On generalization bounds, projection profile, and margin distribution
(Garg, Peled and Roth, 2002)

Presented by Alex Kosolapov

Presentation Outline

- Introduction
- Base definitions and assumptions
- A Margin Distribution based Bound
- Comparison with some other bounds
- Conclusion

Introduction

- Generalization abilities and its dependence on sample complexity
  - Confidence of predictions
  - Understanding generalization
- Relevant for learning in high dimensional spaces

Learning high dimensional data

- High dimensional problems may be constrained in ways that make them lower dimensional problems (but learning is still in the initial, i.e., high dimensional, space)
- For some high dimensional problems generalization may be dependent on lower dimensionality of the problem
- Random projection of sample into lower dimension space preserving distances (Johnson and Lindenstrauss, 1984)

Contribution

- Garg, Har-Peled and Roth (2002):
  - Project sample and linear classifying hypothesis
  - Generalization bounds for linear classifiers in high dimensional space

Presentation Outline

- Introduction
- Base definitions and assumptions
- A Margin Distribution based Bound
- Comparison with some other bounds
- Conclusion

Projection profile

- Projection profile of D $P(D,h) = (a_1(D,h), a_2(D,h), \ldots)$
  - "data dependent, complexity measure for learning"
  - a_k : expected amount of error introduced when h and data are projected into k-dimensions $a_k(D,h) = \int_{x \in D} u(x) dD$
  - $v(x)$: distance between x and classifying hyperplane

$$u(x) = \min\left(3\exp\left(-\frac{(v(x))^2 k}{8(2+|v(x)|)^2}\right), 1\right)$$

Projection profile contd.

- Decreases monotonically
- Tradeoff between dimension and accuracy
- Takes into account distribution of geometric distances from classifier (*margin distribution*)
- Overall performance will depend on
  - Estimation of projection profile
  - "standard" VC component

Definitions

- Classification problem f: R^n -> {1,1}
- S = {(x_1, y_1), …, (x_m, y_m)}
- $h \in R^n$ n dimensional linear classifier
- Assumed to pass through the origin, so
  - For an example x, $\hat{y}(x) = sign(h^T x)$
- Signed distance of x from h: $v(x) = h^T x = sign(v(x))$
- Empirical error: $\hat{E}(h,S) = \frac{1}{m}\sum_i I(\hat{y}(x_i) \neq y_i)$
- Expected error: $\overline{E}(h,S) = E_x[\hat{y}(x) \neq f(x)]$

Random Projection Matrix

- *Random projection matrix:*
  - R is $x \times n$ matrix
  - Each entry is $N(0, 1/k)$
- For $x \in R^n$, projection of x $x' = Rx \in R^k$
- Similar for a classifier h

Presentation Outline

- Introduction
- Base definitions and assumptions
- A Margin Distribution based Bound
- Comparison with some other bounds
- Conclusion

Margin Distribution based Bounds

- Decision of classifier is based on the sign of $v(x) = h^T x = sign(v(x))$
- $|v(x)|$- a geometric distance between x and hyperplane orthogonal to h that passes through the origin
- Given a set of samples with some distribution, induces *margin distribution*

Main Theorem (3.1)

Let $S = \{(x_1, y_1), \ldots, (x_{2m}, y_{2m})\}$ be a set of n-dimensional labeled examples and h a linear classifier. Then, for all constants $0<\delta<1$, $0<k$, with probability at least $1-4\delta$, the ***expected error of h*** is bound by

$$\overline{E} \leq \hat{E}(S,h) + \min_k \left\{ \mu_k + 2\sqrt{\frac{(k+1)\ln\frac{me}{k+1} + \ln\frac{1}{\delta}}{2m}} \right\}$$

$$\mu_k = \frac{6}{m\delta}\sum_{j=1}^{2m} \exp\left(-\frac{v_j^2 k}{8(2+|v_j|)^2}\right) \qquad v_j = v(x_j) = h^T x_j$$

Observations

1. If x is far from h, then projection of x should be far from projection of h
2. Empirical error in projection space: images of datapoints not consistent with image of h
3. Optimal bound: balance between penalty for projection and VC error term in that dimension

- The probability of misclassifying x relative to its classification in original space:

$$P\left[sign\,(h^T x) \neq sign\,(h'^T x')\right] \leq 3\exp\left(-\frac{v^2 k}{8(2+|v|)^2}\right)$$

- Projection error (caused by projection matrix):

$$Err_{proj}(h,R,S) = \frac{1}{|S|}\sum_{x\in S} I\left(sign(h^T x) \neq sign(h'^T x')\right)$$

Bounds on classification error

- With probability $\geq 1-\delta$, the projection error satisfies

$$Err_{proj}(h,R,S) \leq \varepsilon_1(S,\delta) \qquad \varepsilon_1(S,\delta) = \frac{1}{m}\frac{1}{\delta}\sum_{i=1}^{m} 3\exp\left(-\frac{v_i^2 k}{8(2+|v_i|)^2}\right)$$

- Bounds on classification error with probability $\geq 1-4\delta$:

$$\overline{E}(h,S_1) \leq \hat{E}(h,S_1) + \varepsilon$$
$$\leq \hat{E}(h,S_1) + 2\varepsilon_1(S_1,\delta) + 2\varepsilon_1(S_1,\delta) + 2\sqrt{\frac{(k+1)\ln\left(\frac{2em}{k+1}\right) + \ln\frac{1}{\delta}}{2m}}$$

Improved bounds

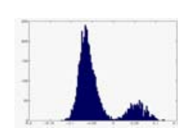
- Important is the distance from classifier
- Expected probability of error for an image of x:

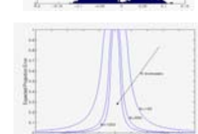
$$\min\left(3\exp\left(-\frac{(v(x)^2)k}{8(2+|(v(x))|)^2}\right), \frac{2}{kv(x)^2}, 1\right)$$

- $$a_k(D,h) = \int_{x\in D} \min\left(3\exp\left(-\frac{(v(x)^2)k}{8(2+|(v(x))|)^2}\right), \frac{2}{kv(x)^2}, 1\right) dD$$
- Also possible to improve if R has entries {-1,+1} (Achlioptas, 2001)

Projection error

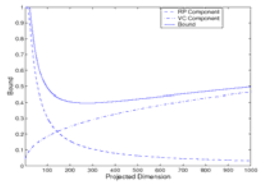
- Histogram of distances from h (for context sensitive spelling correction)
- Contributions of points to generalization error as a function of distance from h





Bounds Tradeoff

- x from Normal distribution with mean 0.3, variance 0.1



VC Bounds

- VC bounds with probability ≥1 δ

$$\varepsilon \le \hat{\varepsilon} + \sqrt{\frac{(n+1)\left(\ln\left(\frac{2m}{n+1}\right)+1\right) - \ln \delta / 4}{m}}$$

  - Worst-case generalization of classifier
  - Depend on the space of the data, independent of the actual data

Bounds via margin

- Deriving bounds via margin:

$$\varepsilon \le \frac{2}{m}\left(f \log_2(32m)\log_2 \frac{8em}{f} + \log_2 \frac{2m}{\delta}\right) \qquad f = afat(\delta / 8)$$

δ – min. margin

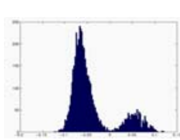
- Linear functions case: $(BR/\delta)^2$
  - B is norm of the classifier
  - R is maximal norm of the data
- Independent of the data space, depends on margin with the given data

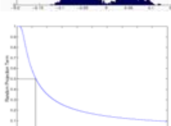
VC, margin based bounds

- Drawback: Large number of observed datapoints before bounds are meaningful (<0.5)
  - i.e., margin-based: need at least 17 times the dimension of a datapoint
    - With 0.9 margin, need about 100,000 datapoints
  - high dimensional data : 17,000 dimensional data (context sensitive spelling correction experiment)
    - VC bounds: 120,000 datapoints before meaningful

Experiment 1

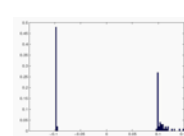
- Context sensitive error correction with winnow based algorithm (Golding and Roth, 1999)
- 17,000 dimensional data

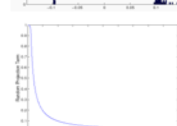




Experiment 2

- Face detection problem
- RBF kernel was used to learn classifier
- No other details





Conclusions

- A new analysis method for linear learning algorithms
- Data dependent complexity measure for learning and bound on error as a function of margin distribution of data relative to the classifier

References

- Achlioptas, D. (2001) Database-friendly random projections. In Symposium on Principles of Database Systems, 274–281.
- Garg, A., Peled, S.H. & Roth, D. (2002) On generalization bounds, projection profile and margin distribution. Proc. of *19th international conference on Machine learning (ICML).*
- Golding, A.R. & Roth, D. (1999) A winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34 (1-3), 107-130

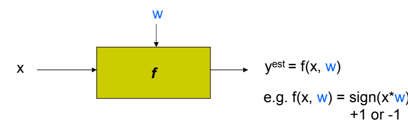
On generalization bounds, projection profile, and margin distribution

Ashutosh Garg, Sariel Har-Peled, Dan Roth

Presented By
Alex Kosolapov

Comments By
Harsh Bapat

VC Dimension



- Different machines have different amounts of "power"
- Tradeoff
  - More power: Can model complex classifiers but might overfit
  - Less power: Do not overfit, but can model simple classifiers
- How do we characterize the amount of power?

Vapnik-Chervonenkis dimension

TrainError(w) = $\frac{1}{R}\sum_{k=1}^{R}\frac{1}{2}\left|y_k - f(x_k, w)\right|$ *TestError(w)* = $E\left[\frac{1}{2}\left|y - f(x,w)\right|\right]$

R is #training data points

•Given a machine *f*, let its VC dimension be *h*
•h is the measure of f's power
•With probability $1-\eta$

$$TestError(w) \le TrainError(w) + \sqrt{\frac{h\log(2R/h)+1)-\log(\eta/4)}{R}}$$

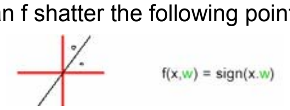
This gives a way to estimate the error on future data based on training error and VC dimension of *f*

How to compute h?

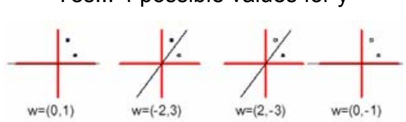
- A machine f can *shatter* a set of points $x_1, x_2, \ldots x_r$ if and only if
  - For every possible training set of the form $(x_1,y_1), (x_2,y_2), \ldots (x_r,y_r)$
  - There exists some value of *w* that gets 0 training error.
- NOTE: There are 2^r such training sets to consider, each with a different combination of +1's and -1's for the y values.

Shattering

Can f shatter the following points?



Yes!!! 4 possible values for y



VC dimension definition

- Given a machine f, the VC dimension h is
  - The maximum number of points that can be arranged so that f shatters them

References

[1] On generalization bounds, projection profile, and margin distribution. Ashutosh Garg, Sariel Har-Peled, Dan Roth.

[2] VC-dimension for characterizing classifiers. Tutorial by Andrew W. Moore, Carnegie Mellon University.

On generalization bounds, projection profile, and margin distributtion

Ashutosh Garg, Sariel Har-Peled, Dan Roth

Comment by: Kai Xu

VC Dimension

- According to the VC theory, a meaningful separating hyper-plane can be found after training by 17n examples.
- However, in most cases, not all attributes affect the classification result.
- Q: How small can we shrink the input dimension?

Margin/Error probability Relationship

- Shawe-Taylor's paper shows there is a relationship between the margin and the error probability.
- The confidence of whether we predict a point correctly can be represent as a function its margin.

Random Projection and Margin Distribution

- This paper proves that
  - the distance distortion can be represent as a function over the projection dimension.
- Thus, given hypothesis h and the dimension of the projection space, we can
  - calculate the error probability for a data example after the projection.

The Main Theorem

- The main theorem (Theorem 3.1) shows the true error probability is bounded by
  - the empirical error probability,
  - plus the sum of
    - The projection penalty, and
    - The VC dimension term.
- We can build the *Projection Profile*, which give us a way to balance between the dimension of the projection space and the accuracy.

Contributions of this paper

- Devise a new linear learning algorithm that uses random projection and margin distribution analysis.
- Pointing out it's possible to reduce the dimension of the training data set while not introducing too much distortion error.
- Giving a way to balance between dimension and accuracy by the projection profile.

References

- Ashutosh Garg, Sariel Har-Peled, Dan Roth , On generalization bounds, projection profile, and margin distribution, Feb. 2, 2002
- A blumer, A Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnick-Chervonenkis dimension. *Journal of ACM, 36(4):929-865*, 1989.
- J. Shawe-Taylor. Classification accuracy based on observed margin. *Algorithmica, 22(1/2):157-172*, 1998