Learning Theory

Ron Parr
CPS 271

With content adapted from Lise Getoor, Tom Dietterich, Andrew Moore & Rich Maclin

What is learning theory?

- Grew from theoretical CS community
- Emphasizes formal results on
  - Amount of data needed
  - Efficiency of algorithm WRT time/data
- Separate community from "practical learning"
- COLT (computational learning theory conference)
- Practical and theoretical influencing each other (Who'd have thought??? ☺)

Motivation

- Originally learning theory was concerned with theories of what was "learnable"
- Different assumptions about models
  - Adversarial
  - Oracle
- Very little turned out to be "learnable" ☹
- PAC learnability more reasonable
  - Probably Approximately Correct
  - Draw training, testing samples from same distribution
  - Try to establish WHP bounds
  - Embodied in current practice

Bias & Variance Review

- Example: Regression
- Suppose we draw m samples from an infinite supply of training data
- What is the right hypothesis space?
  - Linear?
  - Quadratic?
  - Etc?
- What should answer depend on?
  - Background knowledge?
  - Size of m?

Bias

- We (might) want:

$$\lim_{|D|\to\infty} \{E_D[y(\mathbf{x};D) - h(\mathbf{x})]\}^2 = 0$$

- We "eventually get it right" w/enough data
- Otherwise we are said to have bias
- Is bias always bad???

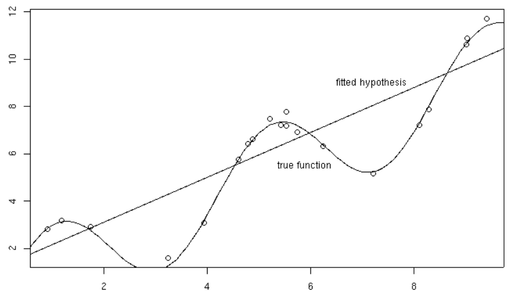
Variance

- We would like (and usually get):

$$\lim_{|D|\to\infty} E_D\left[\{ y(\mathbf{x};D) - E_D[y(\mathbf{x};D)]\}^2\right] = 0$$

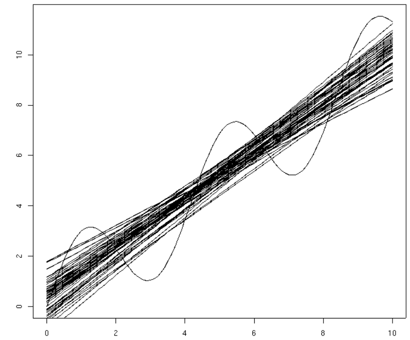
- Compares performance on training set against other draws of same sized set
- Problem: m is finite

Example: 20 points
$y = x + 2 \sin(1.5x) + N(0,0.2)$



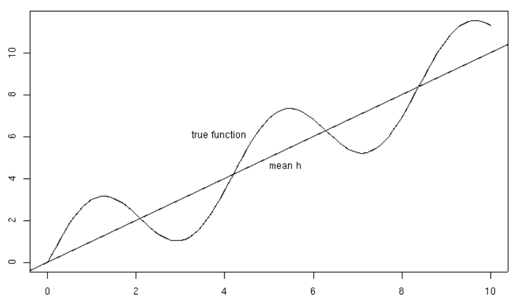
Hypothesis space = linear in x

50 fits (20 examples each)

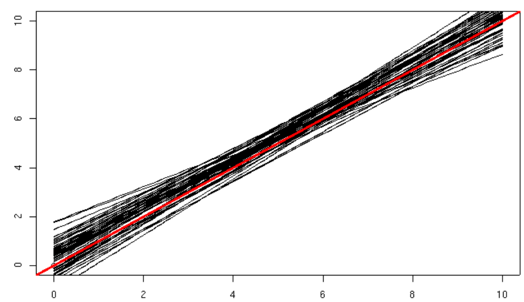


What are we seeing here?

Bias



Variance



Dealing with Bias & Variance

- Real data sets are finite
- Means that bias and variance are positive
- Can we trade one against another?
- Example:
  - Suppose data come from line + noise
  - m=3
  - What is best H?
    - Constants (bias, moderate variance)
    - Lines (no bias, higher variance)

Bias & Variance with real data

- In the real world:
  - Don't know source characteristics
  - Choosing a "fancier" H risks high variance
  - Higher variance=
    - Overfitting
    - Fitting noise
- When can we risk a big H?
- COLT: Theoretical bounds (for discrete cases)
- Practical techniques later
  (not mutually exclusive with COLT!)

Tools of Learning Theory I

- Union bound, for events $e_1 \ldots e_k$

$$P(e_1 \vee e_2 \vee \ldots \vee e_k) \leq \sum_{i=1}^{k} P(e_i)$$

- (Trivial consequence of axioms of prob. theory)

Tools of Learning Theory II

- Let $\hat{\theta}$ be mean of m IID samples of a Bernouli RV w.p. θ (e.g. coin flip)
- Chernoff bound (Hoeffding inequality):

$$P(|\theta - \hat{\theta}| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

- Not a trivial result
- Error drops off:
  - Exponentially in γ^2
  - Exponentially in m

Empirical Risk

- Empirical risk for hypothesis h on D
  (= error on D):

$$\hat{\varepsilon}(y) = \underset{x \in D}{E}\, P(t \neq y(x))$$

- Many learning algorithms are empirical risk minimizers (ML, SSE minimization)

$$\hat{y} = \arg\min_{y \in H} \hat{\varepsilon}(y)$$

Evaluating Hypotheses

- Treat each datum as a test of y_i
- How reliable is $\hat{\varepsilon}(y_i)$?
- IOW: How much do we trust our empirical estimate of the quality of y_i?
- Use Chernoff bound:

$$P(|\hat{\varepsilon}(y_i) - \varepsilon(y_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

Evaluating our learner

- Suppose H is finite
- Learner picks “best” y, so all estimates must be “good”
- What is probability of getting a “bad” estimate:

$$P(\exists y_i \in H s.t. |\hat{\varepsilon}(y_i) - \varepsilon(y_i)| > \gamma) = P(|\hat{\varepsilon}(y_1) - \varepsilon(y_1)| > \gamma \vee \ldots \vee |\hat{\varepsilon}(y_k) - \varepsilon(y_k)| > \gamma)$$
$$\leq \sum_{i=1}^{k} P(|\hat{\varepsilon}(y_i) - \varepsilon(y_i)| > \gamma)$$
$$\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m)$$
$$= 2k\exp(-2\gamma^2 m)$$

How much data???

- If all quality estimates are “good”, then when can we trust that real risk = empirical risk???
- Suppose we want to guarantee answer w.p. 1-δ

$$1 - \delta \geq 1 - 2k\exp(-2\gamma^2 m)$$
$$m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

- “Sample Complexity” of our learner

How much trust?

- Solve for γ
- WP 1-δ

$$|\hat{\varepsilon}(y_i)-\varepsilon(y_i)|\le\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

- Note log dependence on k!

Trust in our choice

- Suppose y* is "best" in H
- We pick something else b/c of finite m

$$\begin{aligned}\varepsilon(\hat{y}) &\le \hat{\varepsilon}(\hat{y})+\gamma \\ &\le \hat{\varepsilon}(y^*)+\gamma \quad \text{(Since we didn't pick y*)} \\ &\le \varepsilon(y^*)+\gamma+\gamma \\ &\le \varepsilon(y^*)+2\gamma\end{aligned}$$

- Even if we didn't pick the best y*, we still didn't do that badly

Putting it all together

- Suppose |H|=k
- Fix δ, γ
- To achieve real performance within 2 γ

$$m\ge O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right)$$

Putting it all Together II

- Learning theory bounds performance on training set as function of performance on test set

$$\varepsilon(\hat{y})\le\hat{\varepsilon}(\hat{y})+\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

- Assuming |H|=k, WP 1-δ
- Log dependence on k

Continuous Spaces

- So far, we have assumed H is finite
- Most algorithms we have studied are smoothly parameterized
  - Perceptron
  - Logistic regression
  - Etc.
- How do these results generalize?

First Cut

- Suppose we have n finite precision numbers
- Use b bits to represent each parameter
- |K| = 2^{bn} (Uh oh…)
- But, log dependence on k saves us:

$$m\ge O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right) \qquad \varepsilon(\hat{h})\le\hat{\varepsilon}(\hat{h})+\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

- Sample complexity linear in n
- Performance bound linear in sqrt(n)

Where bits counting fails

- Suppose we have a perceptron with n inputs
- Duplicating input doesn't change things
  (no increased risk of overfitting)
- Does add one more continuous parameter
- If we're counting bits, for our bound:
  - Leads to double counting
  - Gratuitously loose bounds

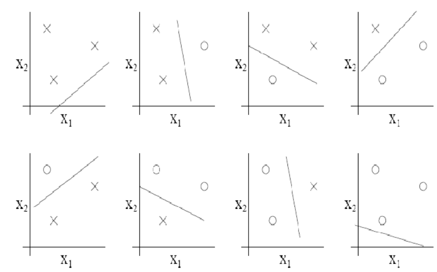
Shattering

- What we need:
  - Way of capturing intrinsic power of classifier
  - Independent of parameterization
- Step 1: "shattering"
- Given set of training data D
- H shatters D if H can correctly classify all possible labelings of D

VC Dimension

- VC = Vapnik-Chervonenkis
- VC(H) = size of largest D shattered by H
- Note quantification:
  - Existence of a single set at given size satisfies
  - Proof typically requires demonstrating impossibility of shattering large sets
- VC(H) can be infinite (nearest neighbor)

Shattering with planes



Can correctly classify all possible labelings of 3 points!

VC Dimension of hyperplanes

- Our example generalizes to d dimensions
- For H = d dimensional hyperplanes
  - Can shatter |D|=d+1
  - Cannot shatter |D|=d+2 (e.g. XOR)
  - VC(H) = d+1

VC Theory - Performance

- Suppose k=VC(H), WP 1-δ

$$\varepsilon(\hat{y}) \le \hat{\varepsilon}(\hat{y}) + O\left(\sqrt{\frac{k}{m}\log\frac{m}{k} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

- Compare with finite case, k=|H|

$$\varepsilon(\hat{y}) \le \hat{\varepsilon}(\hat{y}) + \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

- Remember for n finite precision parameters $k=2^{bn}$

VC Theory – Sample Complexity

- Suppose VC(H)=k, fix δ, γ
- To achieve real performance within 2 γ
- Need O(k) samples
- Compare with finite case:

$$m \geq O(\frac{1}{\gamma^2} \log \frac{k}{\delta})$$

- $k=2^{bn}$ – linear dependence on n

Continuous Hypothesis Spaces Conclusion

- "Natural" parameterization finite set of hypotheses (due to finite precision) leads to linear sample complexity in number of parameters
- VC Theory:
  - Cleaner, more general theory
  - Typically gives similar bounds
- Learning theory bounds:
  - Sometimes loose
  - Sometimes more qualitative than quantitative

Learning Theory Conclusions

- COLT helps us quantify:
  - Power of a hypothesis space
  - How much data we need for given level of trust
- What COLT doesn't do:
  - Tell us to search space of hypotheses
  - How to improve our performance
- In practice:
  - COLT bounds tend to be loose
  - Not a substitute for empirical validation
  - Gives good high level guidance