

Web-Mining Agents

Computational Learning Theory

Prof. Dr. Ralf Möller

Dr. Özgür Özcep

Universität zu Lübeck

Institut für Informationssysteme

Tanya Braun (Exercise Lab)

Computational Learning Theory

(Adapted) Slides by Carla P. Gomes and Nathalie Japkowicz

(Reading: R&N AIMA 3rd ed., Chapter 18.5)

Computational Learning Theory

Inductive learning:

given the **training set**, a **learning algorithm** generates a **hypothesis**.

Run hypothesis on the **test set**. The results say *something* about how **good our hypothesis is**.

But how much do the **results really tell you**? Can we be **certain** about how the learning algorithm **generalizes**?

We would have to see **all the examples**.

Insight: introduce **probabilities** to measure degree of **certainty and correctness** (Valiant 1984).

Computational Learning Theory

Example:

We want to use **height** to **distinguish men** and **women** drawing people from the same distribution for training and testing.

We can never be **absolutely certain** that we **have learned correctly our target (hidden) concept function**. (E.g., there is a non-zero chance that, **so far**, we have only seen a sequence of bad examples)

E.g., relatively tall women and relatively short men...

We'll see that **it's generally highly unlikely to see a long series of bad examples!** (under a stationarity assumption)

Aside: flipping a coin

Experimental data

C program – simulation of flips of a fair coin:

Runs of 100 flips (expect 50 “tails”):

On 1,000 tries reached 66

On 10,000 tries reached 69

On 100,000 tries reached 70

On 1,000,000 tries reached 74 (48% over 50)

Runs of 1000 flips (expect 500 “tails”):

On 1,000 tries reached 564

On 10,000 tries reached 564

On 100,000 tries reached 569

On 1,000,000 tries reached 579 (16% over 500)

Experimental Data Contd.

Runs of 10,000 flips (expect 5000 “tails”):

On 1,000 tries reached 5150

On 10,000 tries reached 5183

On 100,000 tries reached 5231

On 1,000,000 tries reached 5239 (5% over 5000)

With a sufficient number of flips
(set of flips=example of coin bias),
large outliers become quite rare.

Coin example is the key to computational learning theory!

Computational Learning Theory

Intersection of AI, statistics, and theory of computation.

Introduce **Probably Approximately Correct Learning** concerning
efficient learning

For our learning procedures we would like to prove that:

With **high probability** an (efficient) **learning algorithm** will find a hypothesis that is **approximately** identical to the hidden target concept.

Note the double “hedging” – probably and approximately.

Why do we need both levels of uncertainty (in general)?

Probably Approximately Correct Learning

Underlying principle:

Seriously wrong hypotheses can be found out almost certainly (with high probability) using a “small” number of examples

- Any hypothesis that is consistent with a significantly large set of training examples is unlikely to be seriously wrong: it must be probably approximately correct.
- Any (efficient) algorithm that returns hypotheses that are PAC is called a PAC-learning algorithm

Probably Approximately Correct Learning

How many examples are needed to guarantee correctness?

- **Sample complexity** (# of examples to “guarantee” correctness) grows with the size of the Hypothesis space
- **Stationarity assumption**: Training set and test sets are drawn from the same distribution

Notations

ÖÖ:

- Error is ``true error`` . Different from training error
- True error depends on distribution D

Notations:

- X : set of all possible examples
- D : distribution from which examples are drawn
- H : set of all possible hypotheses
- N : the number of examples in the training set
- f : the true function to be learned

Assume: the true function f is in H .

Error of a hypothesis h wrt f :

Probability that h differs from f on a randomly picked example:

$$\text{error}(h) = P(h(x) \neq f(x) \mid x \text{ drawn from } D)$$

Exactly what we are **trying to measure** with our **test set**.

Approximately Correct

A hypothesis h is approximately correct if:

$$\text{error}(h) \leq \varepsilon,$$

where ε is a given threshold, a small constant

Goal:

Show that after seeing a **small (poly) number of examples N** , with **high probability**, all **consistent hypotheses** will be **approximately correct**.

I.e., chance of “bad” hypothesis, (high error but consistent with examples)
is small **(i.e, less than δ)**

ÖÖ: Note that error depends on distribution D

Approximately Correct

Approximately correct hypotheses lie inside

the ϵ -ball around f ;

Those hypotheses that are seriously wrong

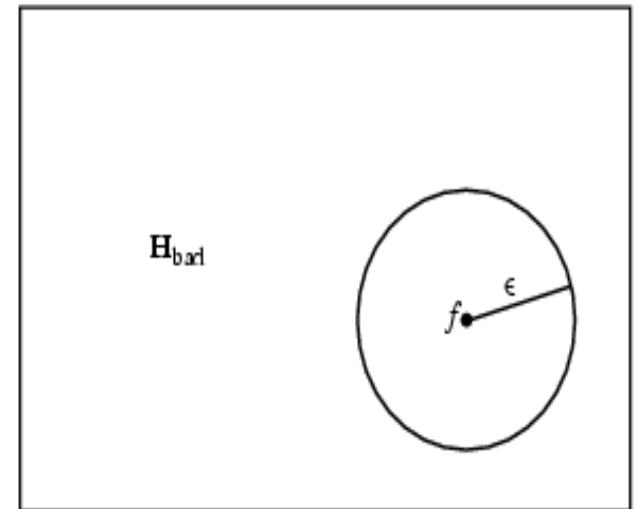
($h_b \in H_{\text{bad}}$) are outside the ϵ -ball,

$\text{Error}(h_b) = P(h_b(x) \neq f(x) \mid x \text{ drawn from } D) > \epsilon$,

Thus the probability that the h_b (a seriously wrong hypothesis) *disagrees* with one example is at least ϵ (definition of error).

Thus the probability that the h_b (a seriously wrong hypothesis) *agrees* with one example is no more than $(1 - \epsilon)$.

So for N examples, $P(h_b \text{ agrees with } N \text{ examples}) \leq (1 - \epsilon)^N$.



Approximately Correct Hypothesis

The probability that H_{bad} contains at least one consistent hypothesis is bounded by the sum of the individual probabilities.

$$P(H_{\text{bad}} \text{ contains a consistent hypothesis, agreeing with all the examples}) \leq |H_{\text{bad}}|(1 - \epsilon)^N \leq |H|(1 - \epsilon)^N$$



h_b agrees with one example is no more than $(1 - \epsilon)$.

$$P(H_{\text{bad}} \text{ contains a consistent hypothesis}) \leq |H_{\text{bad}}|(1 - \epsilon)^N \leq |H|(1 - \epsilon)^N$$

Goal –

Bound the probability of **learning a bad hypothesis** below some small number δ .

Note: $(1 - \epsilon) \leq e^{-\epsilon}$

The **more accuracy** (smaller ϵ), and the **more certainty** (with smaller δ) one wants, the **more examples** one needs.

$$P(H_{\text{bad}} \text{ contains a consistent hypothesis}) \leq |H| (e^{-\epsilon})^N \leq \delta$$

Sample Complexity: Number of examples to guarantee a PAC learnable function class

If the learning algorithm returns a hypothesis that is consistent with this many examples, then with probability at **least $(1 - \delta)$** the **learning algorithm has an error of at most ϵ .**

and the hypothesis is

Probably Approximately Correct.

Probably Approximately correct hypothesis h :

- If the probability of a small error ($\text{error}(h) \leq \epsilon$) is greater than or equal to a given threshold $1 - \delta$
- A bound on the number of examples (sample complexity) needed to guarantee PAC:

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

(The more accuracy (with smaller ϵ), and the more certainty desired (with smaller δ), the more examples one needs.)

- One seeks for (computationally) efficient learning algorithms:
sample complexity N depends only polynomially from some parameter characterizing H and one gets ϵ -correct hypothesis after polynomially many steps

Theoretical results apply to fairly simple learning models (e.g., decision list learning)

ÖÖ:

- Here one assumes that H contains target function f
- In case f is not contained in H one talks about **agnostic PAC learning**.
- Similar bound derivable:

$$N \geq 1/(2\varepsilon^2)(\ln(1/\delta) + \ln(|H|))$$

PAC Learning

Sample complexity – a **polynomial number of examples** suffices to specify a good consistent hypothesis ($\text{error}(h) \leq \varepsilon$) with high probability ($\geq 1 - \delta$).

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

Computational complexity – there is an **efficient algorithm** for learning a **consistent hypothesis** from the small sample.

Let's be more specific with examples.

Example: Boolean Functions

Consider H the set of all Boolean function on n attributes $\rightarrow |H| = 2^{2^n}$

$$N \geq \frac{1}{\epsilon} \left(\ln \frac{1}{\delta} + \ln |H| \right) = O(2^n)$$

So the **sample complexity** grows as 2^n ☹!
(same as the number of all possible examples)

Not PAC-Learnable!

So, any **learning algorithm** will do not better than a lookup table
if it **merely** returns a hypothesis that is **consistent** with all known
examples!

Intuitively what does it say about H ?

Finite H required!

(ÖÖ: or H with finite characteristics)

Coping With Learning Complexity

1. Force learning algorithm to look for **small/simple** consistent hypothesis.

We considered that for Decision Tree Learning, often **worst case intractable** though.

2. **Restrict size of hypothesis space.**

e.g., Decision Lists (DL) → restricted form of Boolean Functions:

Hypotheses correspond to a series of tests, each of which is a conjunction of literals

Good news: only a poly size number of examples

is required for guaranteeing PAC learning K-DL functions (maximal k conjuncts) and there are **efficient algorithms** for learning K-DL

Decision Lists

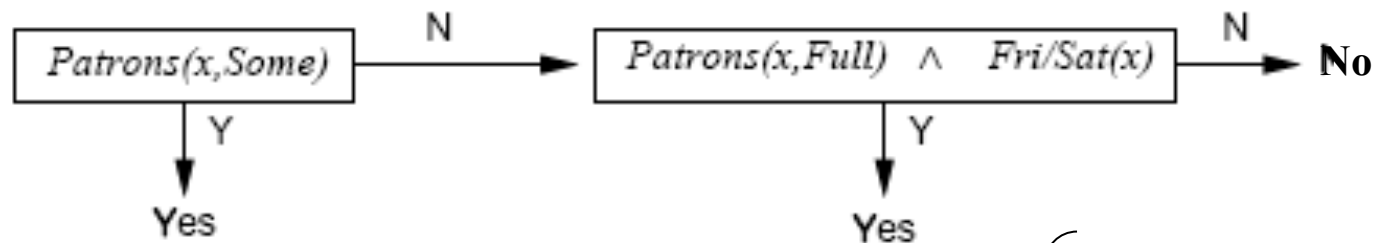
DLs resemble Decision Trees, but with simpler structure:

Series of tests, each test a conjunction of literals;

If a test succeeds, decision list specifies value to return;

If test fails, processing continues with the next test in the list.

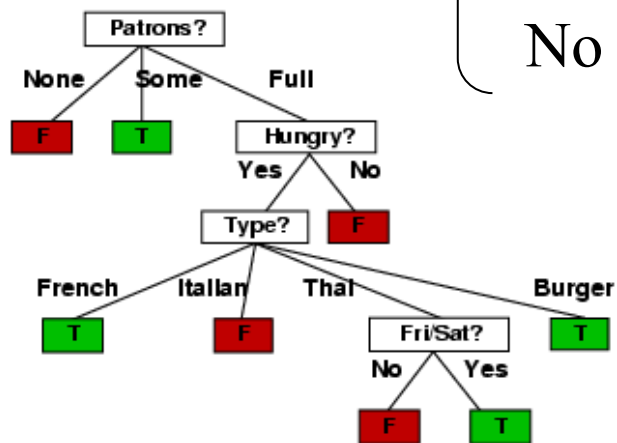
Forall x: Willwait(x) \leftrightarrow Patrons(x,some) or (Patrons(full) & Fri/Sat(x))



a=Patrons(x,Some) b=patrons(x,Full) c=Fri/Sat(x)

$$\begin{pmatrix} (a) & (b \wedge c) \\ Y & Y & N \end{pmatrix}$$

Note: if we allow arbitrarily many literals per test, decision list can express all Boolean functions.



(a)	(b)	(¬d)	(e)	(f)	(h)	(i)	
No	Yes	No	Yes	No	Yes	Yes	No

a=Patrons(x,None) b=Patrons(x,Some)

d=Hungry(x)

e=Type(x,French) f=Type(x,Italian) g=Type(x,Thai) h=Type(x,Burger)

i=Fri/Sat(x)

K Decision Lists

Decision Lists with limited expressiveness (K-DL) – at most k literals per test

$$\text{2-DL} \quad \left(\begin{array}{cc} (a) & (b \wedge c) \\ Y & Y \quad N \end{array} \right)$$

K-DL is PAC learnable!!!

For fixed k literals, the **number of examples** needed for **PAC learning a K-DL function** is **polynomial** in the number of attributes n.

There are **efficient algorithms** for learning K-DL functions.

So how do we show K-DL is PAC-learnable?

2-DL

(x)	(y)	(w ∧ ¬v)	(u ∧ ¬b)	
No	Yes	No	Yes	No

K Decision Lists: Sample Complexity

$$N \geq \frac{1}{\varepsilon} \left(\ln \frac{1}{\delta} + \ln |H| \right) \quad \text{What's the size of the hypothesis space } H, \text{ i.e., } |K\text{-DL}(n)|?$$

K-Decision Lists \rightarrow set of tests: each test is a conjunct of at most k literals

How many possible tests (conjuncts) of length at most k, given n attributes, $conj(n,k)$?

$$|Conj(n,k)| \leq 2n + \binom{2n}{2} + \binom{2n}{3} + \dots + \binom{2n}{k} = O(n^k)$$

A conjunct (or test) can appear in the list as: Yes, No, absent from list

So we have at most $3^{|Conj(n,k)|}$ different K-DL lists (ignoring order)

But the order of the tests (or conjuncts) in a list matters.

$$|k\text{-DL}(n)| \leq 3^{|Conj(n,k)|} |Conj(n,k)|!$$

After some work (ÖÖ: using Stirling formula say), we get (exercise)

$$|K - DL(n)| = 2^{O(n^k \log_2(n^k))}$$

1 - Sample Complexity of K-DL is:

Recall sample complexity formula

$$N \geq \frac{1}{\varepsilon} \left(\ln \frac{1}{\delta} + O(n^k \log_2(n^k)) \right)$$

$$N \geq \frac{1}{\varepsilon} \left(\ln \frac{1}{\delta} + \ln |H| \right)$$

For **fixed k literals**, the number of **examples needed for PAC learning a K-DL function** is **polynomial in the number of attributes n**, ☺!

2 – Efficient learning algorithm – a decision list of length k can be learned in polynomial time.

So K-DL is PAC learnable!!!

Decision-List-Learning Algorithm

Greedy algorithm for learning decisions lists:

- repeatedly finds a test that agrees with some subset of the training set;
- adds test to the decision list under construction and removes the corresponding examples.
- uses the remaining examples, until there are no examples left, for constructing the rest of the decision list.
(Selection strategy not specified)

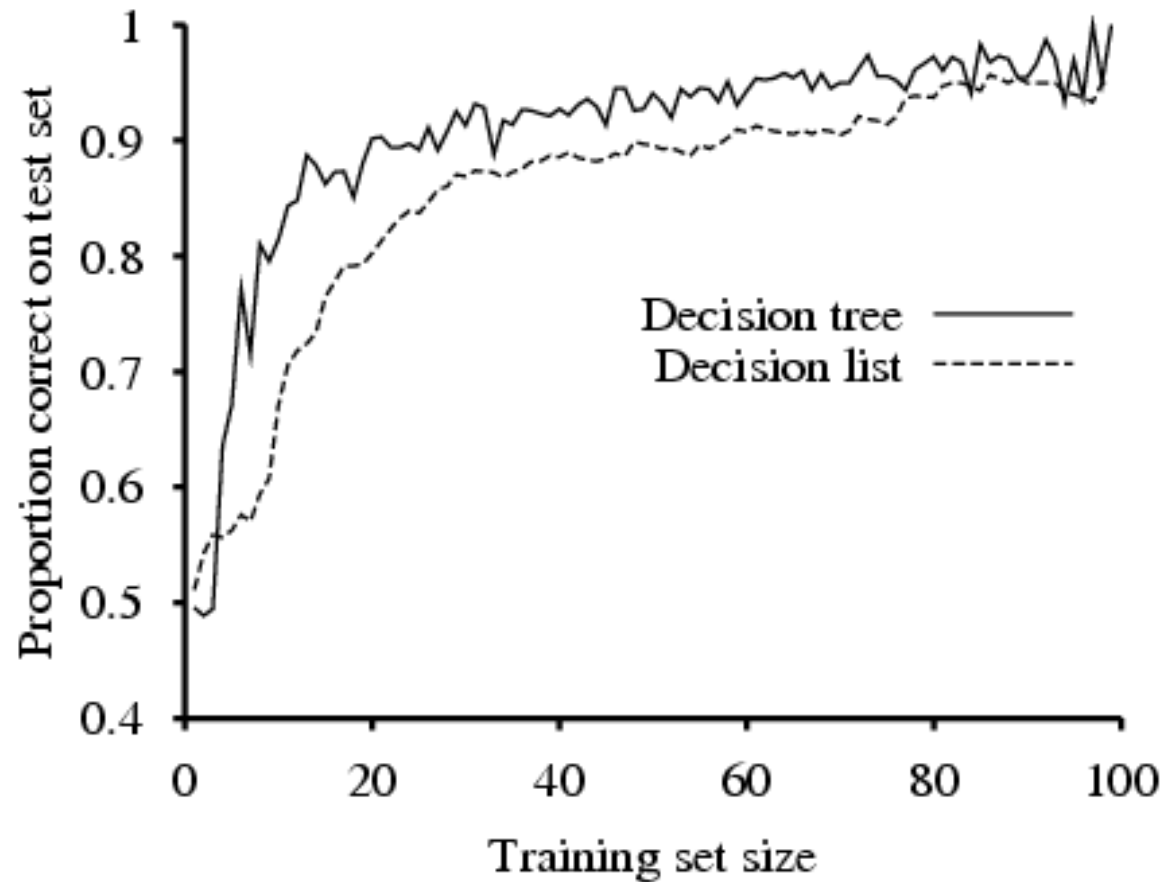
Decision-List-Learning Algorithm

Greedy algorithm for learning decisions lists:

```
function DECISION-LIST-LEARNING(examples) returns a decision list, No or failure  
  
  if examples is empty then return the value No  
   $t \leftarrow$  a test that matches a nonempty subset examplest of examples  
    such that the members of examplest are all positive or all negative  
  if there is no such t then return failure  
  if the examples in examplest are positive then  $o \leftarrow$  Yes  
  else  $o \leftarrow$  No  
  return a decision list with initial test t and outcome o  
    and remaining elements given by DECISION-LIST-LEARNING(examples − examplest)
```

Decision-List-Learning Algorithm

ÖÖ: Here algorithm with selection strategy
Find smallest test set for uniformly classified subset



Restaurant data.

Examples

1. H space of Boolean functions: Not PAC Learnable, hypothesis space too big: need too many examples (sample complexity not polynomial)!
2. K-DL: PAC learnable
3. Conjunction of literals: PAC learnable

ÖÖ:

- PAC-Learnability depends on the hypothesis space
 - Sometimes using a hypothesis space different from space of target functions helps!
- E.g. k-term DNF (k disjuncts of conjuncts with n attributes) learnable with hypothesis space consisting of k-cnfs (conjunctions of arbitrary length with disjunctions up to length k)

Probably Approximately Correct Learning (PAC) Learning (summary)

A class of functions is said to be PAC-learnable if there exists an efficient (i.e., polynomial in size of target function, size of example instances (n), $1/\epsilon$, and $1/\delta$)

Learning algorithm such that for all functions in the class, and for all probability distributions on the function's domain, and for any values of epsilon and delta ($0 < \epsilon, \delta < 1$), using a polynomial number of examples, the algorithm will produce a hypothesis whose error is smaller than ϵ with probability at least δ .

The error of a hypothesis is the probability that it will differ from the target function on a random element from its domain, drawn according to the given probability distribution.

Basically, this means that:

- there is some way to learn efficiently a "pretty good" approximation of the target function.
- the probability is as big as you like that the error is as small as you like.

(Of course, the tighter you make the bounds, the harder the learning algorithm is likely to have to work)

Discussion

Computational Learning Theory studies the tradeoffs between the expressiveness of the hypothesis language and the complexity of learning

Probably Approximately Correct learning concerns efficient learning

Sample complexity --- polynomial number of examples
Efficient Learning Algorithm

Word of caution:

PAC learning results \rightarrow worst case complexity results.

Sample Complexity for Infinite Hypothesis Spaces I: VC-Dimension

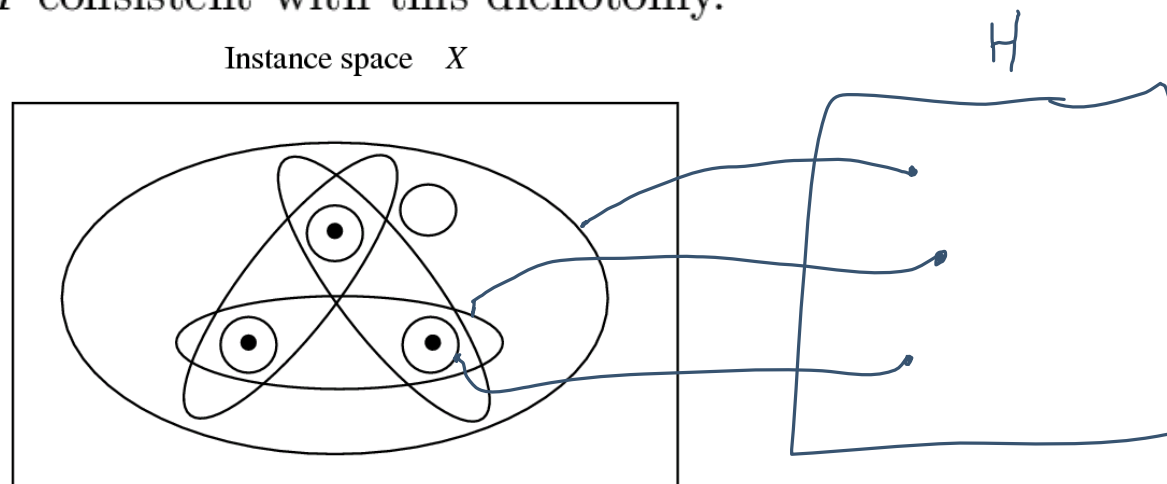
- The PAC Learning framework has 2 disadvantages:
 - It can lead to weak bounds
 - Sample Complexity bound cannot be established for infinite hypothesis spaces (with functions having continuous domain/range, say)
- We introduce new ideas for dealing with these problems:
 - A set of instances S is **shattered** by hypothesis space H iff for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

Shattering a Set of Instances

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

a labeling of each member of S as positive or negative

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.





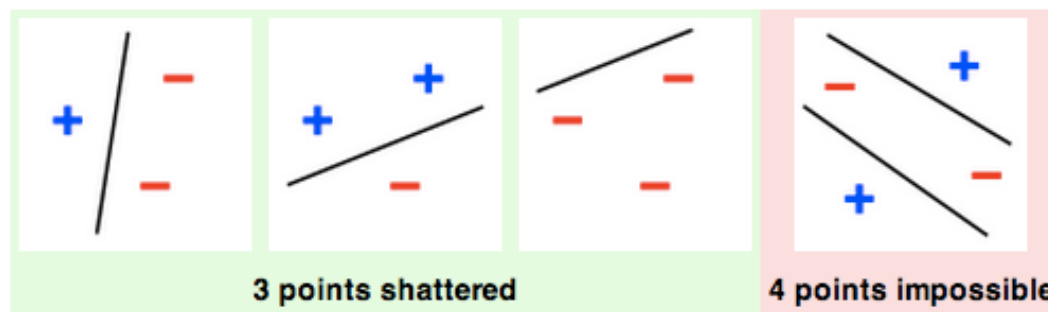
VC Dimension: Example

VC dimension of a classification model [\[edit \]](#)

A classification model f with some parameter vector θ is said to *shatter* a set of data points (x_1, x_2, \dots, x_n) if, for all assignments of labels to those points, there exists a θ such that the model f makes no errors when evaluating that set of data points.

The **VC dimension** of a model f is the maximum number of points that can be arranged so that f shatters them. More formally, it is the maximum integer D such that some data point set of *cardinality* D can be shattered by f .

3. Consider a *straight line* as a classification model on points in a two-dimensional plane (this is the model used by a *perceptron*). The line should separate positive data points from negative data points. There exist sets of 3 points that can indeed be shattered using this model (any 3 points that are not collinear can be shattered). However, no set of 4 points can be shattered: by *Radon's theorem*, any four points can be partitioned into two subsets with intersecting *convex hulls*, so it is not possible to separate one of these two subsets from the other. Thus, the VC dimension of this particular classifier is 3. It is important to remember that while one can choose any arrangement of points, the arrangement of those points cannot change when attempting to shatter for some label assignment. Note, only 3 of the $2^3 = 8$ possible label assignments are shown for the three points.



Sample Complexity for Infinite Hypothesis Spaces I: VC-Dimension

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 $VC(H)=\infty$

Aside: Intuitive derivation of VC dimension

- How to define a natural notion of **dimension** on Hypothesis space $H(X) = \{h \mid h: X \rightarrow \{0,1\}\}$
- Dimension should be monotone
- So define Dimension on simple/small spaces first
- Aim: Define dimension for all subsets H of $H(X)$
 - Simple Case ($H = H(X)$): $\dim(H) = |X|$
(If X infinite, then $\dim(H) = \infty$)
 - Complex case (H proper subset of $H(X)$):
 $\dim(H) = \text{size of biggest simple subset of } H$
 $= \text{size of biggest subset } Y \text{ of } X \text{ s.t. } H(Y) \text{ subset of } H$

Observation: $\dim(H) = VC(H)$

VC dimension and PAC Learning

- PAC Learning possible by restricting hypothesis space H
- Leads to bias
- Remember: Sample complexity for PAC learning

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

- With VC dimension (also applicable for infinite spaces; H_n subclass of hypothesis parameterized by n):

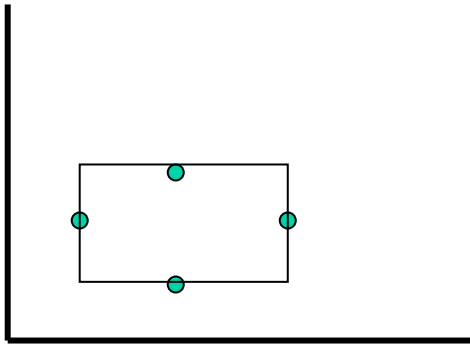
$$N \geq \frac{1}{\epsilon(1 - \sqrt{\epsilon})} \left(2VCdim(H_n) \ln \frac{6}{\epsilon} + \ln \frac{2}{\delta} \right)$$

VC Dimension: Example 2

- $H =$ Axis parallel rectangles in \mathbb{R}^2
- What is the VC dimension of H
- Can we PAC learn?

Learning Rectangles

- Consider axis parallel rectangles in the real plane
- Can we PAC learn it ?
 - (1) What is the VC dimension ?
- Some four instances (points on the rectangle) can be shattered



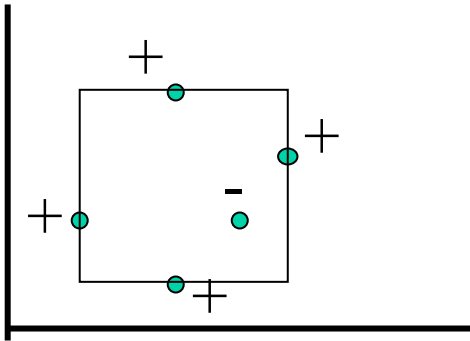
Shows that $VC(H) \geq 4$

Learning Rectangles

- Consider axis parallel rectangles in the real plane
- Can we PAC learn it ?

(1) What is the VC dimension ?

- But, no five instances can be shattered



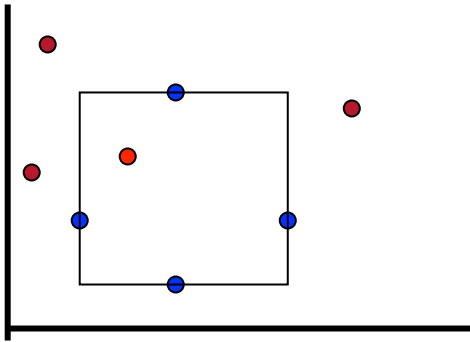
Pick the topmost, bottommost, leftmost and rightmost points and give them the label “+”. The fifth one gets -.

Cannot be shattered.

Therefore $VC(H) = 4$

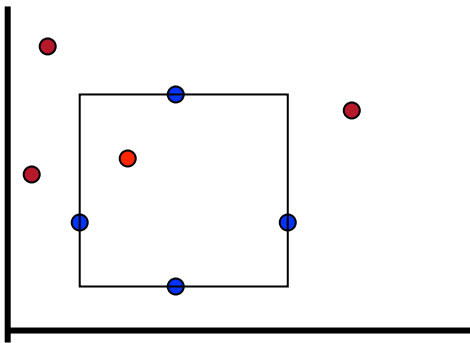
Learning Rectangles

- Consider axis parallel rectangles in the real plane
- Can we PAC learn it ?
 - (1) What is the VC dimension ?
 - (2) Can we give an efficient algorithm ?



Learning Rectangles

- Consider axis parallel rectangles in the real plane
- Can we PAC learn it ?
 - (1) What is the VC dimension ?
 - (2) Can we give an efficient algorithm ?



Find the smallest rectangle that contains the positive examples (necessarily, it will not contain any negative example, and the hypothesis is consistent).

Axis parallel rectangles are efficiently PAC learnable.

Exercise: What is the VC dimension of intervals on \mathbb{R} ?

The *Mistake Bound* Model of Learning

- The *Mistake Bound* framework is different from the PAC framework as it considers learners that receive a sequence of training examples and that predict, upon receiving each example, what its target value is.
(So, it has an incremental, online-flavor)
- The question asked in this setting is: “*How many mistakes M_A will the learner A make in its predictions before it learns the target concept?*”
- This question is significant in practical settings where learning must be done while the system is in actual use.

Optimal Mistake Bounds

- **Definition:** Let C be an arbitrary nonempty concept class. The optimal mistake bound for C , denoted $Opt(C)$, is the minimum over all possible learning algorithms A of $M_A(C)$. $Opt(C) = \min_{A \in \text{Learning_Algorithms}} M_A(C)$
- **Proposition:** For any concept class C , the optimal mistake bound is bound as follows:

$$VC(C) \leq Opt(C) \leq \log_2(|C|)$$