«به نام خدا»

تمرین سری هفتم

حسین ابراهیمی - ۹۵۱۰۵۳۰۲

سوال T24

$$\underset{S\sim D^m}{E}[L_S(h)] = \underset{z_i\sim D}{E}\left[\frac{1}{m}\sum_{i=1}^{m} l(h,z_i)\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \underbrace{\underset{z_i\sim D}{E}[l(h,z_i)]}_{L_D(h)} = \frac{1}{m}\sum_{i=1}^{m} L_D(h) = L_D(h)$$

سوال T25

$x, x' \in \{1,2,\dots,N\}$, $N \in$ Integer Number

$K(x,x') = \min(x,x')$, find $\psi : \{1,2,\dots,N\} \longrightarrow H_{space}$.

نگاشت ψ را بدین صورت تعریف می‌کنیم که به تعداد آن عدد طبیعی ورودی، 1 قرار می‌دهیم و بقیه را صفر می‌گذاریم یعنی:

$$\psi(i) = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} i \\ \\ \\ \end{matrix} \in H$$

حال برای دو عدد دلخواه i و j، ضرب داخلی آن‌ها برابر خواهد بود با:

$$\left\langle \psi(i) = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Big\} i \ , \ \psi(j) = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{bmatrix} \Big\} j \right\rangle$$

$$= \underbrace{1+1+\dots+1}_{\text{مقدار کوچکتر بین i و j}} + 0 + \dots = \min(i,j) = K(i,j)$$

سوال T26

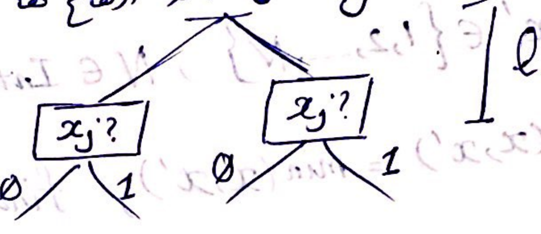
$$h_\epsilon : \{0,1\}^d \to \{0,1\}$$

.1

$$h(x_1, x_2, x_3, \dots, x_d) = 0 \text{ یا } 1$$

باتوجه به اینکه بُعد ورودی ما برابر با d می‌باشد پس 2^d حالت مختلف برای ورودی خواهیم داشت که باینری h - classifier می‌تواند به هر کدام دلخواه یک مقدار صفر یا یک بدهد.

حال برای ساختن درخت تصمیم‌گیری ساختار کافی است در هر گره که در ارتفاع معین l قرار دارد به دلخواه براساس یکی از x_j ها که در ارتفاع های بالاتر تفکیک نشده است، جداسازی انجام دهیم:



حداکثر ارتفاع درخت برابر با $d+1$ خواهد بود چون یا در مراحل بالا مقادیر نسبت داده شده یا با h برابر می‌شود یا نیاز است گره های جدید اضافه کنیم که حداکثر d مرتبه امکان پذیر است.

حال در برگ های درخت با ارتفاع $d+1$، تمام حالت های ممکن ورودی که 2^d حالت داشت وجود دارد و ما می‌توانیم براساس اینکه مقدار h آن ها چیست به آن ها label مناظر را بزنیم و کاملاً h را پیاده سازی کنیم.

.2

ابتدا باید نشان دهیم که Decision Tree شامل توابعی است که می‌تواند تمام حالت های مقداردهی به مجموعه ای به اندازه 2^d را در $Domain = \{0,1\}^d$ بپوشاند یا به عبارتی shatter کند.

بدین منظور همانند حالت قبل، درختی با ارتفاع $d+1$ را درنظر می‌گیریم که در هر سطح از آن گره‌ها $\boxed{x_i\ ?}$ قرار دارند. حال هر برگ‌های این درخت که تعدادشان 2^d است، تمام حالت‌های $(x_1, x_2, ..., x_d)$ را دارد. حال ما می‌توانیم با label قرار دادن برای برگ‌ها تمام حالت‌های مختلف مقدار دهی به این 2^d عضو را داشته باشیم. دقت کنید هر حالت مقدار دهی یک h را مشخص می‌کند در H یعنی DT_s قرار دارد. در نتیجه

2^{2^d} حالت مقدار دهی متفاوت به این مجموعه در H وجود دارد پس:

H shatters C با تعداد

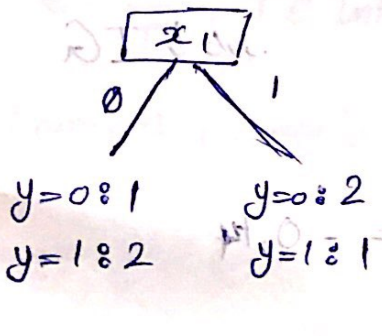
حال از طرفی چون مجموعه‌ای بزرگتر از 2^d در Domain نداریم پس مجموعه‌ای به سایز 2^d+1 وجود ندارد. در نتیجه VC dimension مجموعه توابع Desicion Tree برابر با 2^d خواهد بود.

سوال T27

x_1	x_2	x_3	x_4	y
1	0	1	0	0
1	1	0	0	0
0	0	1	1	0
1	0	0	1	1
0	1	1	1	1
0	0	0	1	1

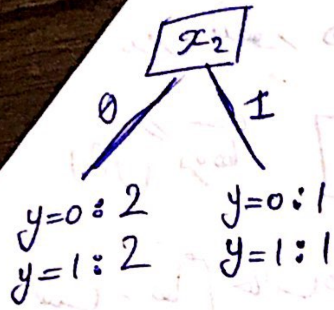
تابع Gain را خود Information Gain قرار می‌دهیم.

الف)



$$C(\mathbb{P}_S[y=1]) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = -\log\frac{1}{2}$$
$$= \log_2 2 = 1$$

$$Gain(S, x_1) = 1 + \left[+\frac{1}{2}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) + \frac{1}{2}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)\right] = 0.08$$



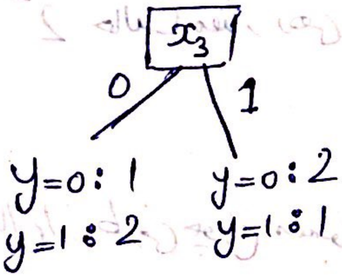
$$C(\mathbb{P}_S(y=1)) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6}$$

$$= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.91$$

$$Gain(S, x_2) = 0.91 + \Big[\frac{4}{6}\Big(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}\Big)$$

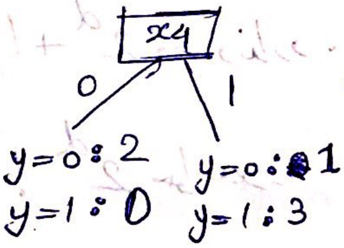
$$+ \frac{2}{6}\Big(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\Big)\Big] = 0.91 + \frac{2}{3}\log\frac{1}{2} + \frac{1}{3}\log\frac{1}{2}$$

$$= 0.91 + \log\frac{1}{2} = 0.91 - \overset{1}{\log 2} = \underline{-0.091}$$



x_1 مانند : $Gain(S, x_3) = 0.08$

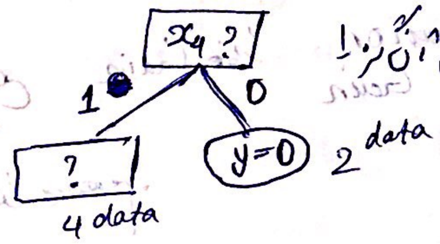
$$C(\mathbb{P}_S[y=1]) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.91$$



$$Gain(S, x_4) = 0.91 + \Big[\frac{4}{6}\Big(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\Big)$$

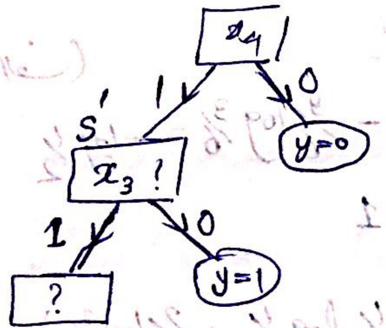
$$+ \frac{2}{6}\big(\overset{0}{\log_2 1} + \cancel{\log 0 \times 0}\big)\Big] = 0.91 + \frac{-0.54}{3/2} = \boxed{0.36}$$

x_4 بیشترین Information Gain را دارد پس براساس آن split میکنیم.



وقتی x_4 برابر با 0 است تمام داده ها y ای برابر با 1 دارند پس آن را برگ میکنیم.

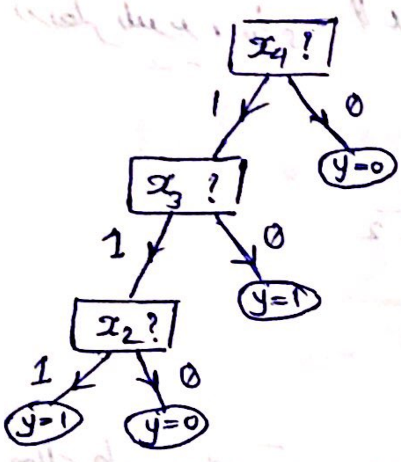
حال در مرحله بعد باید براساس x_3 جداسازی کنیم زیرا بیشترین IG را دارد.



$$IG(S', x_3) = 0.5 \checkmark$$

$$IG(S', x_1) = IG(S', x_2) = -0.14$$

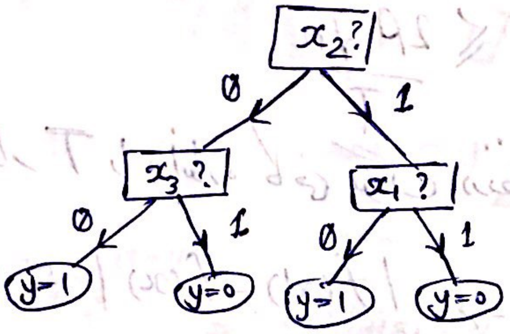
در مرحله آخر براساس x_2 split می‌کنیم.
زیرا جداسازی کامل دارد درصورتی که برای x_1 اینگونه نیست.



داده S کاملاً جداسازی می‌شود در نتیجه: $L_S(h_S) = 0$

ب)

$$y = \bar{x}_1 . x_2 + \bar{x}_3 . \bar{x}_2$$



$L_S(h) = 0$

ج) بنابر 2 دلیل: 1- پیچیدگی محاسباتی زیاد 2- overfiting

پیدا کردن عبارت منطقی‌ای که S را بیان کند از لحاظ محاسباتی سنگین است همچنین می‌تواند تعداد زیادی عبارت منطقی وجود داشته باشد که S را توضیح دهد و پیدا کردن عمیق‌ترین آن‌ها کار دشواری از لحاظ محاسباتی است.

همچنین با این کار، مدل ما کاملاً بر داده آموزشی fit می‌شود در نتیجه احتمال overfiting بسیار بالا می‌رود.

سوال T28

برای مقدار fixed شده ϵ عبارت معادل را در نظر می‌گیریم: $T \in$ Integer و $\epsilon = \frac{1}{T}$

سپس هر یک از n محور $[-1,1]$ را به T قسمت تقسیم می‌کنیم. در این صورت n جعبه با ابعاد $\frac{2}{T}$ خواهیم داشت. حال دو نقطه x و x' را داخل یکی از این جعبه‌ها

درنظر بگیرید، برای $\|x'-x\|$ خواهیم داشت:

$$\|x'-x\|^2 \leqslant \frac{4}{T^2}+\frac{4}{T^2}+\dots+\frac{4}{T^2}=\frac{4n}{T^2}$$

$$\Rightarrow \|x'-x\| \leqslant \frac{2\sqrt{n}}{T}$$

حال طبق خاصیت ρ_lipschitz داریم:

$$|f(x')-f(x)| \leqslant \rho\|x'-x\| \leqslant \frac{2\rho\sqrt{n}}{T}$$

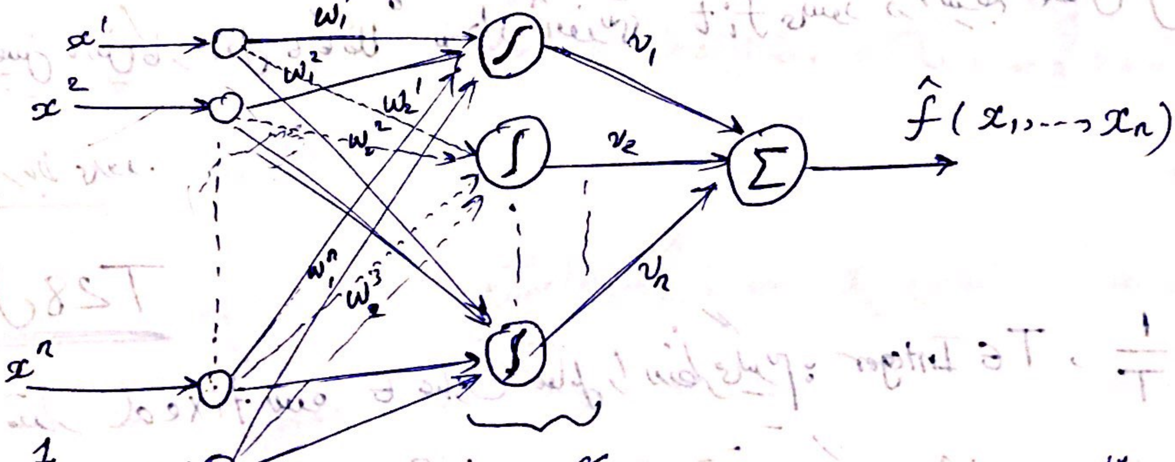
حال اگر مقدار T را به اندازه کافی بزرگ درنظر بگیریم خواهیم داشت که:

$$T\to\infty:\ \lim_{T\to\infty}\frac{2\rho\sqrt{n}}{T}\simeq 0 \Rightarrow |f(x')-f(x)|\simeq 0$$

$$\Rightarrow f(x')\simeq f(x)$$

درنتیجه مقادیر داخل هر جعبه به طور تقریبی ثابت هستند.

Approximate func : $G(x)=\sum_{i=1}^{n} v_i\,\phi(w_i^T x+b_i)$



در لایه پنهان اول در واقع داریم مقادیر مختلف میانگین را به ازای وزن‌ها مختلف حساب می‌کنیم

سپس در لایه $\sum$ از این میانگین‌ها Average می‌گیریم.

سوال T29

در ابتدا داریم:

$$D_i^{(t+1)} = \frac{D_i(t)\, e^{-w_i y_i h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)}\, e^{-w_t y_j h_t(x_j)}}$$

حال

$$\sum_i D_i^{(t+1)}\, \mathbb{1}_{[y_i h_t(x_i) = -1]} = \sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t+1)}$$

الان باید اثبات کنیم که آیا عبارت بالا برابر با $\frac{1}{2}$ است یا خیر.

$$\Rightarrow \sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t+1)} = \frac{\sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t)} e^{w_t}}{\sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t)} e^{w_t} + \sum_{i:\, y_i h_t(x_i) \neq -1} D_i^{(t)} e^{-w_t}}$$

حال برای اینکه کسر بالا برابر با $\frac{1}{2}$ شود، لازم است اثبات کنیم که دو عبارت در مخرج کسر با هم برابراند پس:

$$\sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t)} e^{w_t} = \sum_{i:\, y_i h_t(x_i) \neq -1} D_i^{(t)} e^{-w_t} \quad \div e^{-w_t} \Longrightarrow$$

$$\sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t)} e^{2w_t} = \sum_{i:\, y_i h_t(x_i) \neq -1} D_i^{(t)} = 1 - \sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t)}$$

$$\Rightarrow (1 + e^{2w_t}) \underbrace{\sum_{i:\, y_i h_t(x_i) = -1} D_i^{(t)}}_{\epsilon_t} = 1$$

درنهایت کافی است اثبات کنیم که عبارت $(1+e^{2w_t})\,\epsilon_t$ برابر با یک است.

$$\Big(1+\underbrace{e^{2\left(\frac{1}{2}\log\left(\frac{1}{\epsilon_t}-1\right)\right)}}_{\frac{1}{\epsilon_t}-1}\Big)\,\epsilon_t = \frac{1}{\epsilon_t}\cdot\epsilon_t = 1$$

درنتیجه حکم اثبات شد:

$$\sum_{i:\, y_i h_t(x_i)=-1} D_i^{(t+1)} = \frac{1}{2}$$