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:

$$P(\text{error}_{true}(h) > \varepsilon) \leq |H|e^{-m\varepsilon}$$

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 $m$

$P(\text{error}_{\text{true}}(h) \geq \varepsilon) \leq |H|e^{-m\varepsilon}$

2: Pick $m$ and $\delta$, give you $\varepsilon$

Limitations of Haussler ‘88 bound

- Consistent classifier

$P(\text{error}_{\text{true}}(h) \geq \varepsilon) \leq |H|e^{-m\varepsilon}$

- Size of hypothesis space

when is $|H|$ too big

$|H|$ when $H$ is continuous
PAC bound and Bias-Variance tradeoff

$$P(\text{error}_{true}(h) - \text{error}_{\text{train}}(h) > \epsilon) \leq |H|e^{-2m\epsilon^2}$$

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

$$\text{error}_{true}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

- Important: PAC bound holds for all \(h\), but doesn’t guarantee that algorithm finds best \(h\)!!

---

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…

Number of leaves never more than number data points
PAC bound for decision trees with k leaves – Bias-Variance revisited

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

\[ \text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{(k-1)\ln n + (2k-1)\ln(k+1) + \ln \frac{1}{\delta}} \]

What did we learn from decision trees?

- Bias-Variance tradeoff formalized

- 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?

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

- 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)

How many points can a linear boundary classify exactly? (d-D)
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

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{V\text{C}(H)}{m} \left( \ln \frac{2m}{\delta} + 1 \right) + \ln \frac{4}{\delta}}$$

---

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.
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 $VC(H) \equiv \infty$.

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
  - Bound for infinite dimension hypothesis spaces:

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{VC(H) \left( \ln \frac{2m}{VC(H)} + 1 \right)}{m} + \ln \frac{4}{\delta}}$$
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?

Another VC dim. example - What can we shatter?

- What’s the VC dim. of decision stumps in 2d?
Another VC dim. example - What can’t we shatter?

- What’s the VC dim. of decision stumps in 2d?

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

- Remember: will your algorithm find best classifier?
Handwriting recognition

Character recognition, e.g., kernel SVMs
Webpage classification

Company home page vs Personal home page vs University home page vs ...

Handwriting recognition 2
Today – Bayesian networks

- One of the most exciting advancements in statistical AI in the last 10-15 years
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies
Causal structure

- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?

Possible queries

- Inference
- Most probable explanation
- Active data collection
Car starts BN

- 18 binary attributes
- Inference
  - $P(\text{BatteryAge}|\text{Starts}=f)$

- $2^{16}$ terms, why so fast?
- Not impressed?
  - HailFinder BN – more than $3^{54} = 58149737003040059690390169$ terms

Factored joint distribution - Preview

- Flu
- Allergy
- Sinus
- Headache
- Nose
Number of parameters

Key: Independence assumptions

Knowing sinus separates the variables from each other
(Marginal) Independence

- Flu and Allergy are (marginally) independent

<table>
<thead>
<tr>
<th></th>
<th>Flu = t</th>
<th>Flu = f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy = t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergy = f</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- More Generally:

<table>
<thead>
<tr>
<th></th>
<th>Flu = t</th>
<th>Flu = f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy = t</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergy = f</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Marginal independence: $P(X \perp Y)$

- **Sets** of variables $X, Y$
- $X$ is independent of $Y$ if
  - $P \vdash (X=x \perp Y=y), \ \forall x \in \text{Val}(X), \ y \in \text{Val}(Y)$

- Shorthand:
  - **Marginal independence:** $P \vdash (X \perp Y)$

- **Proposition:** $P$ satisfies $(X \perp Y)$ if and only if
  - $P(X,Y) = P(X) \cdot P(Y)$
Conditional independence

Flu and Headache are not (marginally) independent

Flu and Headache are independent given Sinus infection

More Generally:

Conditionally independent random variables

Sets of variables X, Y, Z

X is independent of Y given Z if

- $P \models (X=x \perp Y=y | Z=z)$, $\forall x \in \text{Val}(X), y \in \text{Val}(Y), z \in \text{Val}(Z)$

Shorthand:

- **Conditional independence**: $P \models (X \perp Y | Z)$
- For $P \models (X \perp Y | \emptyset)$, write $P \models (X \perp Y)$

**Proposition**: $P$ satisfies $(X \perp Y | Z)$ if and only if

- $P(X,Y|Z) = P(X|Z) \cdot P(Y|Z)$
Properties of independence

- **Symmetry:**
  - $(X \perp Y \mid Z) \Rightarrow (Y \perp X \mid Z)$

- **Decomposition:**
  - $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)$

- **Weak union:**
  - $(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)$

- **Contraction:**
  - $(X \perp W \mid Y, Z) \land (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)$

- **Intersection:**
  - $(X \perp Y \mid W, Z) \land (X \perp W \mid Y, Z) \Rightarrow (X \perp Y, W \mid Z)$
  - Only for positive distributions!
  - $P(\alpha)>0, \forall \alpha, \alpha>0$;

---

The independence assumption

**Local Markov Assumption:**
A variable $X$ is independent of its non-descendants given its parents.
Explaining away

Flu  Allergy  Sinus  Headache  Nose

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents

Naïve Bayes revisited

Flu  Allergy  Sinus  Headache  Nose

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents
What about probabilities?
Conditional probability tables (CPTs)

Joint distribution

Why can we decompose? Markov Assumption!
The chain rule of probabilities

- \( P(A,B) = P(A)P(B|A) \)

More generally:

- \( P(X_1,\ldots,X_n) = P(X_1)P(X_2|X_1)\ldots P(X_n|X_1,\ldots,X_{n-1}) \)

Chain rule & Joint distribution

**Local Markov Assumption:**
A variable \( X \) is independent of its non-descendants given its parents
Two (trivial) special cases

- Edgeless graph
- Fully-connected graph

The Representation Theorem – Joint Distribution to BN

BN: Encodes independence assumptions

If conditional independencies in BN are subset of conditional independencies in $P$

Joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Pa}_{X_i})$$