X = input space   Y: output space

$f: X \to Y$ unknown function

↳ f deterministic değil / stochastic

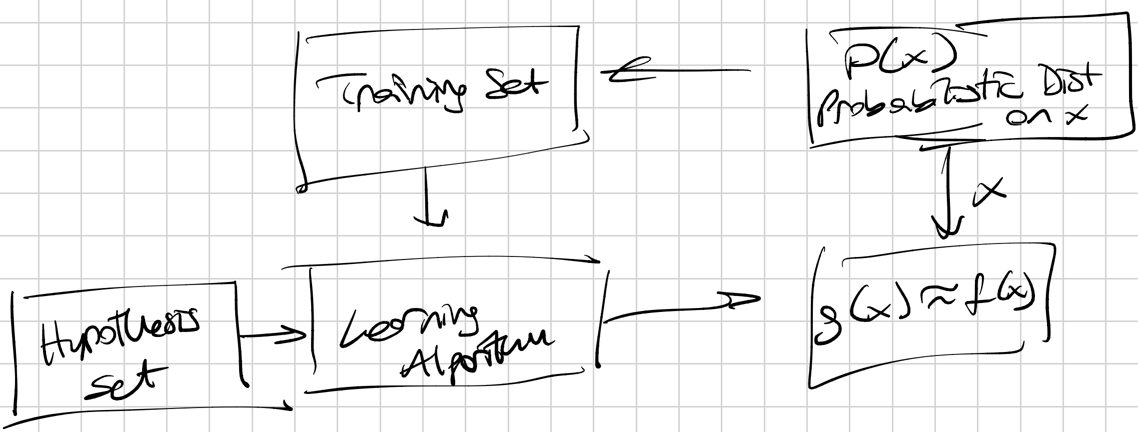
$y = f(x) \to \underbrace{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)}$

bu training set'e

bir "learning algorithm" uyguluyoruz.

* Çok büyük bir küme seçersek sıkıntı çıkabilir (over-fit vs)   Buna ben katılmıyorum

O yüzden bir "hypothesis set" seçiyoruz ve bu set üzerinde learning algorithm uyguluyoruz.



Want to minimize:

$$\mu = \underset{\substack{\downarrow \\ \text{expected value}}}{\mathbb{E}} \left[\mathbb{1} \left(g(x) \neq f(x) \right) \right] \qquad \mathbb{1}(A) = \begin{cases} 1, & A \text{ true} \\ 0, & A \text{ false} \end{cases}$$

Instead we minimize training set'teki ortalama hata:

$$\nu = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left(g(x_i) \neq f(x_i) \right)$$

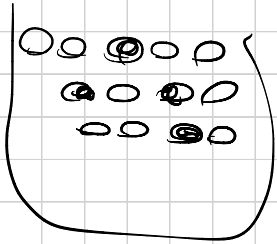
$\nu \sim \mu$ soru şu: ikisi arasındaki fark ne? ν μ'ya ne kadar yakın

Bunu istatistiksel olarak genel garanti, haftaya da ispat

Bu, hipotez kümesinin esnekliğine gidecek.

ex:

diyelim ki elimizde bir cup var:



$$\mu = \frac{\# \bullet}{\# \circ + \bullet}$$ size random bağımsız örnek in cup

$$\nu = \frac{\# \bullet}{\# \circ + \bullet}$$ in sample

If $\mu = 0.9$, then what's the probability of $\nu = 0.1$ where sample size is $N = 10$?

1 top çektim

$\bullet\circ\circ\circ\circ \rightarrow 0.9 \cdot (0.1)^9$

$\circ\bullet\circ\circ\circ \rightarrow 0.1 \cdot 0.9 \cdot (0.1)^8$

$\circ\circ\bullet\circ\circ\circ$

$$P(\nu = 0.1) = \boxed{10 \cdot (0.9) \cdot (0.1)^9}$$

bunlar da eksik

$$P(\nu = 0) = (0.1)^{10}$$

$$P(\nu <= 0.1) = P(\nu = 0.1 + \nu = 0)$$

$$= 9 \cdot (0.1)^9 + (0.1)^{10} = \frac{9}{10^9} + \frac{1}{10^{10}} = \frac{91}{10^{10}}$$

Hoeffding's Inequality:

Theorem: let $z_1, z_2, \ldots, z_N$ be independent random variables on $\mathbb{R}$ such that $a_i <= z_i <= b_i$ with probability one.

If $S_n = \sum_{i=1}^{n} z_i$, then for all $t > 0$

$$\mathbb{P}\left[S_n - \mathbb{E}(S_n) \geq t\right] <= e^{-2t^2 / \Sigma (b_i - a_i)^2}$$

and

$$\mathbb{P}\left[S_n - \mathbb{E}(S_n) <= t\right] <= e^{-2t^2 / \Sigma (b_i - a_i)^2}$$

then

$$\mathbb{P}\left[|S_n - \mathbb{E}(S_N)| \geq t\right] <= 2e^{-2t^2 / \Sigma (b_i - a_i)^2}$$

Unutma: $\mathbb{P}(A \vee B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

Remark: $z_i \sim \text{Bern}(\mu)$ $\quad z_i = \begin{cases} 1 & \text{with prob } \mu \\ 0 & \text{with prob } 1-\mu \end{cases}$

$a_i = 0 \quad b_i = 1$

Bernoulli ve Binomial dağılımını birbirinden ayırmak önemli.

then

$S_N = z_1 + \ldots z_N \sim \text{Binom}(\mu)$ and

$\mathbb{E}[S_N] = N \cdot \mu$ therefore

$$\mathbb{P}[|S_N - \underbrace{\mathbb{E}[S_N]}_{N\cdot\mu}| >= t] <= 2e^{-2t^2 / \sum_{i=1}^{N}(1-0)^2}$$

$$\Rightarrow \mathbb{P}[|S_N - N\cdot\mu| >= t] <= 2e^{-2t^2/N}$$

$$\Rightarrow \mathbb{P}\left[\left|\frac{S_N}{N} - \mu\right| >= \frac{t}{N}\right] <= 2e^{-2(\frac{t}{N})^2/N}$$

$\varepsilon = t/N$

$$\Rightarrow \mathbb{P}\left[\left|\frac{S_N}{N} - \mu\right| >= \varepsilon\right] <= 2e^{-2\varepsilon^2 N} = \frac{2}{e^{2\varepsilon^2 N}}$$

$\frac{S_N}{N}$: sample'deki topun siyah olma olasılığı

μ: cup'taki topun siyah olma olasılığı

$$S_n \sim \text{Binom}(n, \mu)$$

$$\mathbb{P}(S_n = 1) = \binom{10}{1} \cdot \mu^1 (1-\mu)^{10-1}$$

$$\mathbb{P}(S_n = 0) = (1-\mu)^{10}$$

$$\mathbb{P}(S_n = k) = \binom{n}{k} \mu^k (1-\mu)^{n-k}$$

Sample size büyük verirsek olasılık küçülüyor, böylece aradaki farkın azalması sağlanıyor $(\nu - \mu)$

Hoeffding's bize ν ile μ'nün arasındaki farkın büyük olma olasılığının ne kadar küçük olduğunu söylüyor

Soru Use Hoeffding's identity to bound the prob of $\nu <= 0.1$ when $\mu = 0.9$ $(N = 10)$

$$|\nu - \mu| = |\nu - 0.9| >= 0.8$$

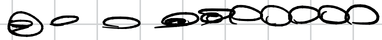
$$P[|\nu - \mu| >= 0.8] <= 2e^{-2(0.8)^2 \cdot 10}$$

$$<= 2e^{-12.8}$$

binomial ile $\approx 5.52 \times 10^{-6}$

Tanım

Sample seçtik:



$\otimes \ \underbrace{g(x)}_{\text{tahmin}} \neq \underbrace{f(x)}_{\text{gerçek}}$

$\bigcirc \ g(x) = f(x)$

$$\mu = \mathbb{E}\left[\mathbb{1}\left(g(x) \neq f(x)\right)\right] = E_{out}(g)$$ yani siyah olma olasılığı

g'nin outside

$$\nu = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(g(x) \neq f(x)\right) = E_{in}(g)$$ sample'da yaptığım hatalar

Hoeffding: $\forall \varepsilon > 0, \ P\left[\left|\underset{\text{error}}{E_{in}(g)} - E_{out}(g)\right| > \varepsilon\right] <= \underbrace{2e^{-2\varepsilon^2 N}}_{\delta}$

$$\frac{\delta}{2} = e^{-2\varepsilon^2 N} \Rightarrow \ln\left(\frac{2}{\delta}\right) = 2\varepsilon^2 N$$

$$\Rightarrow \varepsilon = \sqrt{\frac{1}{2N}\ln\left(\frac{2}{\delta}\right)}$$

$$\forall \varepsilon > 0, \ \mathbb{P}\left[\left|E_{in}(g) - E_{out}(g)\right| >= \sqrt{\frac{1}{2N}\ln\left(\frac{2}{\delta}\right)}\right] <= \delta$$

with probability $1-\delta$ we have:

$$\left|E_{in}(g) - E_{out}(g)\right| <= \sqrt{\frac{1}{2N}\ln\left(\frac{2}{\delta}\right)}$$

↓

$$E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N}\ln\left(\frac{2}{\delta}\right)}$$

and

$$E_{out}(g) \ge E_{in} - \sqrt{\frac{1}{2N}\ln\left(\frac{2}{\delta}\right)}$$

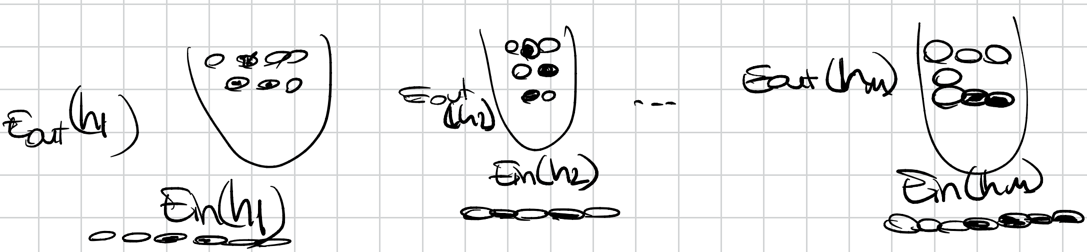
E_{in} → Seçtiğim sampledeki yanlış oranı

E_{out} → tüm kümedeki error

N ne kadar büyük seçersek fark azalacak.

Burada Duygu 1 fonksiyon getirdi, ne kadar kapsayıcı olduğunu hesapladık

Peki ya 100 fonks getirirse, "al birini seç, elekt çeksin" derse?



$g \in H = \{h_1, h_2, \ldots h_m\}$ ben buradan 1 g seçeceğim nasıl seçeceğim önemli değil

-g yerinde birini seçersen artık Hoeffding identity

hesabında kullanamıyoruz.

" $|E_{in}(g) - E_{out}(g)| >= \varepsilon$ "

P_0 olayı diyelim

$\Rightarrow$ (" $|E_{in}(h_1) - E_{out}(h_1)| >= \varepsilon$ or) P_1

(" $|E_{in}(h_2) - E_{out}(h_2)| >= \varepsilon$ or) P_2

⋮ or

(" $|E_{in}(h_m) - E_{out}(h_m)| >= \varepsilon$) P_m

$P_0 \overset{ise}{\Longrightarrow} P_1 \vee P_2 \vee P_3 \vee \ldots \vee P_m$ olma olasılığından fazladır

$$\mathbb{P}(P_0) <= \mathbb{P}(P_1 \vee P_2 \vee \ldots \vee P_m) <= \sum_{i=1}^{M} \mathbb{P}(P_i) <=$$

$$\text{Hoeffding} <= \sum_{i=1}^{M} 2e^{-2\varepsilon^2 N} = 2Me^{-2\varepsilon^2 N}$$

Yani: $\mathbb{P}\left[|E_{in}(g) - E_{out}(g)| \geq \varepsilon \right] \leq 2m\, e^{-2\varepsilon^2 N}$

Hyperparameter seçtin, diyelim gridsearch'te $m = 1000$
yani 1000 farklı hiperparametre denedimiz.

Sample size yeterince büyükse yine $2m$'i yer, yani

bulduğun fonks. iyi çalışır.

Mesela lineer | logistic regression'da bile sonsuz parametre var → $M = \infty$

Tanım: VC dimension tanımına gidiyoruz.

Önemli olan kaç farklı modelim olduğu değil, o modellerle kaç farklı etiketleme yapabildiğimiz

VC (Vapnik–Chervonenkis) – Dimension

$$h \in H, \quad x \in X, \quad h(x) \in \{-1, 1\}$$

Def: $x_1, x_2, \ldots x_N \in X$

bir hipo – kaç farklı şekilde ayırabilir / etiketleyebilir?

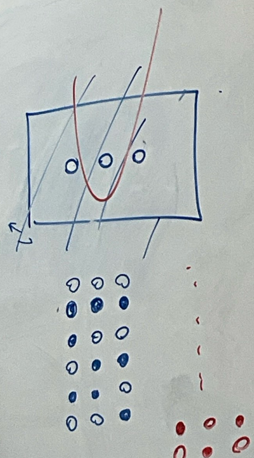
$$\Rightarrow H(x_1, x_2, \ldots, x_N) = \{(h(x_1), h(x_2), \ldots, h(x_N)) : h \in H\}$$

tek birinesi ver

$$\text{Def: } m_H(N) = \max_{x_1, x_2 \ldots x_N} |H(x_1, x_2, \ldots x_N)|$$

en iyi ihtimalle max kaç farklı etiket sunarım

-Dimension (Vapnik - C



$P_0 \Rightarrow P_1 \vee P_2 \vee \ldots \vee P_M$

$\mathbb{P}(P_0) \leq \mathbb{P}(P_1 \vee P_2 \vee \ldots$

H

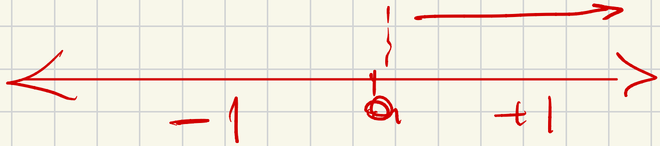
Yani

$\mathbb{P}\left[\left| E_{in}(g) - E_{o} \right. \right.$

$|A| = \#$ of elements in A

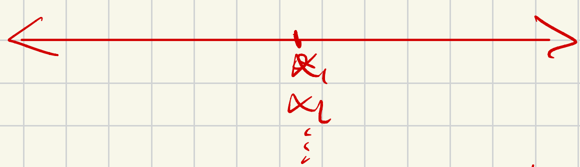
ex: H: positive rays, $x \in \mathbb{R}$

$h_a \in H$



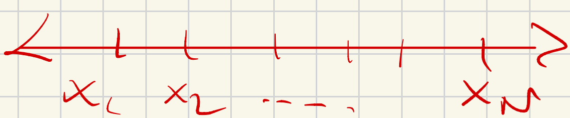
$$h_a(x) = \begin{cases} +1, & \text{if } x \geq a \\ -1, & \text{if } x < a \end{cases}$$

eğer hepsini aynı yerde seçersen:



sadece 2 farklı etiket ayrımı dır

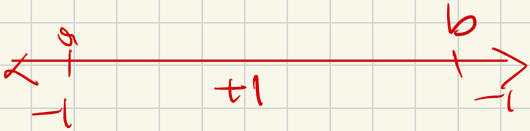
hepsini ayrık koyarsan:

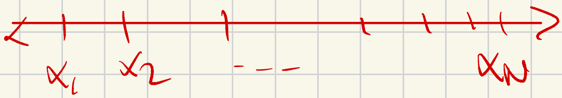


$m_H(N) = N+1$ → $N+1$ farklı şekilde etiketlebilirim ex:

H = positive intervals $\quad X = \mathbb{R}$

$$h_{a,b} \in H \Rightarrow h_{a,b}(x) = \begin{cases} +1, & x \in [a,b] \\ -1, & \text{otherwise} \end{cases}$$



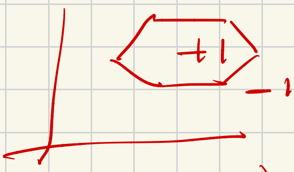


kaç farklı ayrım var?

$$m_H(N) = \binom{N+1}{2} + 1 = \frac{(N+1).N}{2} + 1$$

farklı şekilde etiketlenebilir.

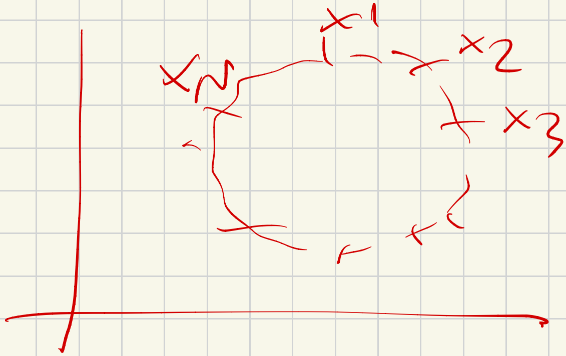
ex: H: positive convex set ; $X \in \mathbb{R}$



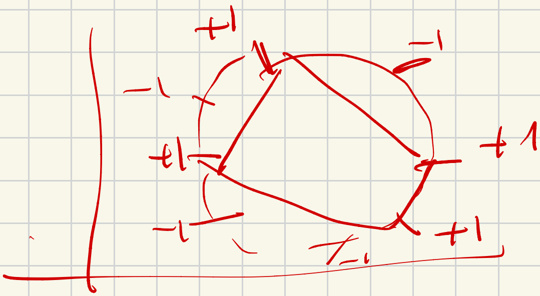
İçinde +1
dışında -1

$m_H(N) = ?$ max olacak değer 2^N

$<= 2^N$ her zaman 2^N'den küçük



mesela:



bu şekilde istediğim kadar farklı şekilde etiketleyebilirim. Yani bu her zaman 2^N

$\Rightarrow$

∴ çember üzerindeyse her zaman 2^N

Hoeffding'e göre dönersek içinden m tane

seçersek:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| >= \varepsilon] <= 2 m e^{-2\varepsilon^2 N}$$

bunun yerine $m_H(N)$ yazsak

mesela

$$2 . \underbrace{m_H(N)}_{} e^{-2\varepsilon^2 N}$$

2^N olursa çok büyük üssü yutar

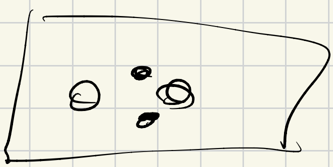
$N^2 + N + \ldots$ gibi bir polinom olsa, o zaman bunu bir polinomla sınırlasak, yeteri kadar veri olursa öğrenme gerçekleşir!

Definition: If $m_H(N) = 2^N$, then

H shatters N points
(parçalar)

2 ve 3 noktayı shatter edebiliyor ama

4'te sıkıntı:



bu şekilde ayırabilen bir doğru yok.

Definition : Smallest number k where

$m_H(k) < 2^k$ is called a

"breaking point" for H

$$\frac{1}{2}N^2 + \frac{1}{2}N + 1 = 2^N$$

→ eşitliğin kırıldığı ilk N

1'de oluyor

2'de oluyor

3'te olmuyor

→ positive ray → breaking point = 2

→ positive interval breaking point = 3

→ positive convex sets → breaking point = ∞

& hangi her türlü shatter ediyor

Theorem: If k is a breaking point for H ($m_H(k) < 2^k$), then $\forall N$

$$m_H(N) <= \sum_{i=0}^{k-1} \binom{N}{i}$$

bu fonksiyon N'ye bağlı bir polinom

Definition: If N is the biggest number such that $M_H(N) = 2^N$, then N is called the VC-Dimension of H.

$k \to$ breaking point $\Rightarrow k-1$: VC dimension

d_{vc} = VC-dimension $= k-1$

$$\Rightarrow \underbrace{M_H(N)}_{\text{growth function}} \leq \sum_{i=0}^{d_{vc}} \binom{N}{i} \to \text{bu bir polinom}$$

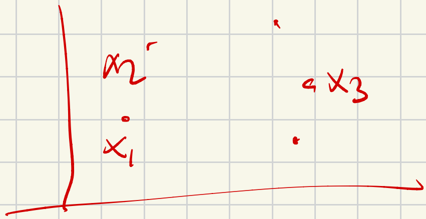
eğer bir breaking point varsa!

$$\mathbb{P}\left[\,|E_{in}(g) - E_{out}(g)| > \varepsilon\,\right] \leq 4\, m_H(2N) \cdot e^{-\frac{1}{8}\varepsilon^2 N}$$

" Dvc sonluysa & yeterince N varsa öğrenme sağlanır. "

Haftaya kanıtlayacağız!

Ex: H: lines with normals, $X \in \mathbb{R}^2$



doğrularla böldüğümde kaç farklı şekilde etiketleyebilirim?

$m_H(N) = ?$

yazdı ama çözmeden bıraktı.