CSE 512 HW 1

Cole Conte

1 Warm up problem

$$\begin{aligned} P[h(x) \neq f(x)] &= P[y = 1|x \leq 0)] * P[x \leq 0] + P[y = -1|x > 0)] * P[x > 0] \\ &= 0.1 * 0.5 + 0.1 * 0.5 = 0.1 \end{aligned} \tag{1}$$

2 Bayes Optimal Predictor

$$\begin{aligned} L_D(h) = P[h(x) \neq y] &= \begin{cases} P[y \neq 0|x] & \text{if } h(x) = 0 \\ P[y \neq 1|x] & \text{if } h(x) = 1 \end{cases} \\ &= \begin{cases} P[y = 1|x] & \text{if } h(x) = 0 \\ P[y = 0|x] & \text{if } h(x) = 1 \end{cases} \\ &= \begin{cases} P[y = 1|x] & \text{if } h(x) = 0 \\ 1 - P[y = 1|x] & \text{if } h(x) = 1 \end{cases} \end{aligned} \tag{2}$$

To minimize this loss function, choose $h(x) = 0$ if $P[y = 1|x] < 1 - P[y = 1|x]$, otherwise choose $h(x) = 1$. We can rewrite the condition $P[y = 1|x] < 1 - P[y = 1|x]$ as $P[y = 1|x] < 1/2$, and we can restate our prediction function as:

$$h(x) = \begin{cases} 1 & \text{if } P[y = 1|x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

which is just $f_d(x)$, the Bayes optimal predictor. Therefore we've shown $f_d(x)$ minimizes the loss function, that is $L_D(f_D) \leq L_D(g)$ for any classifier g.

3 Unidentical Distributions

Consider $H_B = [h \in H : L_{(\bar{D}_m, f)}(h)]$ to be the set of bad hypotheses and $M = [S^{'} : \exists h \in H_B, L_{S'}(h) = 0]$ to be the set of misleading examples.

$D^m[S : L_{(\bar{D}_m, f}] \leq D^m(M) = D^m(\bigcup_{h \in H_B} [S : L_S(h) = 0])$.

Using the union bound,

$P[\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0] \leq \sum_{h \in H_B}$

$= |H_B| \prod_{i=1}^{m} D_i([x_i : h(x_i) = f(x_i)]) = |H_B| \prod_{i=1}^{m} (1 - L_{D_i, f}(h))$

$\leq |H_B| \prod_{i=1}^{m} (1 - L_{\bar{D}_m, f}(h)) \leq |H_B|(1 - \epsilon)^m \leq |H_B| e^{-\epsilon m} \leq |H| e^{-\epsilon m}$

4 VC Dimension

4.1 Part a

Using axis-aligned rectangles, we can capture two points per dimension (the points with the maximum and minimum coordinates in the dimension), so we can shatter a set of size $2d$. With a set of size $2d + 1$, one of the points must

be contained within one of our axis-aligned rectangles, as opposed to on the edge. But we can't label a point within a rectangle differently than the points on the edge of the rectangle, so we can not shatter a set of size $2d+1$. Therefore $VCdim(H^d) = 2d$.

4.2 Part b

This hypothesis function is a square wave function. By manipulating the frequency, we can produce whatever labeling we want for a set of form $\{2^{-n} : n = 0, 1...\}$ of any size. Therefore $VCdim(H) = \infty$.

5 Boosting

$$\begin{aligned}\sum_{i=1}^{m} D_i^{(t+1)} 1_{[y_i \neq h_t(x_i)]} &= \frac{\sum_{i=1}^{m} D_i^{(t)} e^{-w_t y_i h_t(x_i)} 1_{[y_i \neq h_t(x_i)]}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_i h_t(x_i)}} \\ &= \frac{e^{w_t}\epsilon_t}{e^{w_t}\epsilon_t + e^{-w_t}(1-\epsilon_t)} = \frac{\epsilon_t}{\epsilon_t + e^{-2w_t}(1-\epsilon_t)} \\ &= \frac{\epsilon_t}{\epsilon_t + \frac{\epsilon_t}{1-\epsilon_t}(1-\epsilon_t)} = \frac{1}{2}\end{aligned} \tag{4}$$

6 Learnability of logistic regression

H is a convex set, since we can draw a line from any two points in H without leaving H. For all $x \in H$ we have $||x|| \leq B$ by the definition. To show convexity of the loss function, let $y\langle w, x\rangle = z$.

$$\begin{aligned} l(w,(x,y)) &= log[1 + exp(-z)] \\ l^{(1)}(w,(x,y)) &= \frac{-exp(-z)}{1+exp(-z)} \\ l^{(2)}(w,(x,y)) &= \frac{exp(-z)}{(1+exp(-z))^2} \end{aligned} \tag{5}$$

$l^{(2)}(w,(x,y))$ is clearly nonnegative for all values of w,x,y, so $l(w,(x,y))$ is a convex function. To show that the loss function is L-Lipschitz:

$$\begin{aligned}
&||l(w_1,(x,y)) - l(w_2,(x,y))|| \\
&= ||log[1+exp(-y\langle w_1,x\rangle)] - log[1+exp(-y\langle w_2,x\rangle)]|| \\
&= ||log[1+exp(-y\langle w_1,x\rangle)] + log[\frac{1}{1+exp(-y\langle w_2,x\rangle)}]|| \\
&\leq ||log[1+exp(-y\langle w_1,x\rangle)]|| + ||log[\frac{1}{1+exp(-y\langle w_2,x\rangle)}]|| \\
&= ||log[1+exp(-y\langle w_1,x\rangle)]|| + ||log[1+exp(-y\langle w_2,x\rangle)]|| \\
&\leq ||1+exp(-y\langle w_1,x\rangle) - 1|| + ||1+exp(-y\langle w_2,x\rangle) - 1|| \\
&= ||exp(-y\langle w_1,x\rangle)|| + ||exp(-y\langle w_2,x\rangle)|| \\
&= 1 * ||exp(-y\langle w_1,x\rangle) + exp(-y\langle w_2,x\rangle)||
\end{aligned} \tag{6}$$

Therefore the loss function is 1-Lipschitz. We've therefore shown that this problem meets all of the conditions of being convex-Lipschitz-bounded. To show that the loss function is smooth:

$$\begin{aligned}
||\nabla l(v,(x,y)) - \nabla l(w,(x,y))|| &= ||\nabla l(v,(x,y)) - \nabla l(w,(x,y))|| \\
&= ||\frac{-exp(-y\langle v,x\rangle)}{1+exp(-y\langle v,x\rangle)} + \frac{exp(-y\langle w,x\rangle)}{1+exp(-y\langle w,x\rangle)}|| \\
&= ||\frac{exp(-y\langle w,x\rangle) - exp(-y\langle v,x\rangle)}{(1+exp(-y\langle v,x\rangle))(1+exp(-y\langle w,x\rangle))}|| \\
&\leq ||\frac{1+(-y\langle w,x\rangle) - 1 - (-y\langle v,x\rangle)}{(1+exp(-y\langle v,x\rangle))(1+exp(-y\langle w,x\rangle))}|| \\
&\leq ||(-y\langle w,x\rangle) - (-y\langle v,x\rangle)||
\end{aligned} \tag{7}$$

Therefore the loss function is 1-smooth. We've therefore shown that this problem meets all of the conditions of being convex-smooth-bounded.

7 Learnability of halfspaces with hinge loss

First we must show that the set of halfspaces is convex. Take any pair of points$(x_1,y_1),(x_2,y_2)$ in the halfspace where $y \geq 0$. Those points will both have $y_i \geq 0$. The line segment that connects these two points is given by $r(t) = t(x_1,y_1) + (1-t)(x_2,y_2)$ for $0 \leq t \leq 1$. The y-value for this line segment must be positive, since $t, 1-t, y_1, y_2$ are all positive, therefore the line segment must be contained entirely within the halfspace, making the halfspace convex. WLOG we can say that our set of halfspaces is convex. Next we need to show that our loss function is convex. The hinge loss function is made up of convex functions, and the maximum of convex functions is convex, therefore the hinge loss function is convex. Finally we need to show that the loss function is R-Lipschitz, that is $||l(v,(x,y)) - l(w,(x,y))|| \leq ||-y\langle v,x\rangle + y\langle w,x\rangle||$ We can break this into a proof by cases:

Case 1: $1 - y\langle v, x\rangle \leq 0, 1 - y\langle w, x\rangle \leq 0$.

$$||0 - 0|| \leq || - y\langle v, x\rangle + y\langle w, x\rangle|| \tag{8}$$

Case 2: $1 - y\langle w, x\rangle \leq 0, 1 - y\langle v, x\rangle > 0$ (the opposite case follows).

$$||1 - y\langle v, x\rangle|| \leq || - y\langle v, x\rangle + y\langle w, x\rangle|| \tag{9}$$

since our condition implies that $y\langle w, x\rangle \geq 1$.

Case 3: $1 - y\langle v, x\rangle > 0, 1 - y\langle w, x\rangle > 0$.

$$||1 - y\langle v, x\rangle - 1 + y\langle w, x\rangle|| = || - y\langle v, x\rangle + y\langle w, x\rangle|| \tag{10}$$

Therefore the loss function is 1-Lipschitz.

8 Cross-validation

We must break this problem into cases. Let a be the number of ones and b be the number of 0s. The true error is:

$\frac{b}{a+b}$ if a is odd, and $\frac{a}{a+b}$ if a is even.

Each iteration of the LOOCV error for odd a: Pull a 1, error is $\frac{a-1}{a+b-1}$. Pull a 0, error is $\frac{b-1}{a+b-1}$. Therefore the overall LOOCV error for odd a is: $\frac{a\frac{a-1}{a+b-1}+b\frac{b-1}{a+b-1}}{a+b}$

$|\frac{a\frac{a-1}{a+b-1}+b\frac{b-1}{a+b-1}}{a+b} - \frac{b}{a+b}| = |\frac{a(a-b-1)}{a+b-1}|$

Each iteration of the LOOCV error for even a: Pull a 1, error is $\frac{b}{a+b-1}$. Pull a 0, error is $\frac{a}{a+b-1}$. Therefore the overall LOOCV error for odd a is: $\frac{\frac{ab}{a+b-1}+\frac{ba}{a+b-1}}{a+b}$.

$|\frac{\frac{ab}{a+b-1}+\frac{ba}{a+b-1}}{a+b} - \frac{a}{a+b}| = |\frac{a(-a+b+1)}{a+b-1}|$ The two differences in error are equal, and must sum to 1, so we've shown that the difference between the LOO estimate and true error is always 0.5.

9 Local minimum

Take $w = (-2, -2)$ and $w^* = (1, 1)$, and a classifier that labels points 1 if they have y-values greater than 0, and -1 if they have y-values less than or equal to 0. The correct labels of every point in this set are 1 for x-values greater than 0, and -1 for x-values less than or equal to 0.

Take $\epsilon = 1$. Then for any w' such that $||w - w'|| \leq \epsilon = 1$, we have $L_S^{(01)}(w) = L_S^{(01)}(w') = 1$, showing that w is a local minimum.

But there also exists w^*, which has a 0-1 loss of 0. So $L_S^{(01)}(w^*) < L_S^{(01)}(w)$. Therefore w is not a global minimum.