#### PAC-Learning VC-Dimension Errors While Training

#### PAC-Learning VC-Dimension Errors While Training

#### Theoretical Considerations

• What might Fail?

#### 2 PAC-Learning

- Consistent Learners
- Number of Training Examples
- Learning Conjunctions
- Unbiased Learning

#### 3 VC-Dimension

- Example
- Complexity Measure

#### 4 Errors While Training

- Find-S
- List-then-Eliminate

Örjan Ekeberg Machine Learning

Learning Theory

vetical Considerations PAC-Learning VC-Dimension Errors While Training

# Theoretical Considerations What might Fail?

### 2 PAC-Learning

- Consistent Learners
- Number of Training Examples
- Learning Conjunctions
- Unbiased Learning

#### 3 VC-Dimension

- Example
- Complexity Measure

#### 4 Errors While Training

- Find-S
- List-then-Eliminate

Questions suitable for Theoretical Analysis

- How hard is a given learning task?
- How many training examples are needed?

Örjan Ekeberg

tical Considerations

Errors While Training

VC-Dimension

• How many errors should we expect during and after training?

Machine Learning

• How large is the risk of failing to learn?



Assumptions:

- Concept Learning
- $\bullet\,$  Training and test data from same distribution  ${\cal D}\,$

What kind of errors can occur?

• The result of leaning can be bad The resulting hypothesis makes too many errors

Örjan Ekeberg

• Learning itself can fail The learning algorithm may not find any reasonable hypothesis

Machine Learning

True Error

The probability that a given hypothesis gives the wrong answer

 $\operatorname{error}_{\mathcal{D}}(h) \equiv P_{x \in \mathcal{D}} \left[ h(x) \neq c(x) \right]$ 

How bad hypotheses are we prepared to accept?

Approximately Correct

A hypothesis h is called approximately correct if

 $\operatorname{error}_{\mathcal{D}}(h) < \epsilon$ 

Örjan Ekeberg Machine Learning

Theoretical Considerations PAC-Learning VC-Dimension Errors While Training

Quantification of the risk that learning does not find an approximately correct hypothesis

$$P_L[\operatorname{error}_{\mathcal{D}}(h) \geq \epsilon]$$

How often is it acceptable for learning to fail?

Probably Succeeds

The algorithm L is said to probably find a solution if

 $P_L[\operatorname{error}_{\mathcal{D}}(h) \geq \epsilon] < \delta$ 

 Theoretical Considerations PAC-Learning Errors While Training
 Theoretical Considerations

 What might Fail?

 PAC-Learning

 Consistent Learners
 What might Fail?

 PAC-Learning

 Consistent Learners
 What might Fail?

- Number of Training Examples
- Learning Conjunctions
- Unbiased Learning

#### 3 VC-Dimension

- Example
- Complexity Measure
- Errors While Training
  - Find-S
  - List-then-Eliminate

# PAC-learning

Probably Approximately Correct

#### Given

- C the concept to learn
- $\epsilon$  limit on the error
- $\delta$  limit on the risk
- n size of the examples

PAC-learnable: Time to find a solution grows polynomially with respect to size(*C*), *n*,  $\frac{1}{\epsilon}$  and  $\frac{1}{\delta}$ 

Örjan Ekeberg

PAC-Learning

Errors While Training

#### Analysis of a Consistent Learner

- Assumtion: no errors in training examples
- $\bullet\,$  Examples are drawn from the distribution  ${\cal D}$
- The solution is consistent with all training examples
- "Dangerous Hypotheses":

 $\operatorname{error}_{\mathcal{D}}(h) \geq \epsilon$ 

We do not want learning to produce a dangerous hypothesis!

How large is the risk that a dangerous hypothesis is consistent with all training examples?

Örjan Ekeberg Machine Learning

Theoretical Considerations **PAC-Learning** VC-Dimension Errors While Training

ns Consistent Learners ng Number of Training Example on Learning Conjunctions ng Unbiased Learning

Machine Learning

• Probability that one hypothesis *h* is contradicted by one example

#### $\operatorname{error}_{\mathcal{D}}(h)$

• Probability that *h* is not contradicted

#### $1 - \operatorname{error}_{\mathcal{D}}(h)$

Risk that a dangerous hypothesis (error<sub>D</sub>(h) ≥ ε) is not contradicted by a randomly drawn example

$$\leq (1-\epsilon)$$

oretical Considerations Consistent Learners PAC-Learning Number of Training Examples VC-Dimension Learning Conjunctions Errors While Training Unbiased Learning

• Risk that a *dangerous hypothesis* is not contradicted by one randomly drawn example

$$\leq (1-\epsilon)$$

• Risk that a *dangerous hypothesis* is not contradicted my *m* randomly drawn examples

 $\leq (1-\epsilon)^m$ 

• How large is the risk that any dangerous hypothesis in *H* happens to be consistent with all examples:

```
\leq |H| \cdot (1 - \epsilon)^m
\leq |H| \cdot e^{-\epsilon m}
```



PAC-Learning VC-Dimension Errors While Training

How many training examples are needed?

How many examples *m* are needed to make the risk of ending up with a dangerous hypothesis less than  $\delta$ ?

$$\delta \geq |\mathbf{H}| \cdot \mathbf{e}^{-\epsilon \mathbf{m}}$$

$$e^{\epsilon m} \ge rac{|H|}{\delta}$$
 $m \ge rac{1}{\epsilon} \left[ \ln |H| + \ln rac{1}{\delta} 
ight]$ 

Important relation:

$$m \geq rac{1}{\epsilon} \left[ \ln |\mathcal{H}| + \ln rac{1}{\delta} 
ight]$$

Is this PAC-learnable?

Potential problem: |H| might be too large

Örjan Ekeberg Machine Learning

Theoretical Considerations PAC-Learning VC-Dimension Errors While Training Unbiased Learning Unbiased Learning

Machine Learning

Örjan Ekeberg

#### Learning Conjunctions

Example: Sunny  $\land \neg$  Windy  $\land$  Humid *n* attributes  $\Rightarrow 3^n$  possible concepts  $\Rightarrow |H| = 3^n$ 

$$m \geq \frac{1}{\epsilon} \left[ n \ln 3 + \ln \frac{1}{\delta} \right]$$

• Linear w.r.t.  $\frac{1}{\epsilon}$ 

- Linear w.r.t. n
- Logarithmic w.r.t.  $\frac{1}{\delta}$

#### Seems PAC-learnable!

Further, *time for each example* must be polynomial. Find-S: Ok

| Theoretical Considerations<br>PAC-Learning | Consistent Learners<br>Number of Training Examples |
|--|--|
| VC-Dimension                               | Learning Conjunctions                              |
| Errors While Training                      | Unbiased Learning                                  |

Unbiased Learning All subsets of X are hypotheses

 $|X| = 2^{n}$ 

 $|H|=2^{2^n}$ 

$$m \geq \frac{1}{\epsilon} \left[ 2^n \ln 2 + \ln \frac{1}{\delta} \right]$$

Not PAC-learnable! However, this estimate is an upper bound We have not proven that *m* actually grows exponentially w.r.t. *n* However, in this case it *is* true

| Theoretical Considerations<br>PAC-Learning<br>VC-Dimension<br>Errors While Training   | Theoretical Considerations<br>PAC-Learning<br>VC-Dimension<br>Errors While Training   |
|---|---|
| <ul> <li>Theoretical Considerations <ul> <li>What might Fail?</li> </ul> </li> <li>PAC-Learning <ul> <li>Consistent Learners</li> <li>Number of Training Examples</li> <li>Learning Conjunctions</li> <li>Unbiased Learning</li> </ul> </li> <li>VC-Dimension <ul> <li>Example</li> <li>Complexity Measure</li> </ul> </li> <li>Frrors While Training <ul> <li>Find-S</li> <li>List-then-Eliminate</li> </ul> </li> </ul> | Problem with $ H $<br>• Gives too pessimistic estimates<br>• Can't be used when $ H  = \infty$<br>Vapnik — Chervonenkis observation:<br>The important thing is not the <i>number of</i> hypotheses,<br>but how they can form subsets in X |
| Örjan Ekeberg Machine Learning  | Örjan Ekeberg Machine Learning  |
| Theoretical Considerations<br>PAC-Learning<br>VC-Dimension<br>Errors While Training<br>Errors While Training  | Theoretical Considerations<br>PAC-Learning<br>VC-Dimension<br>Errors While Training<br>Errors While Training  |
| ScatteringA finite set S is scattered by the hypotheses H if every subset of S<br>is described by a $h \in H$ The size of S is a measure of the expressive power of HVC DimensionVC(H)<br>Size of the largest subset<br>of X which can be scattered by H  | Example:<br>H Intervals on the real axis<br>X Real numbers<br>• Can 2 points be scattered?<br>• Can 3 points be scattered?<br>Conclusion: VC( $H$ ) = 2   |

#### Example:

- H Separating hyperplane
- X Points in  $\Re^r$
- When r = 1

$$VC(H) = 2$$

VC-Dimension

Errors While Training

• When r = 2

$$VC(H) = 3$$

Example Complexity Measure

• Generally

$$\operatorname{VC}(H) = r + 1$$

#### Number of Training Examples

Previous estimate

$$m \geq rac{1}{\epsilon} \left[ \ln |H| + \ln rac{1}{\delta} 
ight]$$

New estimate

$$m \geq rac{1}{\epsilon} \left[ 4 \log_2 rac{2}{\delta} + 8 \mathrm{VC}(\mathcal{H}) \cdot \log_2 rac{13}{\epsilon} 
ight]$$

Machine Learning

Örjan Ekeberg

PAC-Learning VC-Dimension Errors While Training

Much better (smaller)

Machine Learning

Theoretical Considerations PAC-Learning VC-Dimension Errors While Training

Örjan Ekeberg

### Theoretical Considerations

• What might Fail?

## 2 PAC-Learning

- Consistent Learners
- Number of Training Examples
- Learning Conjunctions
- Unbiased Learning

### **3** VC-Dimension

- Example
- Complexity Measure

## 4 Errors While Training

- Find-S
- List-then-Eliminate

Alternative Performance Measure for Learning Algorithms:

How many errors does the algorithm make during learning

#### Find-S

- Only learns when making errors
- Worst case: generalises only one attribute each time

Find-S List-then-Eliminate

• First example only chooses one specific hypothesis

PAC-Learning VC-Dimension Errors While Training

• Maximally n + 1 changes

Will maximally make n + 1 errors

#### List-then-Eliminate

- Multiple hypotheses: force the algorithm to guess
- Suppose we use a majority vote among all hypotheses remaining in *Version Space*
- Wrong answer only when at least half of VS give the wrong answer
- For each error made, at least half of VS disappears

Maximally  $\log_2 |H|$  errors

Örjan Ekeberg Machine Learning

Örjan Ekeberg Machine Learning