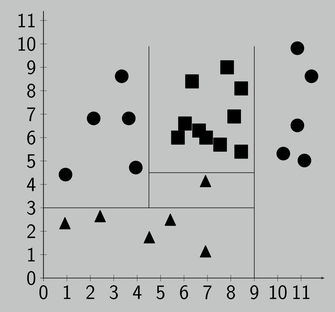
# VC-Dimension of Rule Sets

Olcay Taner Yıldız
Department of Computer Engineering, Işık University, TR-34980, Istanbul, Turkey

## Rule Induction

- A rule set is typically an ordered list of rules.
- A rule contains a conjunction of terms and a class code.
- The terms are of the form $x_i = v$, $x_i < \theta$ or $x_i \geq \theta$.

## Separation of Data and Learned RuleSet




**If** $(x_1 < 9)$ **and** $(x_2 < 3)$
  **Then** class = ▲
**Else**
  **If** $(x_1 > 4.5)$ **and** $(x_2 < 4.5)$
    **Then** class = ▲
  **Else**
    **If** $(x_1 < 4.5)$
      **Then** class = ●
    **Else**
      **If** $(x_1 > 9)$
        **Then** class = ●
      **Else** class = ■

## VC-Dimension

- VC dimension for a class of functions $f(\mathbf{x}, \alpha)$, where $\alpha$ denotes the parameter vector, is defined to be the largest number of points that can be shattered by members of $f(\mathbf{x}, \alpha)$.
- A set of data points is *shattered* by a class of functions $f(\mathbf{x}, \alpha)$ if for all assignments of class labels to those points, one can find a member of $f(\mathbf{x}, \alpha)$ which makes no errors when evaluating that set of data points.
- For example, in two dimensions, we can separate three points with a line, but we can not separate four points. Therefore, the VC dimension of the linear estimator class in two dimensions is 3.

## Definitions

- The rule set construction algorithm uses a sample of $m$ labeled points $S = (\mathbf{X}, \mathbf{Y}) = ((\mathbf{x^{(1)}}, y^{(1)}), \ldots, (\mathbf{x^{(m)}}, y^{(m)}) \in (\mathcal{X} \times \mathcal{Y})^m$.
- $\mathcal{X}$ is the input space and $\mathcal{Y}$ the label set, which is $\{0, 1\}$.
- The input space $\mathcal{X}$ is a vectorial space of dimension $d$, the number of features, where each feature can take values from $\{0, 1\}$.
- Each rule set $\mathcal{R}$ is represented with a vector $\mathcal{R} = [r_1, r_2, \ldots, r_k]_k$, where there are $r_1$, $r_2$, ..., $r_k$ conditions in the first, second, ..., $k^{th}$ rule respectively.
- For example, the rule set given above is an element of the hypothesis class $f(\mathbf{x}, [2, 2, 1, 1]_4)$, where the first, second, third, and fourth rules have 2, 2, 1, and 1 conditions respectively.

## Theorem 1

*The VC-dimension of rule set $\mathcal{R}_1 = [1]_1$ (a single rule composed of a single decision condition) that classifies $d$ dimensional binary data is $\lfloor \log_2(d+1) \rfloor + 1$.*

## Example

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|---|---|---|---|---|---|---|---|
| $\mathbf{x}^{(1)}$ | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| $\mathbf{x}^{(2)}$ | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| $\mathbf{x}^{(3)}$ | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| $\mathbf{x}^{(4)}$ | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

**If** $x_5 = 1$ **Then** $C_1$ **Else** $C_0$ | **If** $x_3 = 1$ **Then** $C_1$ **Else** $C_0$

Example for Theorem 1 with $d = 7$ and $m = 4$. If the class labeling of $S$ is {1, 1, 0, 0} we select feature 5 (left rule set). If the class labeling of $S$ is {0, 0, 1, 0} we select feature 3 (right rule set).

## Theorem 2

*The VC-dimension of of rule set $\mathcal{R}_2 = [h]_1$ (a single rule composed of a $h$ decision conditions) that classifies $d$ dimensional binary data is $\left\lfloor \log_2 \left( \binom{d}{h} + 1 \right) \right\rfloor + 1$.*

## Theorem 3

*The VC-dimension of rule set $\mathcal{R}_3 = [1, 1, \ldots, 1]_k$ that classifies $d$ dimensional binary data is at least $\lfloor \log_2(d - k + 2) \rfloor + k$.*

## Example

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|---|---|---|---|---|---|---|---|
| $\mathbf{x}^{(1)}$ | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| $\mathbf{x}^{(2)}$ | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| $\mathbf{x}^{(3)}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $\mathbf{x}^{(4)}$ | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| $\mathbf{x}^{(5)}$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $\mathbf{x}^{(6)}$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $\mathbf{x}^{(7)}$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**If** $x_7 = 1$ **Then** $C_x$
**Else**
  **If** $x_6 = 1$ **Then** $C_x$
  **Else**
    **If** $x_5 = 1$ **Then** $C_x$
    **Else**
      **If** $x_4 = 1$ **Then** $C_x$
      **Else**
        **If** $x_3 = 0$ **Then** $C_0$
        **Else** $C_1$

Example for Theorem 3 with $d = 7$ and $m = 7$. If the class labeling of $S$ is {1, 1, 0, $x$, $x$, $x$, $x$} we select feature 3 in the bottom rule. The labelings of the last four examples do not matter since they are alone in the rules they reside.

## Theorem 4

*The VC-dimension of rule set $\mathcal{R}_4 = [h, h, \ldots, h]_{2^{h-1}}$ that classifies $d$ dimensional binary data is at least $2^{h-1}(\lfloor \log_2(d - h + 2) \rfloor + 1)$.*

## Example

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| $\mathbf{x}^{(1)}$ | 1 | 0 | 1 | 0 | 0 |
| $\mathbf{x}^{(2)}$ | 0 | 1 | 1 | 0 | 0 |
| $\mathbf{x}^{(3)}$ | 0 | 0 | 0 | 0 | 0 |
| $\mathbf{x}^{(4)}$ | 1 | 0 | 1 | 0 | 1 |
| $\mathbf{x}^{(5)}$ | 0 | 1 | 1 | 0 | 1 |
| $\mathbf{x}^{(6)}$ | 0 | 0 | 0 | 0 | 1 |
| $\mathbf{x}^{(7)}$ | 1 | 0 | 1 | 1 | 0 |
| $\mathbf{x}^{(8)}$ | 0 | 1 | 1 | 1 | 0 |
| $\mathbf{x}^{(9)}$ | 0 | 0 | 0 | 1 | 0 |
| $\mathbf{x}^{(10)}$ | 1 | 0 | 1 | 1 | 1 |
| $\mathbf{x}^{(11)}$ | 0 | 1 | 1 | 1 | 1 |
| $\mathbf{x}^{(12)}$ | 0 | 0 | 0 | 1 | 1 |

**If** $x_4 = 0$ **and** $x_5 = 0$ **and** $x_3 = 0$ **Then** $C_0$
**Else**
  **If** $x_4 = 0$ **and** $x_5 = 1$ **and** $x_1 = 0$ **Then** $C_0$
  **Else**
    **If** $x_4 = 1$ **and** $x_5 = 0$ **and** $x_2 = 0$ **Then** $C_0$
    **Else**
      **If** $x_4 = 1$ **and** $x_5 = 1$ **and** $x_1 = 0$ **Then** $C_0$
      **Else** $C_1$

Example for Theorem 4 with $d = 5$ and $m = 12$. Using features 4 and 5 as the first two features in all rules, one divides the class labelings into 4 subproblems of $m = 3$. Each subproblem can then be shattered with a single condition. For the example rule, the class labeling of $S$ is {1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0}.

## Theorem 5

*The VC-dimension of a rule set with binary features that classifies $d$ dimensional binary data is at least the maximum of the sum of the VC-dimensions of its sub rule sets those classifying $d - 1$ dimensional binary data.*

## Lower bound of the VC-dimension of a rule set for binary data

```
int VC-Dimension1(R = [r_1, r_2, ..., r_k]_k, d)
1  if k = 1 and r_1 = 1
2      return ⌊log_2(d + 1)⌋ + 1
3  if k = 1 and r_1 != 1
4      return ⌊log_2 (C(d, r_1) + 1)⌋ + 1
5  max = 0
6  for i = 1 to k - 1
7      s = VC-Dimension1([r_1 - 1, ..., r_i - 1], d - 1) +
       VC-Dimension1([r_{i+1} - 1, ..., r_k - 1], d - 1)
8      if s > max
9          max = s
10 return max
```

## Corollary 1

*The VC-dimension of rule set $\mathcal{R}_1 = [1]_1$ that classifies $d$ dimensional continuous data is at least $\lfloor \log_2(d+1) \rfloor + 1$.*

## Example

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|---|---|---|---|---|---|---|---|
| $\mathbf{x}^{(1)}$ | 1.3 | 0.2 | 0.1 | 0.4 | 1.7 | 1.4 | 1.7 |
| $\mathbf{x}^{(2)}$ | 0.9 | 1.3 | 0.7 | 0.1 | 1.1 | 0.1 | 0.1 |
| $\mathbf{x}^{(3)}$ | 0.5 | 0.8 | 1.4 | 0.9 | 0.6 | 1.8 | 0.6 |
| $\mathbf{x}^{(4)}$ | 0.6 | 0.6 | 0.3 | 1.8 | 0.3 | 0.3 | 1.8 |

**If** $x_5 \leq 1$ **Then** $C_0$ **Else** $C_1$ | **If** $x_3 \leq 1$ **Then** $C_0$ **Else** $C_1$

Example for the Corollary 1 with $d = 7$ and $m = 4$. If the class labeling of $S$ is {1, 1, 0, 0} we select feature 5 and the split $x_5 \leq 1$ (left rule set). If the class labeling of $S$ is {0, 0, 1, 0} we select feature 3 and the split $x_3 \leq 1$ (right rule set).

## Corollary 2

*The VC-dimension of rule set $\mathcal{R}_2 = [h]_1$ that classifies $d$ dimensional continuous data is at least $\left\lfloor \log_2 \left( \binom{d}{h} + 1 \right) \right\rfloor + 1$.*

## Lower bound of the VC-dimension of a rule set for continuous data

```
int VC-Dimension2(R = [r_1, r_2, ..., r_k]_k)
1  sum = 0
2  for i = 1 to k do
3      sum += ⌊log_2 (C(d - 1, r_i - 2) + 1)⌋ + 1
4  return sum
```

## Example

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ |
|---|---|---|---|---|
| $\mathbf{x}^{(1)}$ | 1.3 | 0.5 | 1.3 | 0.2 |
| $\mathbf{x}^{(2)}$ | 0.8 | 1.5 | 1.6 | 0.3 |
| $\mathbf{x}^{(3)}$ | 0.7 | 0.8 | 0.4 | 0.2 |
| $\mathbf{x}^{(4)}$ | 1.4 | 0.3 | 1.2 | 0.7 |
| $\mathbf{x}^{(5)}$ | 0.6 | 1.3 | 1.5 | 0.9 |
| $\mathbf{x}^{(6)}$ | 0.5 | 0.5 | 0.6 | 0.7 |
| $\mathbf{x}^{(7)}$ | 1.3 | 0.6 | 1.5 | 1.2 |
| $\mathbf{x}^{(8)}$ | 0.7 | 1.6 | 1.3 | 1.4 |
| $\mathbf{x}^{(9)}$ | 0.9 | 0.2 | 0.7 | 1.3 |
| $\mathbf{x}^{(10)}$ | 1.4 | 0.7 | 1.3 | 1.7 |
| $\mathbf{x}^{(11)}$ | 0.7 | 1.6 | 1.6 | 1.8 |
| $\mathbf{x}^{(12)}$ | 0.6 | 0.8 | 0.3 | 1.6 |

**If** $x_4 > 0$ **and** $x_4 \leq 0.5$ **and** $x_3 \leq 1.0$ **Then** $C_0$
**Else**
  **If** $x_4 > 0.5$ **and** $x_4 \leq 1.0$ **and** $x_1 \leq 1$ **Then** $C_0$
  **Else**
    **If** $x_4 > 1.0$ **and** $x_4 \leq 1.5$ **and** $x_2 \leq 1$ **Then** $C_0$
    **Else**
      **If** $x_4 > 1.5$ **and** $x_4 \leq 2.0$ **and** $x_1 \leq 1$ **Then** $C_0$
      **Else** $C_1$

Example for algorithm VC-Dimension2 for continuous data with $d = 4$ and $m = 12$. Using the spared feature 4 in all rules, one divides the class labelings into 4 subproblems of $m = 3$. Each subproblem can then be shattered with the remaining features. For the example rule, the class labeling of $S$ is {1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0}.


olcaytaner,sezin.ata@isikun.edu.tr