in (b) with kernels, notice that we have to calculate dot products of the form $w_j \cdot x$. Also the prediction, contain products in the same form.

So:

$$\begin{aligned} \Phi(w_j)\Phi(x) &= \sum_{(*)} (-1)^{\mathbb{1}_{\{l \neq y_{i_J}\}}} (1-\eta)^{(T-J)} \eta C \Phi(x_{i_J}) \Phi(x) \\ &= \sum_{(*)} (-1)^{\mathbb{1}_{\{l \neq y_{i_J}\}}} (1-\eta)^{(T-J)} \eta C K(x_{i_J}, x) \end{aligned}$$

Therefore, its enough for every w_j to hold a list (call it, L_j), that for iteration t, we append item as follow:
If $s = argmax_{j \in [K]} (\Phi(w_j)\Phi(x_i) - \Phi(w_{y_i})\Phi(x_i) + \mathbb{1}_{j \neq y_i})$, then:

$$\begin{cases} (t, -1, x_i), & j = s \text{ and } s \neq y_i \\ (t, 1, x_i), & j = y_i \text{ and } s \neq y_i \\ None^{(**)}, otherwise & \end{cases}$$

(**) By None, we mean there is no update.

Meaning, for every update that depend on the sample (other updates are just multiplication of $(1-\eta)$), we have to remember the iteration number, the sign of the update and the sample itself.

Now we can calculate the product $\Phi(w_j)\Phi(x)$ at iteration t + 1 according to the above equality.

The new SGD algorithm, using kernels:

1. $\forall j \in [K]$, Initialize empty list $L_j = []$
2. For t in $1, \ldots, T$:
   (a) Choose i u.a.r from $[m]$
   (b) find $s = argmax_{j\in[K]}(\underbrace{\Phi(w_j)\Phi(x_i) - \Phi(w_{y_i})\Phi(x_i)}_{\text{using the method above, for iteration t}} + \mathbb{1}_{j\neq y_i})$
   (c) for every j, update L_j as described above.
3. return $L_1, \ldots, L_K$

predict using $argmax_{j\in[K]} \underbrace{\Phi(w_j)\Phi(x)}_{\text{using the method above, for iteration T+1}}$.

5 Solution

Let h be some binary classifier as defined. We will show that h can be implemented by decision tree of height at most $d+1$, with internal nodes of the form $(x_i = 0?)$, as follows:

root: $(x_1 = 0?)$
node in depth i, will have 2 sons, with decision stump: $(x_i = 0?)$
and in depth $d+1$, we have 2^d leafs.

$\Rightarrow$ we build a "full" binary decision tree

Each leaf will get label according to the path from root that lead to it, and the binary function h on that path.Because each leaf has distinct path lead to it, we have 2^d labels, 1 for each possible input. Therefore, the above decision tree implement h.

Denote by X the domain, i.e. $X = \{0,1\}^d$.

So, the above shows that there is a set $S = X$ that can be shattered by the concept class C of decision trees.

$$\Rightarrow VC - dim(C) \geq 2^d$$

but obviously $VC - dim(C) \leq 2^d$.

$$\Rightarrow VC - dim(C) = 2^d$$