بسم الله الرحمن الرحيم

الدكتور هيعمل مراجعه سريعه كدا علي الليرنجن ثيوري ...

Q&A

Q: Why do we shuffle the examples in SGD?

A: This is how we do sampling *without* replacement

1. **Theoretically** we can show sampling **without replacement** is not significantly worse than sampling with replacement (Shamir, 2016)
2. **Practically** sampling without replacement tends to work better

Q: What is "bias"?

A: That depends. The word "bias" shows up all over machine learning! Watch out...

1. The additive term in a linear model (i.e. b in $w^T x + b$)
2. Inductive bias is the principle by which a learning algorithm generalizes to unseen examples
3. Bias of a model in a societal sense may refer to racial, socio-economic, gender biases that exist in the predictions of your model
4. The difference between the expected predictions of your model and the ground truth (as in "bias-variance tradeoff")

2

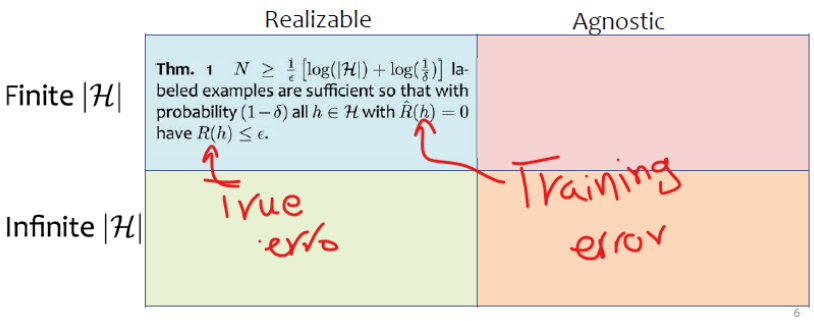
طيب بسم الله .. احنا عاوزين نفكر في السامبل كومبلكستي .. في آخر مره احنا وصلنا لأول نظريه

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	
Infinite $\lvert\mathcal{H}\rvert$		



6

هنا انت عندك upper bound علي ال true error اللي انت متعرفهوش بناءً علي ال training error الي احنا بنبقا عارفينو ... طيب الدكتور حاطط مثال وعاوزنا نفكر في النظريه كإنها مثال .. والمثال هنا هو هيبقا hypotheses class of conjunctions .. فاحنا هنقول ان H is a class of conjunctions over vectors X .. والفيكتورز إكس دول هيبقا اصصفار ووحايد ..

Example: Conjunctions

Question:
Suppose H = class of conjunctions over **x** in $\{0,1\}^M$

Example hypotheses:
$h(\mathbf{x}) = x_1 (1-x_3) x_5$
$h(\mathbf{x}) = x_1 (1-x_2) x_4 (1-x_5)$

If $M = 10$, $\varepsilon = 0.1$, $\delta = 0.01$, how many examples suffice according to Theorem 1?

Answer:
A. $10*(2*\ln(10)+\ln(100)) \approx 92$
B. $10*(3*\ln(10)+\ln(100)) \approx 116$
C. $10*(10*\ln(2)+\ln(100)) \approx 116$
D. $10*(10*\ln(3)+\ln(100)) \approx 156$
E. ~~$100*(2*\ln(10)+\ln(10)) \approx 691$~~
F. $100*(3*\ln(10)+\ln(10)) \approx 922$
G. $100*(10*\ln(2)+\ln(10)) \approx 924$
H. $100*(10*\ln(3)+\ln(10)) \approx 1329$

$|H| = 3^M$

طيب الدكتور بيقول في المثال ده اننا عاوزين نجيب عدد الامثله اللي هتبقا كفايه عشان تحقق theorem 1 .. الدكتور اتكلم علي ان المشكله كلها في حساب ال size of H .. اللي هو في النظريه عندك |H| المهم هنا الدكتور قال هو الحجم اصلا عباره عن ايه ... عباره عن all the possible conjuction function you can come up with ... الدكتور كمل كلامه وقال ان في 3 اوبشنز لكل فاريبل من ال M variables اللي عندك .. كل فاريبل هنا هيبقا ليه 3 حالات . يا إما included او included but negated أو excluded all together من .. x1, x2, x3, ..., x10 .. كل واحد ليه 3 اختيارات زي دول .. فبالتالي الحجم بتاع ال hypothesis space هيبقا 3^10 ... فبالتالي الاجابه الصح هتبقا D ... طيب الدكتور بيقول ده معناه ايه المثال ده .. معناه انك لو عندك 156 examples فبالتالي 156 according to theorem 1 .. ده هيبقا كفايه عشان باحتمالية 99% كل ال hypotheses اللي هو all conjunction functions هيبقا ليهم صفر تريننج ايرور علي الداتا سيت بتاعتك اللي هي 156 مثال ... وهيبقا ليهم true error اقل من 10% .... الحاجه الحلوه هنا ان علي ارض الواقع احنا مش بنستخدم النظريه ديه عشان نحدد عدد الأمثله اللي هنحتاجها عشان نجيب البرفورمنس اللي احنا عاوزينو .. هي بس مجرد بتديك sense كدا انك تقول انت محتاج كام مثال يعني انت ممكن تحتاجهم عشان تجيب برفورمنس كويس تحت شوية assumptions قلناهم المره اللي فاتت تحت ال PAC model ...

طيب الحاجه التانيه اللي هنقول عليها هي ال Realizable vs Agnostic case .. ال Relalizable setting هنا انت عندك ال c star جوا ال hypothesis space ... انما ال Agnostic setting ... هي ان برضو ال c star in the hypotheses space أو انها مش في ال hypotheses space احنا مش متأكدين ... " عشان نبقا ناس كوريكتور" ..

طيب السؤال هنا عشان نخلي نظريه رقم 1 ديه ليها قيمه يعني احنا فعلا بنحتاج نجيب قيم الإبسيلون و قيم الدلتا ... هنا انت بتختارهم بإيدك ... طيب تعال نرجع للفرق ما بين ال Realizable vs Agnostic ... طيب اللي عاوزين نقولو هنا ان ال Agnostic case بتيجي لما انت بتفكر في ان أياً كان الفانكشن اللي عملت جينيريشن للداتا .. هي مش معاك في الهيبوثيسيز سبيس ... فال finite H هنا احنا لو فكرنا مثلا زي في ال class of decision trees ... مثلا لو عندك ماكس ديبث 10 .. ليه clear finite H تقدر تبص علي كل الهيبوثيزيز اللي عندك ... بس فعلياً كل الداتا سيتس اللي هتعمل تريننج عليها .. هتبقا في ال Agnoistic ليه .. عشان اللي عمل ال labeling هو انسان .. مش شجره .. "انا شجره" .. فاللي بنقولو في ال Agnostic .. في السلايد الجايه

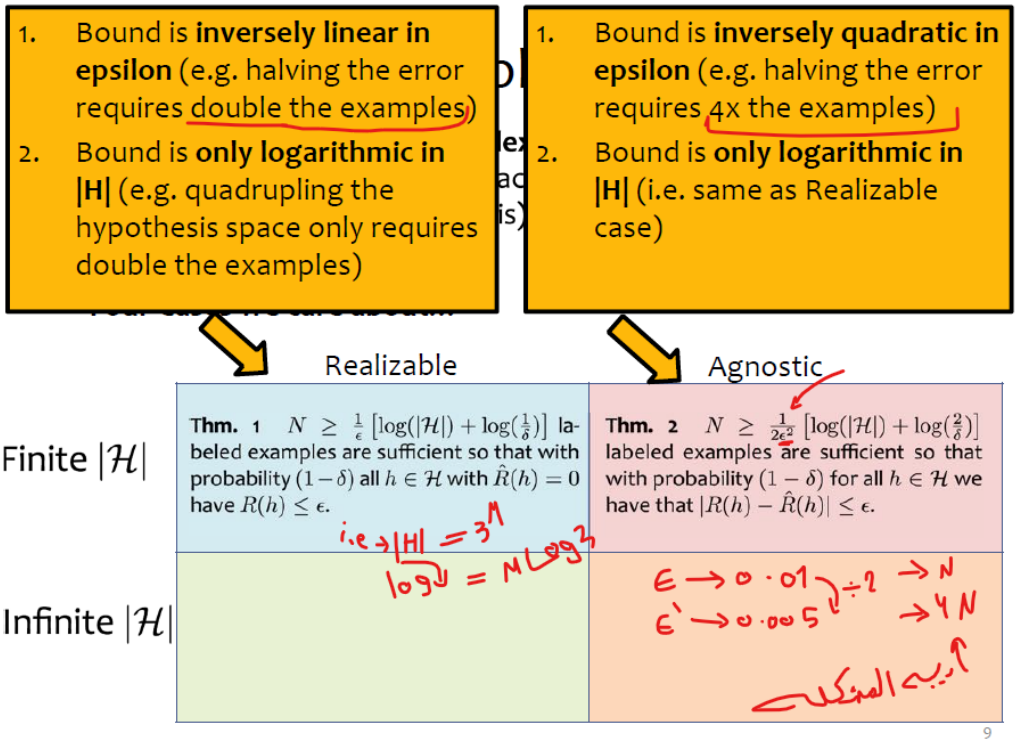
Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable ($c^* \in \mathcal{H}$)	Agnostic ($c^* \in \mathcal{H}$ or $c^* \notin \mathcal{H}$)
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$. (True, Train)
Infinite $\lvert\mathcal{H}\rvert$		

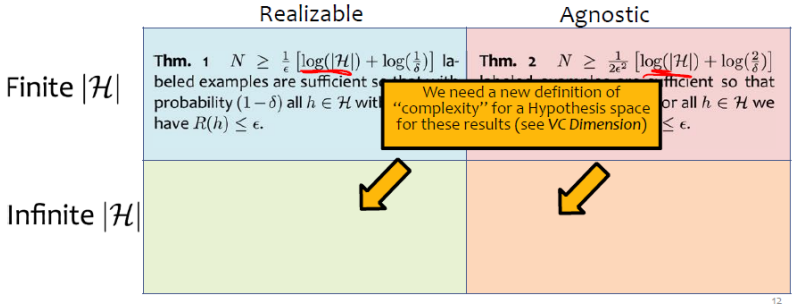
طيب اللي عاوزين نلاحظ هنا هو الاختلاف بتاع ال utility بتاع ال 2 باوندز اللي عندنا ...



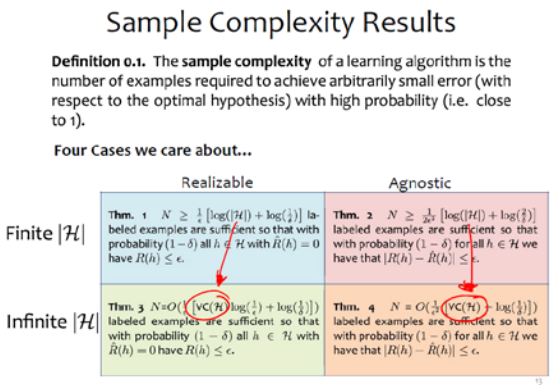
Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

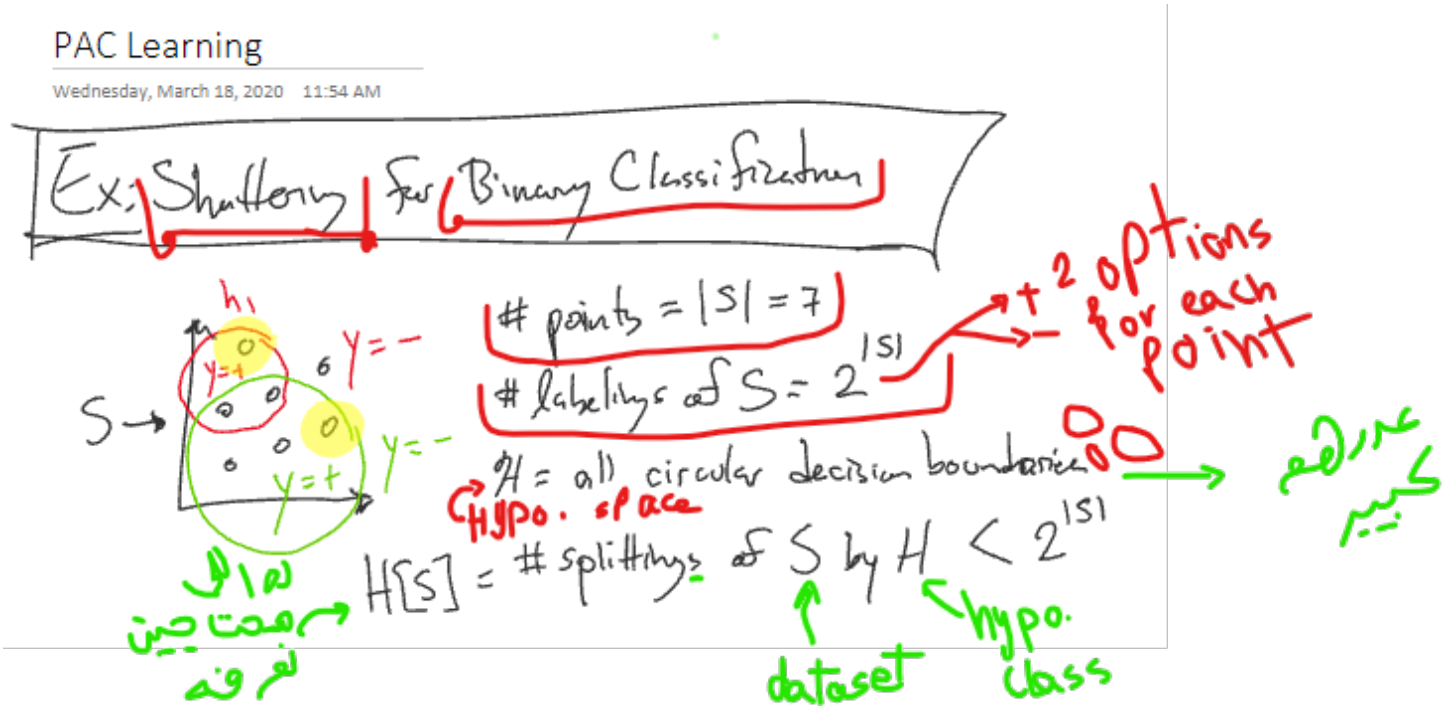
Four Cases we care about...



طيب لو عندنا linear decision boundary عندنا انفنت نامبر اوف هايبوثيزيز ..



طيب ايه هو ال VC dimension ..



الدكتور بيقول ان في labelings .. السيركيولار ديسيجن باوندري مبيقدرش ي replicate .. يعني مثلا لو خدنا النقط اللي متعلم عليها بالاصفر فوق .. لو الاتنين كانو +ve label .. انت معندكش circular decision boundary اللي بيقدر ي label the dataset بالطريقه ديه .. في الحاله ديه عدد ال splitting هيبقا اكيد اكيد اكيد اقل من 2 ^s .. وده اللي احنا وصلنالو في السلايدز ..

لو جينا نفكر في التعريف ده بيقول ايه .. هو بيقول ان ال hypotheses space بي shatter S لو عدد ال splitting المتاح ليك هو بيساوي 2^s ... فال circular decision boundary للنقط الصفرا مقدرش انو ي shatter that particular dataset S ..

طيب هنعرف ايه هو ال VC-dimension .. ال VC-dimension بتاع ال hypotheses space هو حجم أكبر dataset S تقدر تقسّمها باستخدام ال H .. السؤال هنا ازاي نجيب VC-dimension بتاع some particular hypotheses space ..

Shattering, VC-dimension

Definition:

H[S] - the set of splittings of dataset S using concepts from H.

H shatters S if $|H[S]| = 2^{|S|}$.

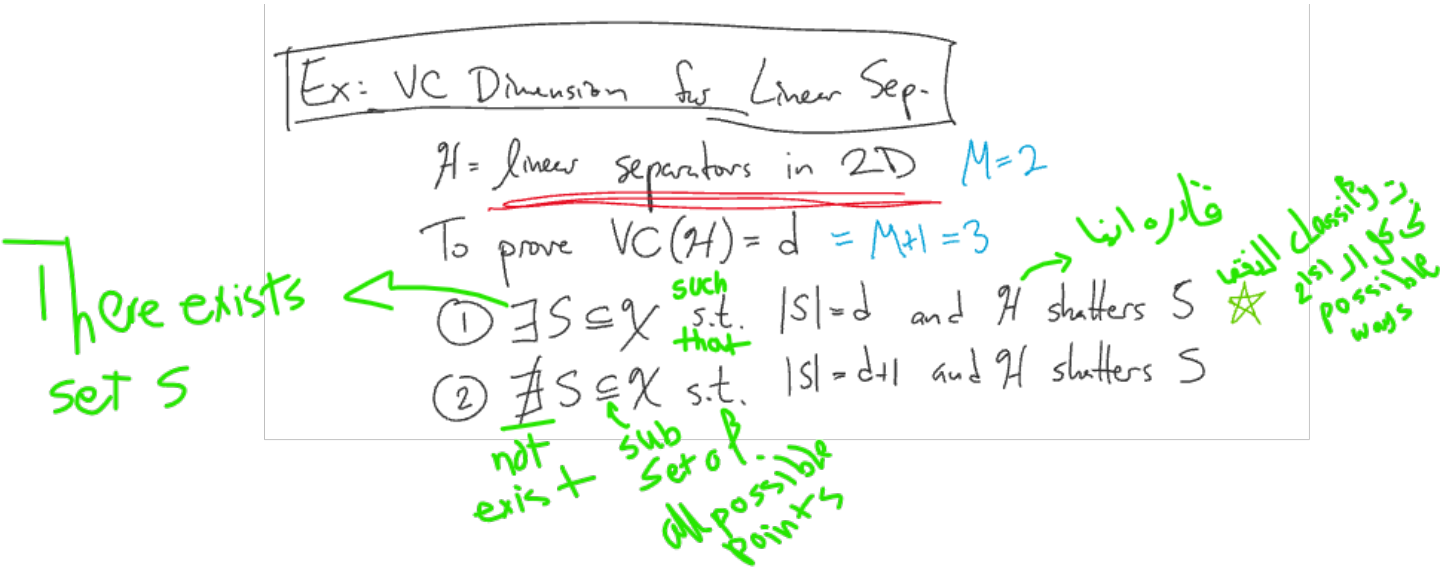
A set of points S is shattered by H is there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in S are achievable using concepts in H.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

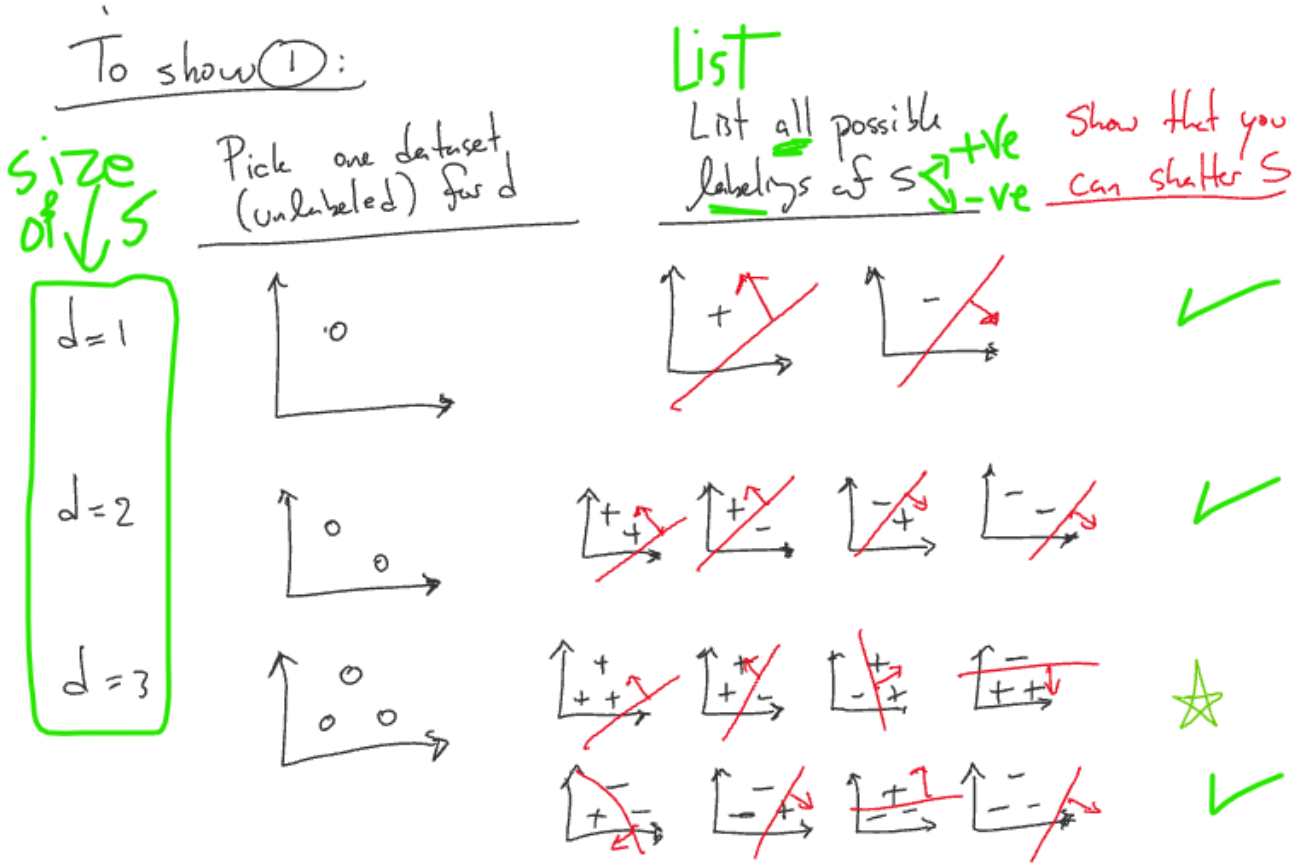
The VC-dimension of a hypothesis space H is the cardinality of the largest set S that can be shattered by H.

If arbitrarily large finite sets can be shattered by H, then $\text{VCdim}(H) = \infty$

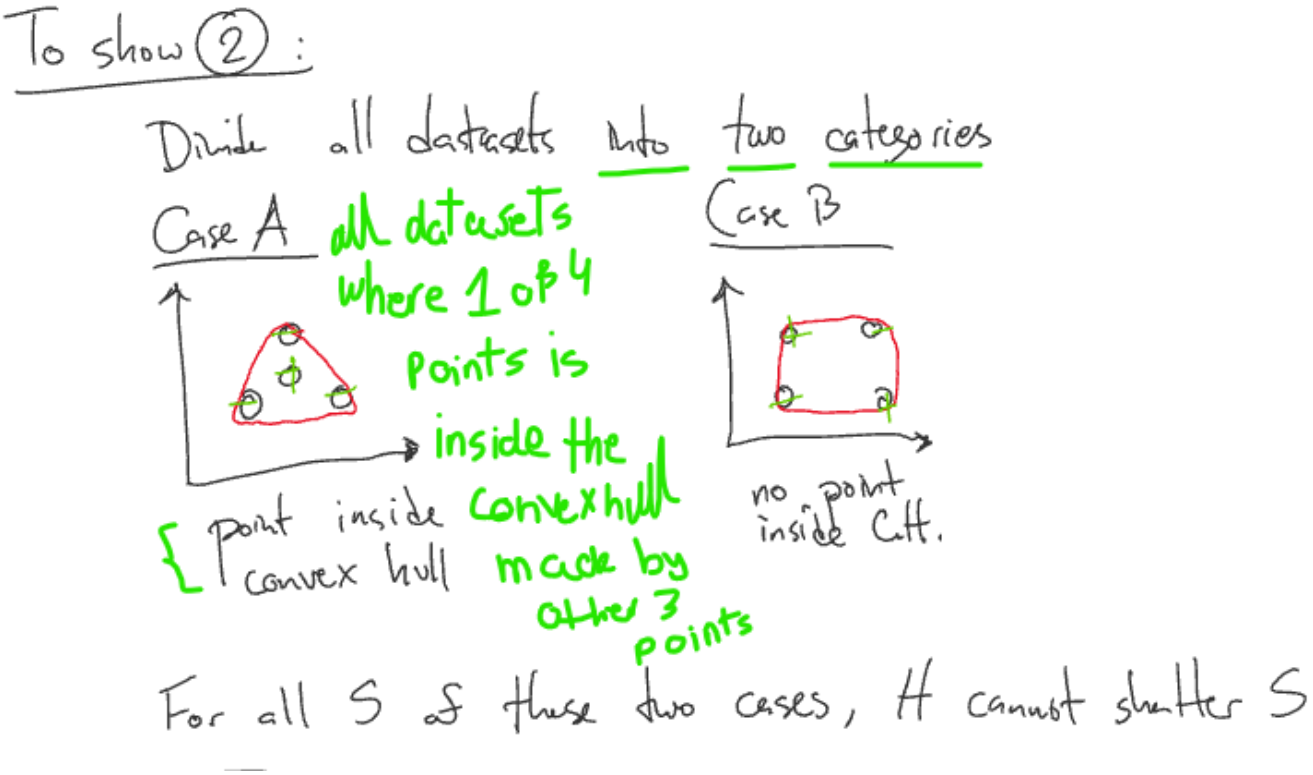
Slide from Nina Balcan



فعشان تثبت ال VC-Dimension اللي انت محتاج تقول انو موجود هو example dataset ال H تقدر ت shatter .. وديه أكبر داتا سيت ال H can shatter .. "انا مش فاهم " ... أول سؤال ازاي نثبت رقم 1 في الصوره اللي فوق:

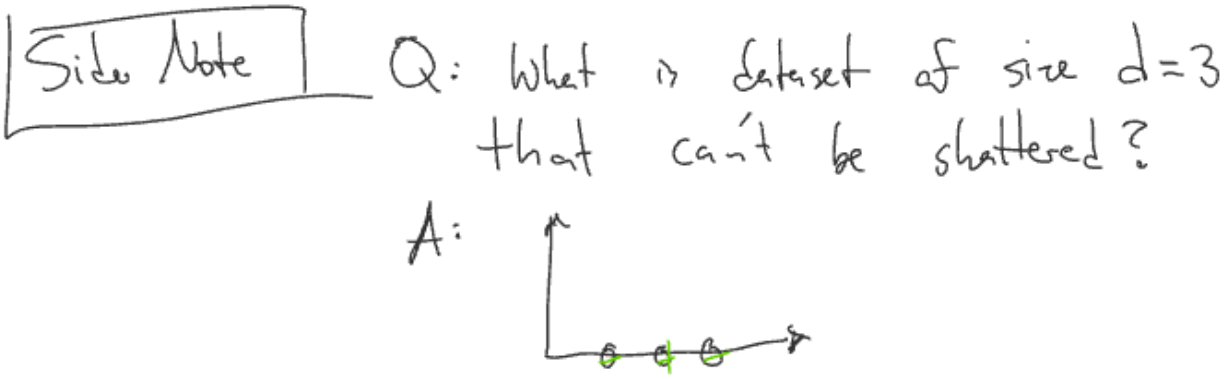


اللي عاوزين نعملو ناو هو نثبت رقم 2:



اللي بنقولو هنا هو ان لو عندنا مثلا في حالة A .. لو انت تخيلت ان عندك داتا سيت من الفورم اللي بالشكل ده .. و البوينتس كانت ان اللي في النص بس هو الموجب و كل اللي الكورنرز كانو سالب ... في الحاله ديه مفيش linear decision boundary بيقدر ي correctly classify that .. ومش مهم ال plus point ديه فين جوا ال convex hull او ايه هو الكونفجريشن بتاع ال 3 نقط التانيين .. فأي دااات سيت هتجيبهالي مش هيبقا في linear decision boundary ..

الدكتور هيحط ساید نوت عباره عن سؤال ..



طيب اللي عاوزين نوضحو ان في فرق ما بين ال distinction between there exists and for all .. في ال VC-dimension .. لما بنيجي نثبت ال VC-Deimension بيحتاجنا نثبت ان في دااتا سيت حجمها d تقدر ت شاتر و كمان انك تثبت ان مفيش داتا سيت حجمها d+1 منقدرش نشاتر .. الدكتور بس حب يوضح السلايد الجايه ... ان ال VC بتعتمد علي exists انما ال shattering بيعتمد علي ال forall ..

∃ vs. ∀

VCDim

- Proving **VC Dimension** requires us to show that **there exists** (∃) a dataset of size d that can be shattered and that **there does not exist** (∄) a dataset of size d+1 that can be shattered

Shattering

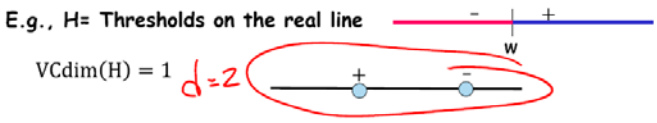
- Proving that a particular dataset can be **shattered** requires us to show that **for all** (∀) labelings of the dataset, our hypothesis class contains a hypothesis that can correctly classify it

طيب في امثله تانيه:

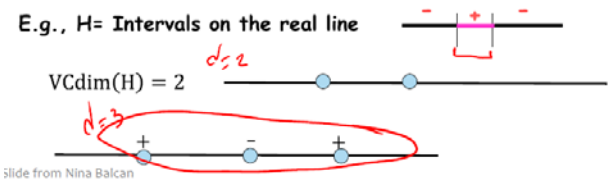
Shattering, VC-dimension

If the VC-dimension is d, that means there exists a set of d points that can be shattered, but there is no set of d+1 points that can be shattered.

E.g., H= Thresholds on the real line



E.g., H= Intervals on the real line



slide from Nina Balcan

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.
Infinite $\lvert\mathcal{H}\rvert$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\lvert R(h) - \hat{R}(h)\rvert \leq \epsilon$.

29

طيب الدكتور بيقول احنا لحد دلوقت بنتكلم علي عدد الأمثله اللي هنحتاجها .. دلوقت هنخش في اتجاه مختلف شويه هو ال slt اللي هو statistical learning theory ... فالدكتور في السلايد الجايه ديه كان بيشرح ازاي نروح من نظريه رقم 1 ل corollary 1 ... انك هتحل للإبسيلون .. ده بيحط upper bound علي التروو إيرور ل fixed number of examples ...

SLT-style Corollaries

Thm. $\frac{1}{\epsilon}\left[\log(|\mathcal{H}|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

Solve the inequality in Thm.1 for epsilon to obtain Corollary 1

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

We can obtain similar corollaries for each of the theorems...

30

SLT-style Corollaries

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)\right]$$

Corollary 2 (Agnostic, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N}\left[\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right]}$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O\left(\frac{1}{N}\left[\mathrm{VC}(\mathcal{H})\ln\left(\frac{N}{\mathrm{VC}(\mathcal{H})}\right) + \ln\left(\frac{1}{\delta}\right)\right]\right) \quad (1)$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1-\delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\mathrm{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \quad (2)$$

طيب السؤال المهم .. ازاي هنعمل model selection:

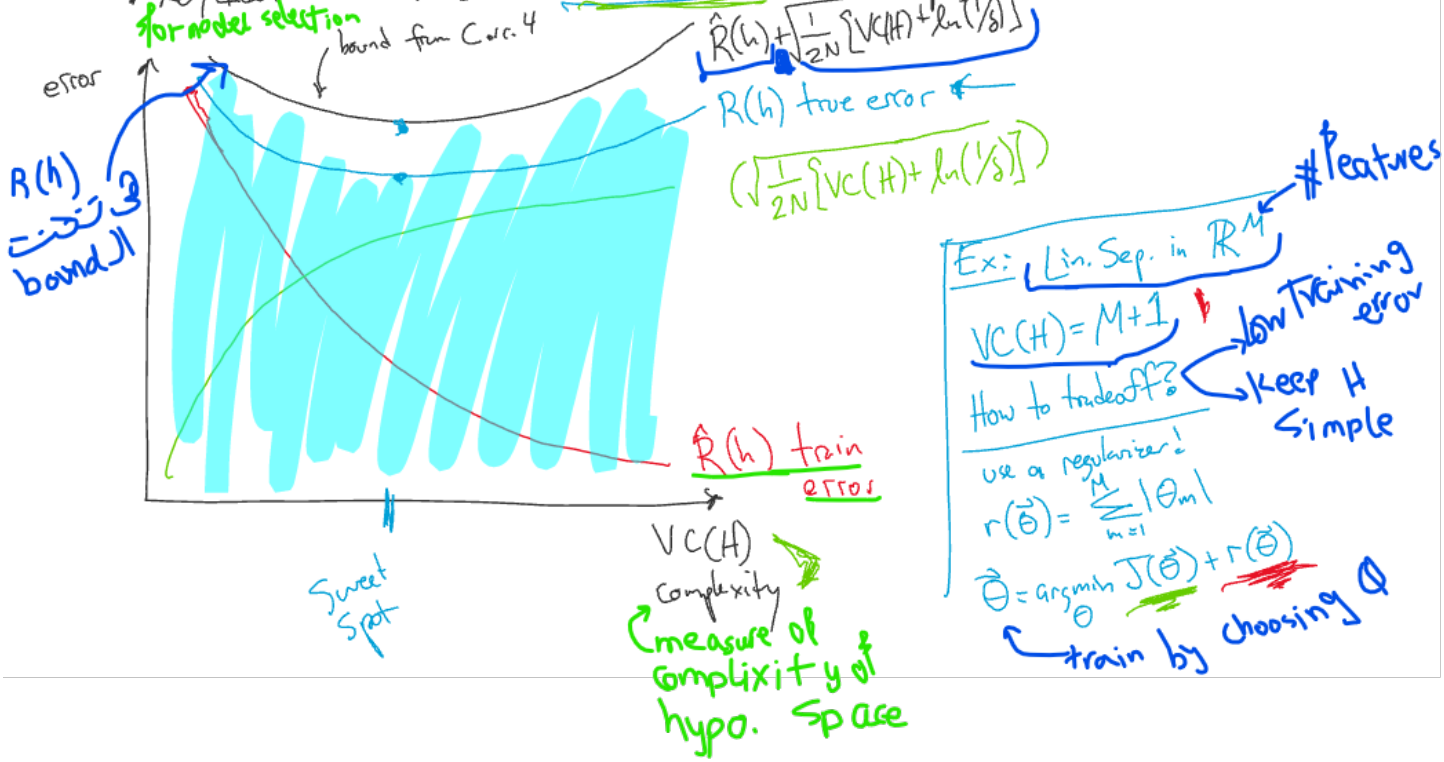
الموديل سيليكيشن هو انت عاوز تفكر في empirical risk minimization ..

Model Selection

Q. Is Corr. 4 useful? A: Yes!

Key Idea: tradeoff between low train error and keeping H simple (low VC Dim)

for model selection



Ex: Lin. Sep. in $\mathbb{R}^4$ — # features

$VC(H) = M+1$

How to tradeoff? → low Training error; Keep H Simple

use a regularizer!

$r(\vec{\theta}) = \sum_{m=1}^{M} |\theta_m|$

$\vec{\theta} = \arg\min_{\theta} J(\vec{\theta}) + r(\vec{\theta})$

↳ train by choosing θ

Recipe for Machine Learning

1. Given data $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$
2. a) Choose a decision function $h_\theta(\vec{x}) = \cdots$ parameterized by $\vec{\theta}$
   b) Choose an objective function $J_D(\vec{\theta}) = \cdots$ relies on data
3. Learn by choosing parameters that optimize the objective $J(\theta)$

$$\hat{\theta} \approx \arg\min_{\vec{\theta}} J_D(\vec{\theta})$$

4. Predict on new test example $\vec{x}_{new}$ using h

$$\hat{y} = h_\theta(\vec{x}^{(new)})$$

Decision Functions

- Perceptron: $h_\theta(\vec{x}) = \text{sign}(\vec{\theta}^T \vec{x})$
- Linear Regression: $h_\theta(\vec{x}) = \vec{\theta}^T \vec{x}$
- Discriminative Models: $h_\theta(\vec{x}) = \underset{y}{\text{argmax}}\ p_\theta(y|\vec{x})$
  - Log. Reg. $p(y=1|\vec{x}) = \sigma(\vec{\theta}^T \vec{x})$
- Generative Models: $h_\theta(\vec{x}) = \underset{y}{\text{argmax}}\ p_\theta(\vec{x}, y)$ → joint distr.
  - Naive Bayes $p(\vec{x}, y) = p(y) \prod_{m=1}^{M} p(x_m|y)$
- NN for classification: $p(y|\vec{x}) = \sigma(W^{(2)T} \sigma(W^{(1)T} \vec{x} + b^{(1)}) + b^{(2)})$

Obj. Function

1. MLE: $J(\theta) = -\sum_{i=1}^{N} \log p(x^{(i)}, y^{(i)})$
2. MCLE: $J(\theta) = -\sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)})$ ← conditional
3. L2 Reg: $J'(\theta) = J(\theta) + \lambda \|\theta\|_2^2$
4. L1 Reg: " " $+ \lambda \|\theta\|_1$

Opt Method

- Gradient Descent $\quad \theta \leftarrow \theta - \gamma \nabla J(\theta)$
- SGD where $J(\theta) = \frac{1}{N} \sum_{i=1}^{N} J^{(i)}(\theta)$ $\quad \theta \leftarrow \theta - \gamma \nabla J^{(i)}(\theta)$ for $i \sim \text{Uniform}(\{1, \ldots, N\})$
- Mini-batch SGD
- Closed form → Partial
  1. compute partial derivatives
  2. set to zero and solve