Foundations for Learning in the Age of Big Data

Maria-Florina (Nina) Balcan Carnegie Mellon University

Brief Overview

ML: automatic methods for extracting info from data and for learning to make accurate predictions or useful decisions based on past observations and experience.

Amazing impact on many application areas. E.g.,

- Document categorization, Natural Language Processing
- Image Classification, Speech Recognition
- Fraud Detection, Spam Detection
- Computational biology.

This lecture series focus: theoretical foundations for learning in the age of big data.

Brief Overview

This lecture series: theoretical foundations for learning in the age of big data.

Lectures 1 & 2: Foundations of classic ML.

• Generalization and Overfitting.

Lectures 3: Active Learning.

Lectures 4: Distributed Learning.

Today's topic: Generalization and Overfitting in Machine Learning

Focus on sample complexity for supervised classification

- Statistical Learning Theory (Vapnik)
- PAC (Valiant)

Supervised Classification

from data to discrete classes

Supervised Learning

• E.g., classify objects as chairs vs non chairs.



• E.g., which emails are spam and which are important.

We are looking forward to your upcoming lecture at our summer school.

Not spam

Click here to get out of debt fast. No charge or tricks, we promise!

spam

Supervised Classification. Example: Spam Detection

spam

Decide which emails are spam and which are important.

Supervised classification

e <u>E</u> dit <u>V</u> iew <u>G</u> o <u>M</u> essage <u>T</u>	ools <u>H</u> elp	(1991) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997) (1997)
Get Mail 🔹 📝 Write 🔲 Addre	ss Book 🔊 Tag • 🔎 Search all messages	×
thesis		
Smart Folders	🚡 🏠 🥔 Subject 🔍 From	실 Date 🔺 🗈
regulations A	😭 🛛 🖑 Congrats on the dissertation award! 🔹 Michelle Leah Goodstein	• 10/12/2009 9:56 🗹
reimbursemrent	SCS Dissertation Award & ACM Dissertatio Randy Bryant	 10/14/2009 10:0
repeated-eq	Re: Congrats on the dissertation award! • Maria Florina Balcan	 10/15/2009 9:34
review	😭 🛛 🐳 Re: SCS Dissertation Award & ACM Dissert 🔹 Doru-Cristian Balcan	 10/15/2009 12:4
seminars-natech	A Day SCS Discontation Award 9: ACM Discont	• 10/15/2000 12/5
cindofrii		
AB 0.0.0 10	Arepty 1 Parts of the Parts of	archive II junk II 👗 delere.
students	from Randy Bryant <randy.bryant@cs.cmu *="" *<="" @reply="" and="" service="" td=""><td>archive Junk Adelete</td></randy.bryant@cs.cmu>	archive Junk Adelete
students students-diverse	from Randy Bryant <randy.bryant@cs.cmu &="" *="" @rebly="" acm="" all="" award="" corrward="" dissertation="" nominees<="" safebiy="" scs="" subject="" td=""><td>10/14/2009 10:00 PM</td></randy.bryant@cs.cmu>	10/14/2009 10:00 PM
students students-diverse	from Randy Bryant <randy.bryant@cs.cmu (<br="" (@rephy)="" (@roward)="" and="">subject SCS Dissertation Award & ACM Dissertation Nominees to You ()</randy.bryant@cs.cmu>	10/14/2009 10:00 PM
students students-diverse talks-accross-gatech	from Randy Bryant «Randy, Bryant®cs.cmu (அमित्रिष्ठ) 1 (agreent) at subject SCS Dissertation Award & ACM Dissertation Nominees to You Ω cc Catherine Copetas «copetas@cs.cmu.edu» Ω	10/14/2009 10:00 PM
students students-diverse talks-across-gatech talks-campus talks-caatech	from Randy Bryant «Randy, Bryant®es.cmu (@rFBW)) @rFBW a) @rFBW add subject SCS Dissertation Award & ACM Dissertation Nominees to You cc Catherine Copetas <copetas@cs.cmu.edu></copetas@cs.cmu.edu>	archive Junk delete 10/14/2009 10:00 PM other actions *
talk-outside	from Randy Blyset -Kanady, Blyset&Cs.cm (@PFBW) [@PFBW] * @PTOWerd] subject SCS Dissertation Award & ACM Dissertation Nominees to You'] cc Catherine Copetas <copetas@cs.cmu.edu> Nina:</copetas@cs.cmu.edu>	archive Junk delete 10/14/2009 10:00 PM other actions •
students students-diverse talks-accross-gatech talks-accross-gatech talks-gatech talks-gatech talks-outside tealks-untside	from Randy Byyant «Randy, Byyant@cs.cmu (@PFBW) [improved] subject SCS Dissertation Award & ACM Dissertation Nominees to You' cc Catherine Copetas <copetas@cs.cmu.edu>① Nina: You micht have already seen this announcement, but I wou</copetas@cs.cmu.edu>	archive Junk Adeete 10/14/2009 10:00 PM other actions -
students students-diverse talks-across-gatech talks-carpus talks-outside talks-outside talks-outside talks-outside	from Randy Bryant <randy, 31="" 33="" 34="" @wf0py="" @wf0py<="" @wf6py=""]="" bryant@cs.cmu,="" td=""><td>archive Junk A delete 10/14/2009 10:00 PM other actions -</td></randy,>	archive Junk A delete 10/14/2009 10:00 PM other actions -
students students talk-accross-gatech talk-accross-gatech talk-compus talk-control talk-coutide teaching teaching teaching	from Randy Byyart «Randy, Byyart Besc.mu, LawrBoy) [LawrBoy a) [Auroward] subject SCS Dissertation Award & ACM Dissertation Nominees to YouT ac Catherine Copetas «copetas@cs.mu.edu> Nina: You might have already seen this announcement, but I wou personally congratulate you for your outstanding dissert like to invite you to return to CMU to give a distinguis	archive Jurk & detect 10/14/2009 10:00 PM other actions - action. I would thed lecture
astudents students talks-across-gatech talks-compus talks-gatech talks-outside teaks-outside teaks-outside teak-report tehrory-talks	<pre>from Randy Bryant <randy, &="" '="" <copetas@cs.cmu.edu="" @rforward'="" acm="" al="" award="" bryant="" c="" cartery="" catherine="" copetas="" dissertation="" nominees="" res.cmu,="" scd="" subject="" to="" you'=""> Nina: You might have already seen this announcement, but I wou personally congratulate you for your outstanding dissert like to invite you to return to CRU to give a distinguis sometime in the winter of 2010. Catherine Copetas will </randy,></pre>	archive jurk & detect 10/14/2009 10:00 PM other actions - atid like to action. I would whed lecture work out the
students students-diverse tables-conso-gatech tables-campus tables-campus tables-campus tables-cateching tables-gateching tab	from Randy Byyart «Randy Byyart @cs.mul [@rBdW] [@rBdW 3] [@rBdW 3] subject SCS Dissertation Award & ACM Dissertation Nominees to You() @ Catherine Copetas <copetas@cs.cmu.edu>() Nina: You might have already seen this announcement, but I wou personally congratulate you for your outstanding dissect like to invite you to return to CHU to give a distinguis sometime in the winter of 2010. Catherine Copetas will timing for you. You'l get to use the new Rashid Audico</copetas@cs.cmu.edu>	archive Jurk detects 10/14/2009 10:00 PM other actions - action. I would bed lecture work out the primer-a big
students students-diverse talks-accose-gatech talks-outade talks-patech talks-patech talks-patech talk-n-part tech-report theory-group theory-talks	<pre>from Randy Bryant <randy, @rf6wy]="" [="" [<="" [@rf6wy]="" bryant="" res.cm,="" td=""><td>archive jurk; A detect 10/14/2009 10:00 PM other actions = ald like to action. I would hed lecture work out the work out the oriuma big</td></randy,></pre>	archive jurk; A detect 10/14/2009 10:00 PM other actions = ald like to action. I would hed lecture work out the work out the oriuma big
students students students tales-accose-patech tales-patech tales-patech tech-report tech-report tech-report theory-goup tech-report theory-tales total-diverse-patech	from Randy Bryant «Randy Bryant@cs.cmu, @wf69y) [@wf69y 3] * @wf69y 3] * @wf69y 3] * @wf69y 3] * @wf69y 43]	ardney Jurk & Geeke 10/14/2009 10:00 PM other actions - and 1 like to action. I would thed lecture work out the riuma big
tudents students-diverse tals-scampus tals-gatech tals-gatech tals-gatech tals-gatech tech-report theory-talls torg	<pre>from Randy Byyart <randy, @cs.mul="" @rbdw="" [="" [@rbdw]="" a]="" a]<br="" byyart="">subject SCS Dissertation Award & ACM Dissertation Nominees to You[] cc Catherine Copetas <copetas@cs.cmu.edu>D Nina: You might have already seen this announcement, but I wou personally congratulate you for your outstanding dissert like to invite you to return to CHU to give a distinguis sometime in the winter of 2010. Catherine Copetas will timing for you. You'l get to use the new Rashid Audito improvement over Wean 7500. Best of wishes to you at Georgia Tech.</copetas@cs.cmu.edu></randy,></pre>	addive jurk & deeke 10/14/2009 10:00 PM other actions - ald like to ination. I would hed lecture work out the riuma big

Not spam

🛎 SPAM for dbalcan@cs.cmu.edu - Thunderbird															
Ele	<u>E</u> dit	View	Go	Mess	age]	Tools	Help							$\langle \rangle$	
Get M	ail	Write	Add	iress B	ook	Reply	Reply All	Forward	🔊 - Tag	X Delete	Not Junk	S Print	- G	•	
All Folders					210	Sub	ject			63	Sender	6	Date 🔹	E.	
•	Mail	Mail Scoala F				student loan debt •						nrie 💧	1/28/20	^	
ė.	🥏 Int	ox			debt consolodation • Emanuel Cor							r 💧	M 1/28/20		
- / Drafts					V	erified You Or	rdered Meds		۰	Tamika Thor	pe 💧	1/28/20			
					G	iive that girl t	he best goft f	or Valentine's		Clare	0	1/28/20			
A SPAM (42)						h	ad cradit dab	t concolidation	n lann	-	Topin Laird	A 1	1 /10 /10		
- 1 Trash Acasa Colegi colegi colegi					Thunderbird thinks this message is junk. This is Not Junk										
					 Subject: debt consolodation 										
					From: Emanuel Cortez <wconvention@heilhecker.de></wconvention@heilhecker.de>										
					Date: 1/28/2008 4:14 PM										
€ 🔜 JK ⊡ 🔜 jobs						Т	o: <u>dayne@c</u>	s.cmu.edu							
					Sav	good	d bye to	debt.							
🚮 Junk 🚮 mike 🛄 nina				Acceptable Unsecured Debt includes All Major Credit Cards.											
					No-collateral Bank Loans, Personal Loans, Medical Bills etc.										
🗈 🔜 Profesori 🗈 🔜 scoala															
					http://www.baddebth.cn										
	۵ 🛄	submiss	ions												
e 🧔	Local	Folder	s												
0											Unread:	42 1	Fotal: 2665		

Goal: use emails seen so far to produce good prediction rule for future data.

Supervised Classification. Example: Spam Detection

Represent each message by features. (e.g., keywords, spelling, etc.)

	"monov"	"pille"	"NAr "	had spolling	known sondor	spam?	
_	money	pins			KIIOWII-Selluel	spanns	-
	Y	N	Y	Y	N	Y	
	Ν	N	N	Y	Y	N	
	N	Y	Ν	N	N	Y	
exam	ple 📉	Ν	Ν	Ν	Y	Ν	label
	Ν	Ν	Y	Ν	Y	N	
	Y	Ν	Ν	Y	Ν	Y	
	Ν	Ν	Y	Ν	Ν	N	
\mathbf{i}						1	

Reasonable RULES:

Predict SPAM if unknown AND (money OR pills)

Predict SPAM if 2money + 3pills -5 known > 0



Linearly separable

Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

• E.g.: Adaboost, logistic regression, SVM, etc.

Confidence Bounds, Generalization

(Labeled) Data

Confidence for rule effectiveness on future data.





- Algo sees training sample S: $(x_1, c^*(x_1)), ..., (x_m, c^*(x_m)), x_i$ independently and identically distributed (i.i.d.) from D; labeled by c^*
- Does optimization over S, finds hypothesis h (e.g., a linear separator).
- Goal: h has small error over D.

- X feature or instance space; distribution D over X e.g., X = R^d or X = $\{0,1\}^d$
- Algo sees training sample S: $(x_1, c^*(x_1)), \dots, (x_m, c^*(x_m)), x_i$ i.i.d. from D
 - labeled examples assumed to be drawn i.i.d. from some distr.
 D over X and labeled by some target concept c*
 - labels \in {-1,1} binary classification
 - Algo does optimization over S, find hypothesis h.
 - Goal: h has small error over D.

 $err_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$



Need a bias: no free lunch.

- X feature or instance space; distribution D over X e.g., $X = R^d$ or $X = \{0,1\}^d$
- Algo sees training sample S: $(x_1, c^*(x_1)), \dots, (x_m, c^*(x_m)), x_i$ i.i.d. from D
 - labeled examples assumed to be drawn i.i.d. from some distr.
 D over X and labeled by some target concept c*
 - labels \in {-1,1} binary classification
 - Algo does optimization over S, find hypothesis h.
 - Goal: h has small error over D.

 $err_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$

Bias: Fix hypotheses space H . (whose complexity is not too large). Realizable: $c^* \in H$. Agnostic: c^* "close to" H.



- Algo sees training sample S: $(x_1, c^*(x_1)), \dots, (x_m, c^*(x_m)), x_i$ i.i.d. from D
- Does optimization over S, find hypothesis $h \in H$.
- Goal: h has small error over D.

True error: $err_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$ How often $h(x) \neq c^*(x)$ over future instances drawn at random from D

• But, can only measure:

Training error: $err_S(h) = \frac{1}{m} \sum_i I(h(x_i) \neq c^*(x_i))$

How often $h(x) \neq c^*(x)$ over training instances

Sample complexity: bound $err_D(h)$ in terms of $err_S(h)$

Consistent Learner

- Input: S: (x₁,c*(x₁)),..., (x_m,c*(x_m))
- Output: Find h in H consistent with the sample (if one exits).

Theorem

$$m \ge \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Contrapositive: if the target is in H, and we have an algo that can find consistent fns, then we only need this many examples to get generalization error $\leq \epsilon$ with prob. $\geq 1 - \delta$

Consistent Learner

- Input: S: (x₁,c*(x₁)),..., (x_m,c*(x_m))
- Output: Find h in H consistent with the sample (if one exits).

Theorem

Bound inversely linear in ϵ

$$m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$. Bound only logarithmic in |H|

- ϵ is called error parameter
 - D might place low weight on certain parts of the space
- δ is called confidence parameter
 - there is a small chance the examples we get are not representative of the distribution

Consistent Learner

- Input: S: (x₁,c*(x₁)),..., (x_m,c*(x_m))
- Output: Find h in H consistent with the sample (if one exits).

Theorem

$$m \ge \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Example: H is the class of conjunctions over $X = \{0,1\}^n$. $|H| = 3^n$ E.g., $h = x_1 \overline{x_3} x_5$ or $h = x_1 \overline{x_2} x_4 x_9$ Then $m \ge \frac{1}{\epsilon} \left[n \ln 3 + \ln \left(\frac{1}{\delta} \right) \right]$ suffice $n = 10, \epsilon = 0.1, \delta = 0.01$ then $m \ge 156$ suffice

Theorem

$$m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Proof Assume k bad hypotheses $h_1, h_2, ..., h_k$ with $err_D(h_i) \ge \epsilon$

1) Fix h_i . Prob. h_i consistent with first training example is $\leq 1 - \epsilon$.

Prob. h_i consistent with first m training examples is $\leq (1 - \epsilon)^m$.

2) Prob. that at least one h_i consistent with first m training examples is $\leq k (1 - \epsilon)^m \leq |H| (1 - \epsilon)^m$.

3) Calculate value of m so that $|H|(1 - \epsilon)^m \le \delta$

3) Use the fact that $1 - x \le e^{-x}$, sufficient to set $|H|(1 - \epsilon)^m \le |H| e^{-\epsilon m} \le \delta$

Sample Complexity: Finite Hypothesis Spaces

Realizable Case

Theorem

$$m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1-\delta$ all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Probability over different samples of m training examples

Sample Complexity: Finite Hypothesis Spaces Realizable Case

1) PAC: How many examples suffice to guarantee small error whp.

Theorem

$$m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

2) Statistical Learning Way:

With probability at least $1 - \delta$, for all $h \in H$ s.t. $err_{s}(h) = 0$ we have

$$\operatorname{err}_{\mathrm{D}}(\mathrm{h}) \leq \frac{1}{\mathrm{m}} \left(\ln |\mathrm{H}| + \ln \left(\frac{1}{\delta} \right) \right).$$

Supervised Learning: PAC model (Valiant)

- X instance space, e.g., $X = \{0,1\}^n$ or $X = R^d$
- S_I={(x_i, y_i)} labeled examples drawn i.i.d. from some distr. D over X and labeled by some target concept c^{*}
 - labels $\in \{-1,1\}$ binary classification
- Algorithm A PAC-learns concept class H if for any target c^* in H, any distrib. D over X, any ε , $\delta > 0$:
 - A uses at most $poly(d,1/\epsilon,1/\delta,size(c^*))$ examples and running time.
 - With probab. 1- δ , A produces h in H of error at $\leq \epsilon$.

What if $c^* \notin H$?

Uniform Convergence

Theorem

$$m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

- This basic result only bounds the chance that a bad hypothesis looks perfect on the data. What if there is no perfect $h \in H$ (agnostic case)?
- What can we say if $c^* \notin H$?
- Can we say that whp all $h \in H$ satisfy $|err_D(h) err_S(h)| \le \epsilon$?
 - Called "uniform convergence".
 - Motivates optimizing over S, even if we can't find a perfect function.

Sample Complexity: Uniform Convergence Agnostic Case

Empirical Risk Minimization (ERM)

- Input: S: (x₁,c*(x₁)),..., (x_m,c*(x_m))
- Output: Find h in H with smallest $err_{S}(h)$

Theorem

$$m \ge \frac{1}{2\varepsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

labeled examples are sufficient s.t. with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$. $1/\epsilon^2$ dependence [as opposed to $1/\epsilon$ for realizable]

Sample Complexity: Finite Hypothesis Spaces Agnostic Case

Theorem

$$m \ge \frac{1}{2\varepsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

labeled examples are sufficient s.t. with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

Important Conclusion:

W.h.p. $\geq 1 - \delta \operatorname{,err}_{D}(\hat{h}) \leq \operatorname{err}_{D}(h^{*}) + 2\epsilon$, \hat{h} is ERM output, h^{*} is hyp. of smallest true error rate.



Sample Complexity: Finite Hypothesis Spaces Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

 $m \geq \frac{1}{2\varepsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right] \text{ something stronger.}$ labeled examples are sufficient s.t. with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

2) Statistical Learning Theory style:

Theorem

With prob. at least $1 - \delta$, for all $h \in H$:

 $\sqrt{\frac{1}{m}}$ as opposed to $\frac{1}{m}$ for realizable

 $1/\epsilon^2$ dependence [as opposed to $1/\epsilon$]

for realizable], but get for

$$\operatorname{err}_{\mathrm{D}}(\mathrm{h}) \leq \operatorname{err}_{\mathrm{S}}(\mathrm{h}) + \underbrace{\frac{1}{2\mathrm{m}}\left(\ln\left(2|\mathrm{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)}_{2\mathrm{m}}}_{2\mathrm{m}}$$

Sample Complexity: Finite Hypothesis Spaces

Realizable Case

Theorem

$$m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Agnostic Case

What if there is no perfect h?

Theorem After *m* examples, with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

$$m \ge \frac{1}{2\varepsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

To prove bounds like this, need some good tail inequalities.

Hoeffding bounds

Consider coin of bias p flipped m times. Let N be the observed # heads. Let $\epsilon \in [0,1]$. Hoeffding bounds:

- $\Pr[N/m > p + \varepsilon] \le e^{-2m\varepsilon^2}$, and $\Pr[N/m .$

Exponentially decreasing tails

Tail inequality: bound probability mass in tail of distribution (how concentrated is a random variable around its expectation).

Sample Complexity: Finite Hypothesis Spaces Agnostic Case

Theorem After *m* examples, with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

$$m \ge \frac{1}{2\varepsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

- Proof: Just apply Hoeffding.
 - Chance of failure at most $2|H|e^{-2|S|\epsilon^2}$.
 - Set to δ . Solve.
 - So, whp, best on sample is ϵ -best over D.
 - Note: this is worse than previous bound (1/ ϵ has become 1/ ϵ^2), because we are asking for something stronger.
 - Can also get bounds "between" these two.

What if H is infinite?

E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



Sample Complexity: Infinite Hypothesis Spaces

H[m] - max number of ways to split m points using concepts in H

Theorem For any class H, distrib. D, if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon} \left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right) \right]$$

then with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $H[m] = O(m^{VCdim(H)})$

Theorem

$$m = O\left(\frac{1}{\varepsilon} \left[VCdim(H) \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \ge \varepsilon$ have $err_S(h) > 0$.

Summary

- PAC/SLT models for supervised learning.
- Notion of sample complexity.
- Sample complexity bounds for finite H (realizable and agnostic).