Review: Generalization error in finite hypothesis spaces [Haussler ’88]

**Theorem**: Hypothesis space $H$ finite, dataset $D$ with $m$ i.i.d. samples, $0 < \varepsilon < 1$ : for any learned hypothesis $h$ that is consistent on the training data:

$$P(\text{error}_\mathcal{X}(h) > \varepsilon) \leq |H|e^{-me\varepsilon}$$

Consistent with $D \Rightarrow \text{Error}_D(h) = 0$

*Zero errors in test set*

$\text{error}_\mathcal{X}(h) \rightarrow$ expected error $x \in \mathcal{X}$

Even if $h$ makes zero errors in training data, may make errors in test
Using a PAC bound

Typically, 2 use cases:

1. Pick $\varepsilon$ and $\delta$, give you $m$
2. Pick $m$ and $\delta$, give you $\varepsilon$

\[ P(\text{error}_x(h) > \varepsilon) \leq |H|e^{-m\varepsilon} \]

\[ \log |H|e^{-m\varepsilon} \leq \delta, \text{ log on both sides} \]

\[ \ln |H| - m\varepsilon \leq \ln \delta \]

\[ m \geq \frac{1}{\varepsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]

\[ \varepsilon \geq \ln |H| + \ln \frac{1}{\delta} \]

\[ \text{error}_x(h) \leq \frac{\ln |H| + \ln \frac{1}{\delta}}{m} \]

with prob. at least $1 - \delta$
Limitations of Haussler ‘88 bound

- Consistent classifier
  \[ P(\text{error}_{\mathcal{X}}(h) > \epsilon) \leq |H|e^{-m\epsilon} \]
  - Zero training error!

- Size of hypothesis space
  \[ |H| \]
  - what if it's too large
    - continuous
What if our classifier does not have zero error on the training data?

- A learner with zero training errors may make mistakes in test set.
- A learner with $\text{error}_D(h)$ in training set, may make even more mistakes in test set.

$\text{error}_x(h)$ relates $\text{error}_D(h)$?
Simpler question: What’s the expected error of a hypothesis?

- The error of a hypothesis is like estimating the parameter of a coin!

- Chernoff bound: for \( m \) i.d.d. coin flips, \( x_1, \ldots, x_m \), where \( x_i \in \{0, 1\} \). For \( 0<\varepsilon<1 \):

\[
P \left( \theta - \frac{1}{m} \sum_{i} x_i > \varepsilon \right) \leq e^{-2m\varepsilon^2}
\]
Using Chernoff bound to estimate error of a single hypothesis

\[ P \left( \theta - \frac{1}{m} \sum_{i} x_i > \epsilon \right) \leq e^{-2m\epsilon^2} \]

Given hypothesis h, how well will it do on test data?

\[
\text{error}_X(h) \equiv \Theta \\
\text{error}_D(h) \equiv \frac{1}{m} \sum_{i} x_i \\
P(\text{error}_X(h) - \text{error}_D(h) > \epsilon) \leq e^{-2m\epsilon^2}
\]
But we are comparing many hypothesis: **Union bound**

For each hypothesis $h_i$:

$$P(\text{error}_X(h_i) - \text{error}_D(h_i) > \epsilon) \leq e^{-2m\epsilon^2}$$

What if I am comparing two hypothesis, $h_1$ and $h_2$?

Choose $h_1$, because $\text{error}_D(h_1) \leq \text{error}_D(h_2)$.
Theorem: Hypothesis space $H$ finite, dataset $D$ with $m$ i.i.d. samples, $0 < \epsilon < 1$ : for any learned hypothesis $h$:

$$P (\text{error}_X(h) - \text{error}_D(h) > \epsilon) \leq |H|e^{-2me^2}$$
PAC bound and Bias-Variance tradeoff

\[ P(\text{error}_x(h) - \text{error}_D(h) > \epsilon) \leq |H|e^{-2m\epsilon^2} \]

or, after moving some terms around, with probability at least 1-\(\delta\):

\[ \text{error}_x(h) \leq \text{error}_D(h) + \sqrt{\frac{\text{Var} + \text{bias}}{2m} \ln \frac{|H| + 1}{\delta}} \]

- Important: PAC bound holds for all \(h\), but doesn’t guarantee that algorithm finds best \(h\)!!!
What about the size of the hypothesis space?

\[ m \geq \frac{1}{2\epsilon^2} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]

- How large is the hypothesis space?

|H| is large \( \Rightarrow \) need many training examples
Boolean formulas with $n$ binary features

$$m \geq \frac{1}{2\epsilon^2} \left( \ln |H| + \ln \frac{1}{\delta} \right)$$

- Look up table:
  - $x_1, \ldots, x_n, y$
  - $\ln |H| = \Theta(2^n)$
  - $|H| = \sqrt{2^n}$

- Conjunctions:
  - $\langle 1, 0, ?, ?, ?, 1, \ldots \rangle$
  - $|H| = 3^n$
  - $\ln |H| = O(n)$

- Pretty good:
  - Look up table for $k$ conjunctions:
  - $|H| = 2^k \cdot 3^{n-k}$
  - $\ln |H| = O(2^k + (n-k))$

- Grow fast with $k$
Number of decision trees of depth $k$

Recursive solution

Given $n$ attributes

$H_k =$ Number of decision trees of depth $k$

$H_0 = 2$

$H_{k+1} =$ (#choices of root attribute) *

(# possible left subtrees) *

(# possible right subtrees)

$= n \times H_k \times H_k$

Write $L_k = \log_2 H_k$

$L_0 = 1$

$L_{k+1} = \log_2 n + 2L_k$

So $L_k = (2^k-1)(1+\log_2 n) + 1$

$m \geq \frac{1}{2e^2} \left( \ln |H| + \ln \frac{1}{\delta} \right)$
PAC bound for decision trees of depth $k$

$$m \geq \frac{\ln 2}{2\epsilon^2} \left( (2^k - 1)(1 + \log_2 n) + 1 + \ln \frac{1}{\delta} \right)$$

- Bad!!!
  - Number of points is exponential in depth!

- But, for $m$ data points, decision tree can’t get too big…
  - only reach $m$ leaves

---

Number of leaves never more than number data points
Number of decision trees with k leaves

\[ H_k = \text{Number of decision trees with } k \text{ leaves} \]
\[ H_0 = 2 \]

\[ m \geq \frac{1}{2\epsilon^2} \left( \ln |H| + \ln \frac{1}{\delta} \right) \]

Loose bound:

\[ H_k \leq n^{k-1}(k+1)^{2k-1} \]

Reminder:

\[ |\text{DTs depth } k| = 2 \times (2n)^{2^k-1} \]

\[ \ln |H| = O(nk^2) \]
\[ \ln |\text{DTs depth } k| = O(2^k n) \]

a lot better
PAC bound for decision trees with $k$ leaves – Bias-Variance revisited

\[ H_k = n^{k-1}(k + 1)^{2k-1} \]

\[ \text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}} \]

\[ \text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{(k - 1) \ln n + (2k - 1) \ln (k + 1) + \ln \frac{1}{\delta}}{2m}} \]

Suppose $k = m$

\[ \emptyset \]

If $k = \alpha m$

\[ \alpha < 1 \]

\[ > 0 \]

\[ \uparrow \text{really big} \]

\[ \downarrow \text{smaller} \]
What did we learn from decision trees?

- Bias-Variance tradeoff formalized

\[
\text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{(k - 1) \ln n + (2k - 1) \ln(k + 1) + \ln \frac{1}{\delta}}{2m}}
\]

- Moral of the story:

Complexity of learning not measured in terms of size hypothesis space, but in maximum *number of points* that allows consistent classification

- Complexity \( m \) – no bias, lots of variance
- Lower than \( m \) – some bias, less variance
What about continuous hypothesis spaces?

\[ \text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}} \]

- Continuous hypothesis space:
  - \(|H| = \infty\)
  - Infinite variance???

- As with decision trees, only care about the maximum number of points that can be classified exactly!
How many points can a linear boundary classify exactly? (1-D)
How many points can a linear boundary classify exactly? (2-D)

+ → yes

+ → yes

- → yes

complexity 3

+ -
- + no!
How many points can a linear boundary classify exactly? (d-D)

\[ wx + b = 0 \]

\[ -wx + b < 0 \]

\[ d + 1 \]

\[ d+1 \text{ variables need } d+1 \text{ constraints} \]

\[ \Rightarrow d+1 \text{ points} \]
PAC bound using VC dimension

- Number of training points that can be classified exactly is VC dimension!!!
  - Measures relevant size of hypothesis space, as with decision trees with k leaves

\[
\text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}}
\]
Shattering a set of points

Definition: a dichotomy of a set $S$ is a partition of $S$ into two disjoint subsets.

Definition: a set of instances $S$ is shattered by hypothesis space $H$ if and only if for every dichotomy of $S$ there exists some hypothesis in $H$ consistent with this dichotomy.

Question: classify exactly
VC dimension

Definition: The *Vapnik-Chervonenkis dimension*, $VC(H)$, of hypothesis space $H$ defined over instance space $X$ is the size of the largest finite subset of $X$ shattered by $H$. If arbitrarily large finite sets of $X$ can be shattered by $H$, then $VC(H) \equiv \infty$.
Examples of VC dimension

- **Linear classifiers:**
  - \( VC(H) = d+1 \), for \( d \) features plus constant term \( b \)

- **Neural networks**
  - \( VC(H) = \#\text{parameters} \)
  - Local minima means NNs will probably not find best parameters

- **1-Nearest neighbor?**
  - \( VC(1-\text{NN}) = \infty \)
PAC bound for SVMs

- SVMs use a linear classifier
  - For $d$ features, $VC(H) = d+1$:

$$\text{error}_X(h) \leq \text{error}_D(h) + \sqrt{(d + 1) \left( \ln \frac{2m}{d+1} + 1 \right) + \ln \frac{4}{\delta}}$$
VC dimension and SVMs: Problems!!!

Doesn’t take margin into account

\[ \text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{(d + 1) \left( \ln \frac{2m}{d+1} + 1 \right) + \ln \frac{4}{\delta}}{m}} \]

- What about kernels?
  - Polynomials: num. features grows really fast = Bad bound
    
    \[ \text{num. terms} = \binom{p+n-1}{p} = \frac{(p+n-1)!}{p!(n-1)!} \]
    
    \( n \) – input features
    \( p \) – degree of polynomial

  - Gaussian kernels can classify any set of points exactly
Margin-based VC dimension

- **H**: Class of linear classifiers: $w \cdot \Phi(x)$ (b=0)
  - Canonical form: $\min_j |w \cdot \Phi(x_j)| = 1$
  - $\text{VC}(H) = R^2 \cdot w \cdot w$
    - Doesn’t depend on number of features!!!
    - $R^2 = \max_j \Phi(x_j) \cdot \Phi(x_j)$ – magnitude of data
    - $R^2$ is bounded even for Gaussian kernels $\rightarrow$ bounded VC dimension

- Large margin, low $w \cdot w$, low VC dimension – Very cool!
Applying margin VC to SVMs?

\[ \text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right) + \ln \frac{4}{\delta}}{m}} \]

- \( VC(H) = R^2 \mathbf{w} \mathbf{w} \)
  - \( R^2 = \max_j \Phi(x_j) \cdot \Phi(x_j) \) – magnitude of data, doesn’t depend on choice of \( \mathbf{w} \)
- SVMs minimize \( \mathbf{w} \cdot \mathbf{w} \)

- SVMs minimize VC dimension to get best bound?
  - Not quite right: 😞
    - Bound assumes VC dimension chosen before looking at data
    - Would require union bound over infinite number of possible VC dimensions…
    - But, it can be fixed!
Structural risk minimization theorem

\[
\text{error}_x(h) \leq \text{error}_D^\gamma(h) + C \sqrt{\frac{R^2}{\gamma^2} \ln m + \ln \frac{1}{\delta}}
\]

as \( \delta \rarr 0 \) variance goes down

\[
\text{error}_D^\gamma(h) = \text{num. points with margin} < \gamma \leq \frac{1}{2}
\]

more training errors

- For a family of hyperplanes with margin \( \gamma > 0 \)
  - \( \mathbf{w} \cdot \mathbf{w} \leq 1 \)
- SVMs maximize margin \( \gamma \) + hinge loss
  - Optimize tradeoff between training error (bias) versus margin \( \gamma \) (variance)
Reality check – Bounds are loose

Bound can be very loose, why should you care?

- There are tighter, albeit more complicated, bounds
- Bounds gives us formal guarantees that empirical studies can’t provide
- Bounds give us intuition about complexity of problems and convergence rate of algorithms

\[
\text{error}_X(h) \leq \text{error}_D(h) + \sqrt{\frac{(d + 1) \left( \ln \frac{2m}{d+1} + 1 \right) + \ln \frac{4}{\epsilon}}{m}}
\]
What you need to know

- Finite hypothesis space
  - Derive results
  - Counting number of hypothesis
  - Mistakes on Training data

- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – decision trees
  - Infinite case – VC dimension

- Bias-Variance tradeoff in learning theory
- Margin-based bound for SVM
- Remember: will your algorithm find best classifier?