On generalization bounds, projection profile, and margin distribution

Ashutosh Garg
Sariel Har-Peled
Dan Roth

University of Illinois, Urbana-Champaign
danr@cs.uiuc.edu
http://L2R.cs.uiuc.edu/~danr
Learning with high dimensional data

- Identifying phrase structure
  \[ [NP \text{ He}] [VP \text{ reckons}] [NP \text{ the current account deficit}] [VP \text{ will narrow}] [PP \text{ to}] [NP \text{ only # 1.8 billion}] [PP \text{ in}] [NP \text{ September}] \]

- Information Extraction Tasks
  \[ \text{afternoon, Dr. Ab C will talk in Ms. De. F class..} \]

- Prepositional Phrase Attachment

- Context Sensitive Spelling Correction
  \[ \text{Illinois’ bored of education} \] board
Learning with high dimensional data

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

[afternoon] [Dr. Ab C] will talk in Ms. De. F class..

buy shirt with sleeves, buy shirt with a credit card

Illinois’ bored of education board

Features include: (patterns of) words; POS tags; relational information (location; order; structure…)

In many of these problems dimensionality is $10^5$ or more

Bounds introduction Snowbird’02
Easiness of Learning

We learn well from relatively small number of examples in very high dimensional spaces? Should we believe it?

Some high dimensional problems are naturally constrained and become, effectively, low dimensional problems.

[Roth, Zelenko’00; Garg, Roth’01, Vempala’00]

In these cases, although learning is done in high dimension, generalization ought to depend on the true, lower dimensionality of the problem.

Not exploited by current theories
This work

Introduces a way to analyze learning in high dimension in a way that exploits the lower, effective dimensionality of the data.

Random projection methods are used to explicitly exploit the margin distribution.

Exhibits generalization bounds that are (sometimes) realistic (< 0.5) for real problems in NLP and vision.
Standard Bounds

**VC dimension based bounds** (hyperplanes)

\[
ERR_D \leq ERR_S + \sqrt{\frac{n(\ln(2m/n) + 1) - \ln(\delta/4)}{m}}
\]

**Margin Based bounds** (data dependent; \(\gamma - \text{margin}\))

\[
ERR_D \leq ERR_S + \frac{2}{m} \left(\frac{1}{\gamma^2} \log(32m) \log(8em\gamma^2) + \log(8m/\delta)\right)
\]

\(VC(n,m)\)
Intuition

**Hard Problem**

**Easy Problem**

hyperplane $h$

$\gamma = \min_{s} h^t x$
Standard Bounds

**VC dimension based bounds** (hyperplanes) \[ \text{ERR}_D \leq \text{ERR}_S + \sqrt{\frac{n(\ln(2m/n) + 1) - \ln(\delta/4)}{m}} \]

**Margin Based bounds** (data dependent; $\gamma$ – margin) \[ \text{ERR}_D \leq \text{ERR}_S + \frac{2}{m} \left( \frac{1}{\gamma^2} \log(32m) \log(8em\gamma^2) + \log(8m/\delta) \right) \]

**Typically:** \[ 1 \ll \text{VC bounds} < \text{Margin Based bound} \]
Real Data

17,000 dimensional context sensitive spelling
Histogram of distance of points from the hyperplane
Standard Bounds

**VC dimension based bounds** (hyperplanes)

\[ ERR_D \leq ERR_S + \sqrt{\frac{n(\ln(2m/n) + 1) - \ln(\delta/4)}{m}} \]

**Margin Based bounds** (data dependent; \( \gamma \) – margin)

\[ ERR_D \leq ERR_S + \frac{2}{m} \left( \frac{1}{\gamma^2} \log(32m) \log(8em\gamma^2) + \log(8m/\delta) \right) \]

**Typically:** 1 \( \ll \) VC bounds < Margin Based bound

**Value of bounds:** algorithmic insight; model selection
This work

Even for:
17,000 dimensional
context sensitive spelling

Can get bounds
that are < 0.5,
using a 1000–5000 examples.
Key Idea: Projection Profile (I)

Learn a Hyperplane $h$ from sample $S$, in high dimension $n$

Analysis: Project $S$ and $h$ randomly to low dimension ($k$) 

w.h.p $(k,S)$: small distortion of distances. (Johnson-Lindenstrauss)

Small error in the lower dimension
Key Idea: Projection Profile (II)

Expected amount of error introduced in projection captured by:  
\[ a_k(D, h) = \int_{x \in D} u(x) dD \]

where: \( u_k(x) = \min \left\{ \exp \left( -\frac{k}{|l(x)|} \right), \frac{1}{kl^2(x)}, 1 \right\} \) 
\[ l(x) = h^t x \]

the profile:
\[ P(D, h) = (a_1(D, h), a_2(D, h), \ldots a_k(D, h), \ldots) \]
Key Idea: Projection Profile (II)

Expected amount of error introduced in projection captured by:

\[ a_k(D, h) = \int_{x \in D} u(x) dD \]

where:

\[ u_k(x) = \min \left\{ \exp \left( -\frac{k}{l(x)} \right), \frac{1}{kl^2(x)}, 1 \right\} \]

the profile:

\[ P(D, h) = (a_1(D, h), a_2(D, h), ... a_k(D, h), ...) \]

gives the tradeoff between dimensionality and accuracy

Resulting bound:

\[ ERR_D \leq ERR_S + \min_k \left\{ \hat{u}_k + VC(k, m) \right\} \]
Rest of the talk

- Some details
  - Random projection
  - Random projection for classification
  - Projection profile of a sample
- Analysis

- Future/Questions
Random Projection

Random Matrix: \( R[k \times n] \) with \( r_{ij} \sim N(0,1/k) \)
\[ x \in \mathbb{R}^n, \quad x' = Rx \in \mathbb{R}^k \]

Theorem [Johnson-Lindenstrauss 84]:
\( u,v \in \mathbb{R}^n \); \([u',v'] = R[u,v]\), projections to \( \mathbb{R}^k \). For any \( c \)
\[
\Pr \left[ (1 - c) \leq \frac{||u' - v'||^2}{||u - v||^2} \leq (1 + c) \right] \geq 1 - e^{-c^2k/8}
\]
where the probability is over the selection of the random matrix \( R \).
Plan

- Project a sample and the hyperplane
- Bound empirical error in the projected space (k)
Random Projection: A Classification Version

Lemma:

$h$: n-dimensional classifier, $x \in \mathbb{R}^n; ||h||=||x||=1, l(x)=h^T x$

The probability of misclassifying $x$ due to the random projection $R$, is

$$P\left[\text{sgn}(h^T x) \neq \text{sgn}(h'^T x')\right] \leq \min\left\{ \exp\left(-\frac{l^2(x)k}{8(2+|l(x)|)^2}\right), \frac{1}{kl^2(x)}\right\}$$
Intuition: (A Classification Version of RP)

\[ P[\text{sgn}(h^T x) \neq \text{sgn}(h'^T x')] \leq \exp\left(-\frac{l^2(x)k}{8(2+|l(x)|)^2}\right) \]

Since \(\|h\| = \|x\| = 1\), \(l = h^T x\) \(l' = h'^T x'\)
we have \(\|h-x\|^2 = \|h\|^2 + \|x\|^2 - 2h^T x = 2-2l\)
\(\|h'-x'\|^2 = \|h'\|^2 + \|x'\|^2 - 2l'\)

JL: With probability at least \(1-\exp(c^2 k/8)\)
\((1-c) \|h\|^2 \leq \|h'\|^2 \leq (1+c) \|h\|^2,\)
\((1-c) \|x\|^2 \leq \|x'\|^2 \leq (1+c) \|x\|^2\)
\((1-c) \|h-x\|^2 \leq \|h'-x'\|^2 \leq (1+c) \|h-x\|^2.\)

Can find \(c\) in JL so that \(l\) and \(l'\) have same sign.
Contribution of points to error

\[ u_k(x) = \min \left\{ \exp \left( -\frac{k}{l(x)} \right), \frac{1}{kl^2(x)} \right\} \]

\( k \) increases

Bounds

Details

Snowbird’02

20
Projection Error for a Sample (I)

**Definition (projection error):**
Given a classifier $h$, a sample $S$, and a random matrix $R$, the classification error caused by $R$ is defined by:

$$\text{Err}_{proj}(h, R, S) = \frac{1}{|S|} \sum_{x \in S} I(\text{sign}(h^T x) \neq \text{sign}(h'^T x')).$$

**Lemma:** With probability $> 1 - \delta$ (over the choice of $R$)

The projection error for sample $S$, $|S|=m$ is bounded by:

$$\text{Err}_{proj}(h, R, S) \leq \frac{1}{m\delta} \sum_{i=1}^{m} 3 \exp\left( - \frac{l^2(x)k}{8(2+|l(x)|)^2} \right)$$
Proof idea

- Bound the expectation of the projection error with respect to the choice of the random matrix

\[ E[Err_{\text{proj}}(h, R, S)] \]

- Use Markov inequality
Projection Error for a Sample (II)

Can now establish: The difference between the classification performance on two samples in high dimension is similar to difference in low dimension.

Lemma:

Let $S_1, S_2$ be two samples in $\mathbb{R}^n$, $|S_1| = |S_2| = m$; $S'_1, S'_2$ the projected sets. Then, with probability $P[|\text{Err}(h, S_1) - \text{Err}(h, S_2)| > \varepsilon] < P[|\text{Err}(h', S'_1) - \text{Err}(h', S'_2)| > \rho]$

Where $\rho = \varepsilon - \text{Err}(h, S_1) - \text{Err}(h, S_2)$
Final Bound

Using Vapnik’s doubling trick –
- once on the n dimensional data and
- once on the projected data, can now bound

\[ \Pr[\sup_{h \in H} | \overline{Err}(h) - Err(h, S_1) | > \varepsilon] \]

To yield the final bound.

\[ ERR_D \leq ERR_S + \min_k \{ \hat{u}_k + VC(k, m) \} \]
Analysis

- The expected probability of error for a k-dimensional image of x of distance \( l(x) \) from an n-dimensional hyperplane:

\[
\min \left\{ \exp \left( -\frac{l^2(x)k}{8(2+|l(x)|)^2} \right), \frac{1}{kl^2(x)^k} \right\}
\]

- Given a probability distribution over the instance space, can compute the distribution over the margin:

\[
\int_{x \in D} \min \left\{ \exp \left( -\frac{l^2(x)k}{8(2+|l(x)|)^2} \right), \frac{1}{kl^2(x)^k} \right\}
\]

- E.g., if \( l \sim N(0.3,0.1) \) can compute this analytically
Generalization Bound, $1 \sim N$

Bound dominated by VC component in the projected dimension
Real Data (I)

17,000 context
Sensitive spelling

Projected dimension

Bounds Analysis Snowbird’02
Real Data (II)

RBF kernel face detection
Infinite dimension

Projected dimension

RP error term

Bounds

Analysis

Snowbird’02
Conclusions

- Understanding learning in high dimensional spaces
- Analysis of error based on
  - Prediction preserving projection into low dimension
  - Standard VC argument at low dimension
- Projection profile
  depends on distribution of distance of points to hyperplane
- Gives informative bounds for some real world (very) high dimensional problems

- Algorithmic implications? Better than random proj.?
Puzzle

- Is it really the margin?

- Example: Winnow vs. Perceptron.
  - Perceptron tries to maximize the margin; Winnow does not.
  - Indeed, Winnow’s margin distribution is worse.

- Yet, Winnow performs consistently better.
Puzzle
Comparison

Bounds

Conclusion

Snowbird’02
Real Generalization