سوال 1 ها

در VC-theory یک اندازه گیری از ظرفیت (پیچیدگی) VC-dim، مجموعه ای از توابع است که می تواند توسط یک الگوریتم طبقه بندی باینری آموخته شود.
برای آن که VC-dim را به صورت عمیق تر بررسی کنیم باید بدانیم که:

ما می گوییم $\mathcal{H}$ یک مجموعه می رود نقاط C داخل domain set را shatter می کند اگر $\mathcal{H}_C$ تمامی توابعی که از این مجموعه ی C به صفر یا یک می روند را بسازد یعنی تمامی حالت ها را داشته باشیم. ($\mathcal{H}_C = \{(h(c_1), h(c_2), \dots, h(c_m)) : h \in \mathcal{H}\}$)

حال برای VC-dim داریم: بیش ترین عضو مجموعه ی C به صورتی که $\mathcal{H}$ ها بتواند C را shatter کند. برای آن که نشان دهیم $VC\text{-}dim(\mathcal{H}) = d$ می باشد داریم:

① ∃ set C of size d that is shatterd by $\mathcal{H}$

② ∀C of size d+1 that is not shattered by $\mathcal{H}$

حداکثر مقدار VC-dim زمانی است که برای هر h یکی از حالات output را داشته باشیم بنابراین خواهیم داشت:

$$VC\text{-}dim(\mathcal{H}) \leq \lceil \log_2(|\mathcal{H}|) \rceil$$

حال که با موضوع VC-dim آشنا شدیم می خواهیم برویم سراغ تئوری اصلی یادگیری ماشین و ارتباط آن با VC-dim. براساس قضیه ی اساسی یادگیری آماری داریم:

Let $\mathcal{H}$ be a hypothessis class of function from $\mathcal{X}$ to $\{0,1\}$ and the loss function be 0-1 loss then the followings are equivalent:

① $\mathcal{H}$ is Agnostic PAC learnable

② $\mathcal{H}$ is PAC learnable

③ Any ERM rule is PAC-learnable for $\mathcal{H}$

④ $\mathcal{H}$ has a finite VC-dim

همچنین از theory of learning (quantative V) داریم:

Assume that $VC\text{-}dim(\mathcal{H}) = d < \infty$ then there are constants C_1, C_2, C_1', C_2' such that

① $\mathcal{H}$ is Agnostic PAC-learnable with sample complexity

$$C_1 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$

② $\mathcal{H}$ is PAC-learnable with sample complexity:

$$C_1' \frac{d + \log(\frac{1}{\delta})}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2' \frac{d + \log(\frac{1}{\delta})}{\varepsilon^2}$$

درواقع اگر به من یک $\mathcal{H}$ ای دهند اول باید چک کنم که قابل PAC یا Agnostic PAC یادگیری هست یا نه. کلید این کار $VC\text{-}dim$ است. اگر محدود بود ✓ اوکی و سپس می فهمیم از چه اردری sample نیاز داریم. حال که با اصل قضیه آشنا شدیم داریم

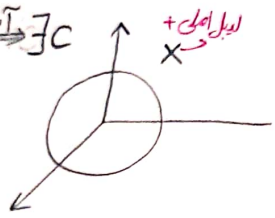
الف) این classifier دوواقع مدل کننده‌ی یک کره است. همچنین از بخش قبل می‌دانیم برای آن که بگوییم $VC\text{-}dim(H)=d$ باید (این classifier داده‌های داخل کره را label-ی یکسان و با خارج را label-ی دیگر می‌زند)

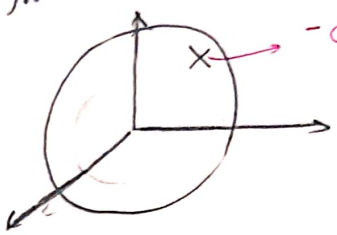
$\exists$ set C of size d that is shatterd by $\mathcal{H}$

$\forall C$ of size $d+1$ " " not " " "

بنابراین از اثبات می‌کنیم و جلو می‌رویم...

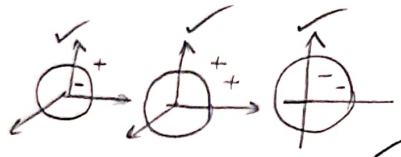
$d=1 \xrightarrow{\text{یک}} \exists C$

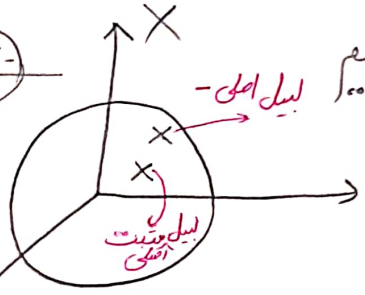




$\forall C \Longleftarrow$ ✓

حالا برای $d=2$ داریم... (باید چک کنیم $\forall C$ داریم یا نه)

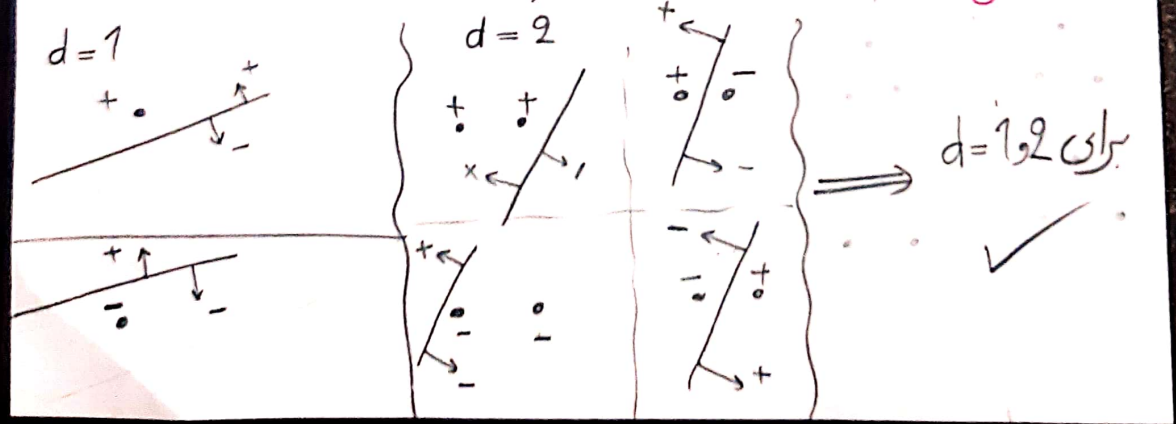




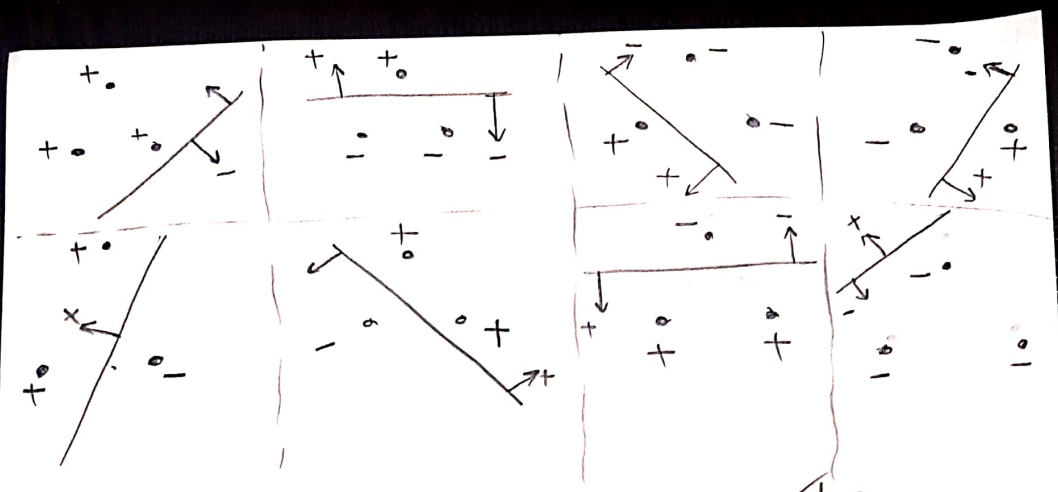
$d=2 \xrightarrow{\text{یک}}$

$VC\text{-}dim=1 \Longleftarrow$ classifier داده‌های بالا را نمی‌تواند از هم جدا کند بنابراین

ب) برای $\operatorname{sgn}(x_1\theta_1+x_2\theta_2+\theta_0)$ داریم:

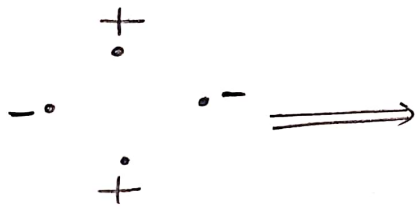


✓



$\Longleftarrow$ بر $d=3$ ✓

اما برای $d=4$ داریم:



به هیچ روشی نمی‌تواند $\Longrightarrow$

$\Longleftarrow VC\_dim=3$

ج) برای حالت کلی N داریم $\Longleftarrow N+1$

اثبات: فرض می‌کنیم $d+1$ نقطه در فضای d بعدی داریم

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix} \quad \text{for } i = 1, \dots, d+1$$

حال می‌خواهیم نشان دهیم آرایش دلخواهی از این $d+1$ نقطه وجود دارد که shattering پذیر باشد. می‌دانیم $h(x)$ در Half-space به این گونه predict می‌کند

$$h(x_i) = sgn\left(w_1 x_{i1} + \dots + w_d x_{id} + b\right)$$

فرض کنید خروجی‌های هدف y_i باشند، $i = 1, \dots, d+1$

فرض کنید y_1 تا y_{d+1} یک labeling دلخواه برای x_1 تا x_{d+1} باشند

$d+1$ معادله خواهیم داشت:

$$\text{sgn}(w_1 x_{11} + \dots + w_d x_{1d} + b) = y_1$$

$$\text{sgn}(w_1 x_{(d+1)1} + \dots + w_d x_{(d+1)d} + b) = y_{d+1}$$

اگر معادله را به صورت زیر بنویسیم و جواب داشته باشد معادله‌ی با الاهم جواب خواهد داشت:

$$w_1 x_{11} + \dots + w_d x_{1d} + b = y_1$$

$$w_1 x_{(d+1)1} + \dots + w_d x_{(d+1)d} + b = y_{d+1}$$

که معادله‌ی بالا را می‌توان به صورت $AX = Y$ نوشت. اگر آرایش نمونه‌ها طوری باشد که ماتریس A به ابعاد $(d+1) \times (d+1)$ وارون‌پذیر باشد w_1 تا w_d و b به صورت یکتا به دست خواهند آمد

$\Longleftarrow$ برای $d+1$ نقطه آرایش خاصی وجود دارد که امکان shattering Half Space پذیر است

حال فرض کنید $d+2$ نقطه داشته باشیم. طبق قضیه‌ی Radon هر مجموعه‌ی $d+2$ تایی از نقاط در $\mathbb{R}^d$ می‌تواند به دو مجموعه‌ی disjoint تقسیم شود به طوری که convex hull آن‌ها متقی باشد. حال اگر یکی از این مجموعه‌ها را label + و دیگری را label منفی بزنیم نمی‌توان با هیچ half space ای نقاط داخل اشتراک را درست label زد

$$\Longrightarrow \quad VC\text{-}dim(\mathcal{H}) < d+2$$

طبق توضیحات بالا $VC\text{-}dim(\mathcal{H}) = d+1$

سوال ۲:

الف) برای به دست آوردن رگرسیون خطی بر روی داده‌ها داریم:

$$\hat{y}^{(i)} = \theta^T x^{(i)} \rightarrow y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}, \quad where \quad \epsilon^{(i)} \sim N(0, \sigma^2), \qquad \epsilon^{(i)} \; are \; i.i.d.$$

با توجه به قاعده maximum likelihood داریم:

$$P(Y|X,\theta) = \prod_{i=1}^{n} P\left(y^{(i)}|x^{(i)},\theta\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - \theta^T x^{(i)}\right)^2\right)$$

باید Y به گونه‌ای انتخاب شود که این احتمال را ماکزیم کند؛ یا به طور معادل عبارت $-\ln\left(P(Y|X,\theta)\right)$ را مینیمم کند:

$$-\ln\left(P(Y|X,\theta)\right) = -\sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

با توجه به این‌که برای همه مقادیر Y مقدار σ یکسان است؛ کافیست عبارت زیر را مینیمم کنیم که همان مینیمم کردن خطای MSE در رگرسیون خطی است:

$$\min \sum_{i=1}^{n}\left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

سوال ۳:

$$f_X(x,\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i,\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)$$

$$\rightarrow \quad \ln f_X(x,\mu,\sigma^2) = -\sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

سوال ۴:

الف) در رگرسیون خطی باید تابع هزینه که به صورت MSE تعریف می‌شود را کمینه کنیم: یعنی:

$$\min J(\theta) = \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2$$

می‌توانیم $h_\theta(x_i)$ را برای رگرسیون خطی به صورت زیر بنویسیم:

$$h_\theta(x_i) = \theta^T x^{(i)} \rightarrow \min \sum_{i=1}^{n} (\theta^T x^{(i)} - y_i)^2$$

$$\rightarrow J(\theta) = (X\theta - y)^T (X\theta - y)$$

$$\frac{\partial J(\theta)}{\partial \theta} = 2X^T X\theta - 2X^T y = 0 \rightarrow \theta = (X^T X)^{-1} X^T y$$

ب) با اضافه کردن L2 Regularization در تابع هزینه داریم:

$$J(\theta) = \sum_{i=1}^{n} (\theta^T x^{(i)} - y_i)^2 + \lambda ||\theta||_2^2 = (X\theta - y)^T (X\theta - y) + \lambda \theta^T \theta$$

$$\frac{\partial J(\theta)}{\partial \theta} = 0 \rightarrow 2X^T X\theta - 2X^T y + 2\lambda\theta = 0 \rightarrow (X^T X + \lambda I)\theta = X^T y$$

$$\rightarrow \theta = (X^T X + \lambda I)^{-1} X^T y$$