Foundations for Learning in the Age of Big Data

Maria-Florina (Nina) Balcan Carnegie Mellon University

Brief Overview

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

Lectures 1 & 2: Foundations of classic ML.

• Generalization guarantees. Sample complexity for supervised classification

Lectures 3: Active Learning.

Lectures 4: Distributed Learning.

Supervised Classification. Example: Spam Detection

spam

Decide which emails are spam and which are important.

Supervised classification

Elle Edit View Go Message Iools Help "													
Get Mail 🔹 📝 Write 🔲 Addre	ss Book 📎 Tag * 🔎 Search all messages	2											
thesis													
Smart Folders	🖹 🏠 🖉 Subject 🔍 From	🔒 Date 🔺 I											
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	 10/16/2000 12/6 											
cindofrii													
	Arepty 1 Provide Arepty at 1 Microward 1 ar	rchive II junk II 👗 delere											
students	from Randy Bryant <randy.bryant@cs.cmu *="" 200="" ar<="" reply="" ronward="" td=""><td></td></randy.bryant@cs.cmu>												
students students-diverse	from Randy Bryant «Randy.Bryant@cs.cmul @reply and subject SCS Dissertation Award & ACM Dissertation Nominees	10/14/2009 10:00 PM											
students students-diverse	from Randy Bryant «Randy.Bryant@cs.cmu (@reply) * (@reply) at * @ronward ar subject SCS Dissertation Award & ACM Dissertation Nominees to You	10/14/2009 10:00 PM											
students students-diverse talks-accross-gatech talks-campus	from Randy Bryant eAndy, Bryant@cs.cmul @PEPD [@PEPD] @PEPD at a subject SCS Dissertation Award & ACM Dissertation Nominees to You cc Catherine Copetas ccopetas@cs.cmu.edu> û	10/14/2009 10:00 Pt other actions *											
students students-diverse talks-across-gatech talks-campus talks-caatech	from Randy Bryant & Kandy, Bryant@cs.cmul @dfby) `@dfbyy al `@dfbyy al ` subject SCS Dissertation Nominees to You' cc Catherine Copetas <copetas@cs.cmu.edu>`</copetas@cs.cmu.edu>	rchive Junk Adeete 10/14/2009 10:00 Ph other actions -											
students students-diverse talks-accross-gatech talks-campus talks-gatech talks-outside	from Randy Bryant & Randy, Bryant@sc.mul @wFDPy ``@wFDPy```@wFDPy```@wFDPy```@wFDPy```@wFDPy```@wFDPy```@wFDPy```@ subject SEC Dissertation Award & ACM Dissertation Nominees to You? cc Catherine Copetas <copetas@ks.cmu.edu>?? Nina:</copetas@ks.cmu.edu>	other actions *											
students students-diverse talks-accross-gatech talks-gatech talks-gatech talks-outside talks-ustide	from Randy Bryant early Bryant@ec.cmu @#PDP ` @#DPD ` @#DDD ` @#DPD ` @#DDD ` @#DDD ` @#DDD ` @#DPD ` @#DPD ` @#DDD ` @#DD ` @# ` #DD ` @#DD ` DD `	other actions -											
students students-diverse talks-accross-gatech talks-campus talks-outside talks-outside talks-outside talks-outside	<pre>from Randy Bryant <& Randy, Bryant @ex.andl @wfDPy ``@wfDPy ``@wfDPy al ``@vfDPy al ``@vfDPyy al ``@vfDPy al ``@vfDPyy al</pre>	tonive junk A deeter 10/14/2009 10:00 PM other actions - d like to tion. I would											
students students talks-accoss-gatech talks-gatech talks-gatech talks-outside teaching teaching teach-report	from Randy Byent & Randy Byent @ Kandy Byent@ Kandy Byen	tonve junk detection of the sector of the se											
students students talk-across-gatech talk-across-gatech talk-gatech talk-outside teak-outside teak-outside tech-report teh-report	<pre>from Randy Byzet & Randy, Byzet & Randy, Byzet & Byzet &</pre>	d like to tion. I would ed lecture ork out the											
students students-diverse tables-conso-gatech tables-conso-gatech tables-canpus tables-catech tables-catech tables-patech tables	from Randy Byant & Arandy Byant@sc.mul @#PDP ` @#PENY` a * @*PONWard a subject SCS Dissertation Award & ACM Dissertation Nominees to You? c Catherine Copetas <copetas@cc.mu.edu>? Nina: You might have already seen this announcement, but I would personally congratulate you for your outstanding dissectat like to invite you to return to CHU to give a distinguishe sometime in the winter of 2010. Catherine Copetas will wo timing for you. Tou'll get to use the new Rashid Auditori</copetas@cc.mu.edu>	d like to tion. I would d letture ork out the iuma big											
atudents students-diverse tale-accoss-gatech tale-scampus tale-spacech tale-space ta	from Randy Byant deandy, Byant@c.cmu (@PED) ` @PEDD' (@PEDD') ` @PEDD' (` ` ` ` ` ` ` ` ` ` ` ` ` ` ` ` ` ` `	Init actions - 10/14/2009 10:00 PP other actions - d like to tion. I would at lecture ork out the iuma big											
students students-diverse stale-sconse-gatech tale-sconse-gatech tale-scanpus tale-scanpus tale-subside tech-report theory-group theory-group theory-tales total-diverse-gatech	<pre>from Randy Bryant & Arandy, Bryant@sc.mul @wFDPy `` @wFDPy `` @wFDPy at `` @wFDPyy at `` @wFDPyyyyyyyyyyyyyyyyyyyyyyy</pre>	Init a detection of the second											
students students-diverse stale-scores-gatech stale-scores-gatech stale-space	from Randy Byant & Arandy Byant @c. cmul @ PEDY ` @ PEDY i @ PEDY at @ Forward & subject SCS Dissertation Award & ACM Dissertation Nominees to You? cc Catherine Copetas <copetas@cs.cmu.edu>? Nina: You might have already seen this announcement, but I would personally congratulate you for your outstanding dissertat like to invite you to return to CHU to give a distinguishe sometime in the winter of 2010. Catherine Copetas will wo timing for you. You'll get to use the new Rashid Auditori improvement over Wean 7500. Best of wishes to you at Georgia Tech.</copetas@cs.cmu.edu>	International and the second s											

Not spam

🏐 S	🔄 SPAM for dbalcan@cs.cmu.edu - Thunderbird													X
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 Doru Inbox				s	tudent loan d	ebt		•	Ahmed Guth	nrie 💧	1/28/20	^
ė.	🥏 Int			-		d	ebt consoloda	ation		•	Emanuel Co	r 💧	1/28/20	
						V	erified You Or	rdered Meds		۰	Tamika Thor	pe 💧	1/28/20	
	-	Sent				G	iive that girl t	he best goft f	or Valentine's		Clare	0	1/28/20	
	- 6	SPAM	(42)			h	ad cradit dab	t concolidation	n lann	-	Topin Laird	A 1	1 /10 /10	
Trash					1	Thun	derbird thir	nks this me	sage is jun	k.		Th	iis is Not Ju	nk
🕀 🔜 Acasa					0									
🗉 🔜 colegi					<u>s</u>	Subjec	t: debt cor	isolodation						
🖲 📶 etc 					Fron	n: Emanuel (Cortez <wcon< th=""><th>vention@heil</th><th>hecker.de</th><th>≥</th><th></th><th></th><th></th></wcon<>	vention@heil	hecker.de	≥				
					Dat	e: 1/28/2008	3 4:14 PM							
						Т	o: <u>dayne@c</u>	s.cmu.edu						
					Sav	good	d bye to	debt.						
					Acce	eptak	ble Unsed	cured Del	bt inclu	des All	l Major	Credi	it Card	s,
	🔜	mike			No-c	colla	ateral Ba	ank Loans	s, Perso	nal Loa	ans,			
	··· 🛄	nina			Medi	ical	Bills et	tc.						
	۵ 🛄	Profeso	ri											
	۵ 🛄	scoala			http	p://v	www.badde	ebth.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.

PAC/SLT models for Supervised Learning

- X feature/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 drawn i.i.d. from D and labeled by target 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 hypothesis space H [whose complexity is not too large]

- Realizable: $c^* \in H$.
- Agnostic: c^* "close to" H.

PAC/SLT models for Supervised Learning

- 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: $\operatorname{err}_{D}(h) = \Pr_{x \sim D}(h(x) \neq c^{*}(x))$ How often $h(x) \neq c^{*}(x)$ over future instances drawn at random from D

• But, can only measure:

Training error: $\operatorname{err}_{S}(h) = \frac{1}{m} \sum_{i} I(h(x_{i}) \neq c^{*}(x_{i}))$

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

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

Sample Complexity for Supervised Learning

Consistent Learner

Theorem

- Input: S: $(x_1, c^*(x_1)), \dots, (x_m, c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Bound only logarithmic in |H|, linear in $1/\epsilon$

 $m \geq \frac{1}{\varepsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$ labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$. Probability over different samples of m training examples

So, if $c^* \in H$ and can find consistent fns, then only need this many examples to get generalization error $\leq \epsilon$ with prob. $\geq 1 - \delta$

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 $\cdot \epsilon$.

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

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

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

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

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

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



- H[S] the set of splittings of dataset S using concepts from H.
- H[m] max number of ways to split m points using concepts in H

 $H[m] = \max_{|S|=m} |H[S]|$

- H[S] the set of splittings of dataset S using concepts from H.
- H[m] max number of ways to split m points using concepts in H

 $H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \le 2^m$



In general, if |S|=m (all distinct), $|H[S]| = m + 1 \ll 2^m$

- H[S] the set of splittings of dataset S using concepts from H.
- H[m] max number of ways to split m points using concepts in H



In general, |S|=m (all distinct), $H[m] = \frac{m(m+1)}{2} + 1 = O(m^2) \ll 2^m$

There are m+1 possible options for the first part, m left for the second part, the order does not matter, so ((m+1) choose 2) + 1 (for empty interval).

- H[S] the set of splittings of dataset S using concepts from H.
- H[m] max number of ways to split m points using concepts in H

 $H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \le 2^m$

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

Sample Complexity: Infinite Hypothesis Spaces Realizable Case

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

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

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

• Not too easy to interpret sometimes hard to calculate exactly, but can get a good bound using "VC-dimension"

If $H[m] = 2^m$, then $m \ge \frac{m}{\epsilon}(....)$

• VC-dimension is roughly the point at which H stops looking like it contains all functions, so hope for solving for m.

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 $err_S(h) > 0$.

Definition: 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, 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 $VCdim(H) = \infty$

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 $VCdim(H) = \infty$

To show that VC-dimension is d:

- there exists a set of d points that can be shattered
- there is no set of d+1 points that can be shattered.

Fact: If H is finite, then $VCdim(H) \le log(|H|)$.

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.



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= Union of k intervals on the real line VCdim(H) = 2k





E.g., H= linear separators in R^2

VCdim(H) < 4

Case 1: one point inside the triangle formed by the others. Cannot label inside point as positive and outside points as negative.

Case 2: all points on the boundary (convex hull). Cannot label two diagonally as positive and other two as negative.

Fact: VCdim of linear separators in R^d is d+1





Sauer's Lemma

Sauer's Lemma:

Let d = VCdim(H)

- $m \le d$, then $H[m] = 2^m$
- m>d, then $H[m] = O(m^d)$

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

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 $err_S(h) > 0$.

Sample Complexity: Infinite Hypothesis Spaces Realizable Case

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 $err_S(h) > 0$.

E.g., H= linear separators in
$$\mathbb{R}^d$$
 $m = O\left(\frac{1}{\varepsilon}\left[d\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$

Sample complexity linear in d

So, if double the number of features, then I only need roughly twice the number of samples to do well.

Data Dependent Generalization Bounds

- Distribution/data dependent. Tighter for nice distributions.
- Apply to general classes of real valued functions & can be used to recover the VC-bounds for supervised classification.
- Prominent technique for generalization bounds since 2000.

Covering Numbers Generalization Bounds

See Anthony-Bartlett, "Neural Network Learning: Theoretical Foundations", 1999.

Rademacher Complexity Generalization Bounds

See Bousquet-Boucheron-Lugosi, "Introduction to Statistical Learning Theory", 2014.

Summary

- PAC/SLT models for supervised learning.
- Notion of sample complexity.
- Shattering, VC dimension as measure of complexity, Sauer's lemma, form of the VC bounds .