Machine Learning in the Data Revolution Era

Shai Shalev-Shwartz

School of Computer Science and Engineering The Hebrew University of Jerusalem



Machine Learning Seminar Series, Google & University of Waterloo, July 2010 10 Years Ago ...

 ALLWEIN, SCHAPIRE & SINGER, 2000 (received Best 10-year paper at ICML 2010)

	#Examples			
Problem	Train	Test	#Attributes	#Classes
dermatology	366	-	34	6
satimage	4435	2000	36	6
glass	214	-	9	7
segmentation	2310	-	19	7
ecoli	336	-	8	8
pendigits	7494	3498	16	10
yeast	1484	-	8	10
vowel	528	462	10	11
soybean	307	376	35	19
thyroid	9172	-	29	20
audiology	226	-	69	24
isolet	6238	1559	617	26
letter	16000	4000	16	26

Table 1: Description of the datasets used in the experiments.

< 🗗 🕨 🔸

Data Revolution

In 2009, more data has been generated by individuals than in the entire history of mankind through 2008 \dots

WikipediA













Shai Shalev-Shwartz (Hebrew U)

Learning in the Data Revolution Era

Google-UW, July'10 3 / 40

Document

If only Tiger Woods could bond with fans as well as he did with his caddie on the Old Course practice range on Monday he might find the road to redemption all the more simple...

About Sport?



E.g.

- \mathcal{X} instance domain (e.g documents)
- \mathcal{Y} label set (e.g. document category is about sport?)
- Goal: learn a predictor $h: \mathcal{X} \to \mathcal{Y}$
- \mathcal{H} Set of possible predictors
- Learning rule (e.g. "find $h \in \mathcal{H}$ with minimal error on training set")

10k training examples

1 hour



2.3% error

< 🗇 🕨

3



1 hour

2.3% error



2.29% error

< 🗇 🕨

3







• Can always sub-sample and get error of 2.3% using 1 hour







- Can always sub-sample and get error of 2.3% using 1 hour
- Say we're happy with 2.3% error What else can we gain from the larger data set?



2



-Google-UW, July'10

A 🖓

3

How can more data reduce runtime?

- Learning using Stochastic Optimization (S. & Srebro 2008)
- Injecting Structure (S, Shamir, Sirdharan 2010)

How can more data compensate for missing information?

- Attribute Efficient Learning (Cesa-Bianchi, S., Shamir 2010) Technique: Rely on Stochastic Optimization
- Ads Placement (Kakade, S., Tewari 2008) Technique: Inject Structure by Exploration

Background on ML: Vapnik's General Learning Setting

Probably Approximately Solve:

$$\min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{z \sim \mathcal{D}} \left[f(h, z) \right]$$

where:

- \mathcal{D} is an unknown distribution
- Can only get samples $z_1, \ldots, z_m \overset{\text{i.i.d.}}{\sim} \mathcal{D}$
- $\mathcal H$ is a known hypothesis class
- f(h,z) is the loss of using hypothesis h on example z

Background on ML: Vapnik's General Learning Setting

Probably Approximately Solve:

$$\min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{z \sim \mathcal{D}} \left[f(h, z) \right]$$

where:

- \mathcal{D} is an unknown distribution
- Can only get samples $z_1, \ldots, z_m \overset{\text{i.i.d.}}{\sim} \mathcal{D}$
- ${\mathcal H}$ is a known hypothesis class
- f(h, z) is the loss of using hypothesis h on example z

Main question studied by Vapnik:

How many examples are required to find $\hat{h} = A(z_1, \ldots, z_m)$ s.t.

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[f(\hat{h},z)
ight] - \min_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[f(h,z)
ight] \ \le \ \epsilon \quad ext{ w.p. } \ge 1-\delta$$

Examples

Document Categorization:

- $z = (\mathbf{x}, y)$, x represents a document and $y \in [k]$ is the topic
- $h \in \mathcal{H}$ is a predictor $h : \mathcal{X} \to [k]$
- $f(h, z) = \mathbf{1}[h(\mathbf{x}) \neq y]$
- k-means clustering:
 - $z \in \mathbb{R}^d$
 - $h \in \mathbb{R}^{k,d}$ specifies k cluster centers
 - $f(h,z) = \min_{i} \|h_{i,\rightarrow} z\|$
- Density Estimation:
 - h is a parameter of a density $p_h(z)$
 - $f(h,z) = -\log p_h(z)$

Ads Placement:

- $z = (\mathbf{x}, \mathbf{c})$, x represents context and $\mathbf{c} \in [0, 1]^k$ represents the gain of placing each ad
- $h: \mathcal{X} \to \mathbb{S}^k$ returns the probability to place each ad
- $f(h, z) = \langle h(x), \mathbf{c} \rangle$ is the expected gain of following policy h

JOC E

Empirical Risk Minimization:

Return a hypothesis with smallest risk on the training examples:

$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} rac{1}{m} \sum_{i=1}^m f(h, z_i)$$

Empirical Risk Minimization:

Return a hypothesis with smallest risk on the training examples:

$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} rac{1}{m} \sum_{i=1}^m f(h, z_i)$$

Sample Complexity: If $m \ge \Omega\left(\frac{C(\mathcal{H}) \log(1/\delta)}{\epsilon^2}\right)$, where $C(\mathcal{H})$ is some complexity measure, then \hat{h} is Probably Approximately Correct.

SVM is a very popular Machine Learning algorithm



[PDF] Probabilistic outputs for SVMs and comparisons to regularized likelihood methods J Platt - Advances in large margin classifiers, 1999 - cs.cornell.edu ... Issue: How to choose the training set? Using the output of the SVM for the training set. Biased estimate both for linear and non linear SVMs. (Re)using a hold out set. Using cross-validation. John Platt Probabilistic Outputs for SVMs and Comparisons to Regularized Page 12. ... Cited by 1123 - Related articles - View as HTML - All 8 versions

Making large scale SVM learning practical

T Joachims - 1999 - eldorado.uni-dortmund.de

... The results give guidelines for the application of **SVMs** to large domains. ... 1 SV M light is available at http://www-ai.cs.uni-dortmund.de/**svm** light Page 5. ... Vapnik 1995] shows how training a **support vector machine** for the pattern recognition problem leads to the following quadratic ...

Cited by 3590 - Related articles - View as HTML - All 14 versions

12 / 40

- 4 同 ト - 4 三 ト - 4 三 ト

SVM learning rule:

• Given samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ solve:

$$\min_{\mathbf{w}} \ \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \max\{\mathbf{0}, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}$$

Casting in the General Learning Setting:

•
$$z = (\mathbf{x}, y)$$

- $h = \mathbf{w}$
- $f(h,z) = \max\{0, 1 y\langle \mathbf{w}, \mathbf{x} \rangle\} + \frac{\lambda}{2} \|\mathbf{w}\|^2$
- SVM learning rule is exactly ERM for the above loss function

- SVM learning rule can be rewritten as a Quadratic Optimization Problem.
- Solvable in polynomial time by standard solvers. Dedicated solvers also exist. End of story?

- SVM learning rule can be rewritten as a Quadratic Optimization Problem.
- Solvable in polynomial time by standard solvers. Dedicated solvers also exist. End of story?
- $\bullet\,$ Not quite... Runtime of standard solvers increases with m
- What if *m* is very large ?

- Our real goal is not to minimize the SVM optimization problem
- SVM optimization problem is just a proxy for achieving the real goal

 find a hypothesis with small error on unseen data
- Should study runtime required to solve the *Learning Problem*:
 - $T(m, \epsilon)$ runtime required to find ϵ -accurate solution w.r.t. unseen data when number of training examples available is m
 - $T(\infty, \epsilon)$ data laden analysis data is plentiful and computational complexity is the real bottleneck



3



Google-UW, July'10

- ∢ ≣ →

16 / 40

3

Data Laden Analysis of SVM Solvers

Method	Optimization runtime	$T(\infty, \epsilon)$
Interior Point	$m^{3.5}\log\lograc{1}{\epsilon}$	$\frac{1}{\epsilon^7}$
Dual decompositoin	$m^2 \log \frac{1}{\epsilon}$	$\frac{1}{\epsilon^4}$
SVM-Perf (Joachims '06)	$rac{m}{\lambda\epsilon}$	$\frac{1}{\epsilon^4}$
SGD (S, Srbero, Singer '07)	$\frac{1}{\lambda\epsilon}$	$\frac{1}{\epsilon^2}$

"Best" optimization method is the worst learning method

Shai Shalev-Shwartz (Hebrew U)

Google-UW, July'10

3 D (3 D)

< □ > < ---->

More Data Less Work



Shai Shalev-Shwartz (Hebrew U)

Google-UW, July'10

More Data Less Work



Intuition: Better to do simple operations on more examples than complicated operations on fewer examples

Google-UW, July'10

How can more data reduce runtime?

- Learning using Stochastic Optimization (S. & Srebro 2008) \checkmark
- Injecting Structure (S, Shamir, Sirdharan 2010)

How can more data compensate for missing information?

- Attribute Efficient Learning (Cesa-Bianchi, S., Shamir 2010) Technique: rely on Stochastic Optimization
- Ads Placement (Kakade, S., Tewari 2008) Technique: Inject Structure by Exploration

- Replace original hypothesis class with a larger hypothesis class
- On one hand: Larger class has more structure \Rightarrow easier to optimize
- On the other hand: Larger class \Rightarrow larger estimation error \Rightarrow need more examples



• Goal: learn a Boolean function $h: \{0,1\}^d \rightarrow \{0,1\}$



- Goal: learn a Boolean function $h : \{0,1\}^d \rightarrow \{0,1\}$
- Hypothesis class: 3-term DNF formulae:



- Goal: learn a Boolean function $h: \{0,1\}^d \rightarrow \{0,1\}$
- Hypothesis class: 3-term DNF formulae:

• E.g.
$$h(\mathbf{x}) = (x_1 \land \neg x_3 \land x_7) \lor (x_4 \land x_2) \lor (x_5 \land \neg x_9)$$



- Goal: learn a Boolean function $h: \{0,1\}^d \rightarrow \{0,1\}$
- Hypothesis class: 3-term DNF formulae:

• E.g.
$$h(\mathbf{x}) = (x_1 \wedge \neg x_3 \wedge x_7) \vee (x_4 \wedge x_2) \vee (x_5 \wedge \neg x_9)$$

• Sample complexity is order d/ϵ



- Goal: learn a Boolean function $h: \{0,1\}^d \rightarrow \{0,1\}$
- Hypothesis class: 3-term DNF formulae:

• E.g.
$$h(\mathbf{x}) = (x_1 \land \neg x_3 \land x_7) \lor (x_4 \land x_2) \lor (x_5 \land \neg x_9)$$

- Sample complexity is order d/ϵ
- Kearns & Vazirani: If RP \neq NP, it is not possible to efficiently find an ϵ -accurate 3-DNF formula



- Goal: learn a Boolean function $h: \{0,1\}^d \rightarrow \{0,1\}$
- Hypothesis class: 3-term DNF formulae:

• E.g.
$$h(\mathbf{x}) = (x_1 \land \neg x_3 \land x_7) \lor (x_4 \land x_2) \lor (x_5 \land \neg x_9)$$

- Sample complexity is order d/ϵ
- Kearns & Vazirani: If RP≠NP, it is not possible to efficiently find an ε-accurate 3-DNF formula
- Claim: if $m \geq d^3/\epsilon$ it is possible to find a predictor with error $\leq \epsilon$ in polynomial time



Proof

- Observation: 3-DNF formula can be rewritten as $\wedge_{u \in T_1, v \in T_2, w \in T_3} (u \lor v \lor w)$ for three sets of literals T_1, T_2, T_3
- Define: $\psi : \{0,1\}^d \to \{0,1\}^{2(2d)^3}$ s.t. for each triplet of literals u, v, w there are two variables indicating if $u \lor v \lor w$ is true or false
- Observation: Each 3-DNF can be represented as a single conjunction over $\psi(\mathbf{x})$
- Easy to learn single conjunction (greedy or LP)



Trading samples for runtime

Algorithm	samples	runtime
3-DNF over x	$rac{d}{\epsilon}$	2^d
Conjunction over $\psi(\mathbf{x})$	$\frac{d^3}{\epsilon}$	$\operatorname{poly}(d)$



Google-UW, July'10

3

More complicated example: Agnostic learning of Halfspaces with 0-1 loss



Google-UW, J<u>uly'10</u>

How can more data reduce runtime?

- Learning using Stochastic Optimization (S. & Srebro 2008)
- Injecting Structure (S, Shamir, Sirdharan 2010)

How can more data compensate for missing information?

- Attribute Efficient Learning (Cesa-Bianchi, S., Shamir 2010) Technique: rely on Stochastic Optimization
- Ads Placement (Kakade, S., Tewari 2008) Technique: Inject Structure by Exploration

Attribute efficient regression

- Each training example is a pair $(\mathbf{x}, y) \in \mathbb{R}^d imes \mathbb{R}$
- Partial information: can only view O(1) attributes of each individual example



Sponsored Advertisement (contextual multi-armed bandit)

- Each training example is a pair $(\mathbf{x},\mathbf{c})\in\mathbb{R}^d imes [0,1]^k$
- Partial information: Don't know c. Can only guess some y and know the value of cy

	DTUTO 6101	were hotels in israel
	252.041	 חפש באינטרנט () חפש דפים בעברית
נוך כ- 58,400,000 עבור hotels in israel עבור ס.25) איניות)	תוצאות 1 - 10 מח	אינטרנט 🛨 ה <u>גג אפשרויות</u>
קישורים ממומנים RAMOT RESORT HOTEL	קישורים ממומנים	75% Off in Tel Aviv Hotels Book your room online Cheap Tel Aviv Hotels Galil-Hotels.com/Engy-Tel Aviv
Overlooking the Sea of Galliee www.ramot-nofesh.com		Specials from \$60 in Tel Aviv Israe Three minutes walk to the beach www.hostelstelawiv.com Hilton Hotels in Israel
Jerusalem newest & most fashionable Lifestyle 5 Stars Luxury Hotel . www.MamillaHotel.com		A choice of superb Hilton hotels in Israel. Book online today! www.Hilton.com/Israel
<u>Tel Aviv Hotel</u> Tel Aviv - Book Now! Amazing Discounts - Up To 80% Off HotelsCombined.com/Tel_Aviv		strade i notes: - מתונות בישראא - בתי מתן אתר המרכז מתומת כא אזורי האידע המידע לגי לאון לא ליאור, מיקום, מבצעים אפשרות הזמנה. מלופת באינות - מתונות בים המלח - מלופת בתל אביב - יבושלים והרי-הודב www.hotels.co.W - <u>שרמ שמור</u> - דמה - ⊘ו€וצ
Accomodation in Tel Aviv The Ultimate Tel Aviv Experience For Short&Medium Length Stays I Tel-Aviv.ipp-pages.co.il		HOTELS.CO.Lt. בישראל - Linu בישראל - HOTELS.CO.Lt. את 11 הספויעם בישראל שאת הים האתר המוביל בישראל לשיוק מלום ווחפשות בשיוק ישיר. כל הגורמים המופיעם בו מגרהלים את פעילות במערטר באפון עבמאי יאמיני בו מגרהלים את פעילות במערטר באפון עבמאי
Tel Aviv Hotel Compare more than 99000 hotels. Low Prices Guaranteel Hotels.com/tel-sviv		[ang nama] - <u>Hotels in Israel</u> Reviews on hotels in Israel Hotels in Israel Wooffer a huge choice of Israel hotels, Israel Special Hotel Deals and reviewed hotels in Israel Tal Aviv Hotels - Jenuzatiem Hotels - Deal See Hotels [전] 은 · · · <u>and · rane grape vuer</u> hotels - Israel.com/
Internet Access in Israel Connect your Laptop to the Internet with a cellular modern for €10/day www.teiroam.com		רתאחדות המלונות התאחדות המלומת ביטראל, ו-HA, מלומת בארץ: אילת,ירושלים,תל אביב, אירועים במלומת, חדשות מלומת, האתר הרששים להמחדות המלומת בשראל. ע ≹מסים ארשאי - <u>עתש שמור - דמה</u> - ⊘ או א
Postouropto of level		A state of the second

Learning in the Data Revolution Era

Google-UW, July'10

28 / 40

Shai Shalev-Shwartz (Hebrew U)

Three main techniques:

- Missing information as noise
- Active Exploration try to "fish" the relevant information
- Inject structure problem hard in the original representation but becomes simple in another representation (different hypothesis class)

More data helps because:

- It reduces variance compensates for the noise
- It allows more exploration
- It compensates for larger sample complexity due to using larger hypotheses classes

Formal problem statement:

- Unknown distribution $\mathcal D$ over $\mathbb R^d\times\mathbb R$
- \bullet Goal: learn a linear predictor $\mathbf{x}\mapsto \langle \mathbf{w},\mathbf{x}\rangle$ with low risk:
- Risk: $L_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{\mathcal{D}}[(\langle \mathbf{w}, \mathbf{x} \rangle y)^2]$
- Training set: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$
- Partial information: For each (x_i, y_i) , learner can view only k attributes of x_i
- Active selection: learner can choose which k attributes to see

Similar to "Learning with restricted focus of attention" (Ben-David & Dichterman 98)

- Usually difficult exponential ways to complete the missing information
- Popular approach Expectation Maximization (EM)

Previous methods usually do not come with guarantees (neither sample complexity nor computational complexity) • Simply set the missing attributes to a default value



• Use your favorite regression algorithm for full information, e.g.,

$$\min_{\mathbf{w}} \sum_{i=1}^{m} (\langle \mathbf{w}, \tilde{\mathbf{x}}_i \rangle - y_i)^2 + \lambda \|\mathbf{w}\|_p^p$$

- Ridge Regression: p = 2
- Lasso: p = 1

- Simply set the missing attributes to a default value
- Use your favorite regression algorithm for full information, e.g.,

$$\min_{\mathbf{w}} \sum_{i=1}^{m} (\langle \mathbf{w}, \tilde{\mathbf{x}}_i \rangle - y_i)^2 + \lambda \|\mathbf{w}\|_p^p$$

- Ridge Regression: p = 2
- Lasso: p = 1
- Efficient
- Correct? Sample complexity? How to choose attributes ?



• Observation:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \frac{1}{d} \begin{pmatrix} dx_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \frac{1}{d} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ dx_d \end{pmatrix}$$

• Therefore, choosing i uniformly at random gives

$$\mathbb{E}_i[dx_i\mathbf{e}^i] = \mathbf{x} \; .$$

- If $\|\mathbf{x}\| \leq 1$ then $\|dx_i\mathbf{e}^i\| \leq d$ (i.e. variance increased)
- Reduced missing information to unbiased noise
- Many examples can compensate for the added noise

- True goal: minimize over \mathbf{w} the true risk $L_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}[(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2]$
- Loss on one example, $(\langle {\bf w}, {\bf x} \rangle y)^2$, gives an unbiased estimate for $L_{\mathcal{D}}({\bf w})$
- Averaging loss on many examples reduces variance
- In our case, we construct an unbiased estimate of the loss of each single example
- Variance increases but many examples still reduces it back

Theorem (Cesa-Bianchi, S, Shamir)

Let $\hat{\mathbf{w}}$ be the output of LaBAER. Then, with overwhelming probability

$$L_{\mathcal{D}}(\mathbf{\hat{w}}) \leq \min_{\mathbf{w}: \|\mathbf{w}\|_1 \leq B} L_D(\mathbf{w}) + \tilde{O}\left(rac{d^2 B^2}{\sqrt{m}}
ight) \; ,$$

where d is dimension and m is number of examples.

Theorem (Cesa-Bianchi, S, Shamir)

Let $\hat{\mathbf{w}}$ be the output of LaBAER. Then, with overwhelming probability

$$L_{\mathcal{D}}(\mathbf{\hat{w}}) \leq \min_{\mathbf{w}: \|\mathbf{w}\|_1 \leq B} L_D(\mathbf{w}) + \tilde{O}\left(rac{d^2 B^2}{\sqrt{m}}
ight) \;,$$

where d is dimension and m is number of examples.

• Factor of d^4 more examples compensates for the missing information !

Theorem (Cesa-Bianchi, S, Shamir)

Let $\hat{\mathbf{w}}$ be the output of LaBAER. Then, with overwhelming probability

$$L_{\mathcal{D}}(\mathbf{\hat{w}}) \leq \min_{\mathbf{w}: \|\mathbf{w}\|_1 \leq B} L_D(\mathbf{w}) + \tilde{O}\left(rac{d^2 B^2}{\sqrt{m}}
ight) \;,$$

where d is dimension and m is number of examples.

Factor of d⁴ more examples compensates for the missing information !
Can we do better?

- The factor d^2 in the bound because we estimate a matrix $\mathbf{x}\mathbf{x}^T$
- Can we avoid estimating the matrix ?
- Yes ! Estimate the gradient of the loss instead of the loss
- Follow the Stochastic Gradient Descent approach
- Leads to a much better bound



- Full information classifiers (top line) \Rightarrow error of $\sim 1.5\%$
- Our algorithms (bottom line) \Rightarrow error of $\sim 4\%$

Demonstration



Google-UW, July'10

- Partial supervision
- Technique: Active Exploration
- Larger regret can be compensated by having more examples



- Learning theory: Many examples \Rightarrow smaller error
- This work: Many examples \Rightarrow
 - higher efficiency
 - compensating for missing information
- Techniques:
 - Stochastic optimization
 - Inject structure
 - Missing information as noise
 - 4 Active Exploration