On the VC-dimension of univariate decision trees

Olcay Taner Yıldız

Dept of Computer Engineering, Işık University, TR-34980, Istanbul, Turkey

ICPRAM 2012

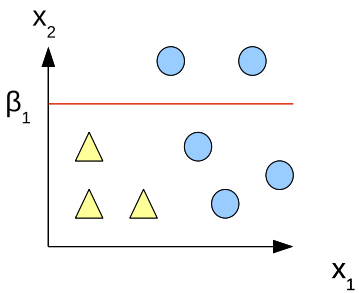
Outline

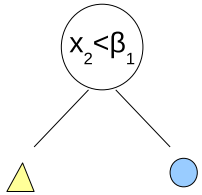
Outline

Introduction

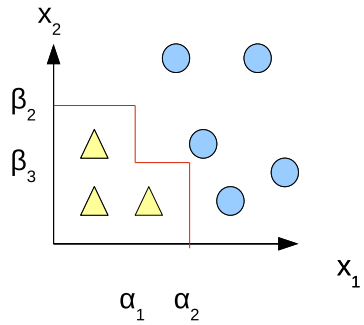
- In pattern recognition, the main goal of the learner is to extract the optimal model from a training set.
- Optimal model $\rightarrow$ the model with least generalization error
- Penalization approaches: $E_g = E_t + \lambda C_{model}$

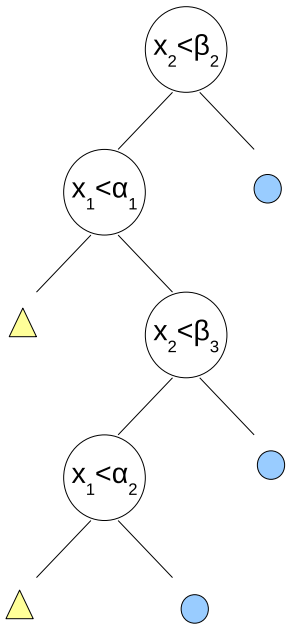
Model Selection



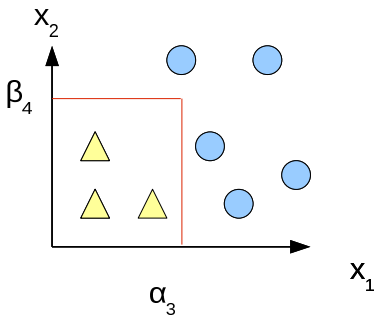


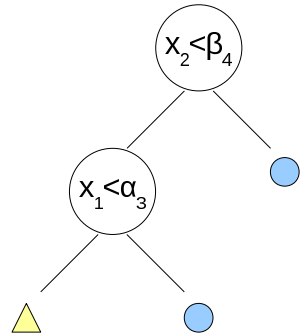
Underfit





Overfit





Best Fit

Structural Risk Minimization

- Possible models are ordered according to their complexity

$$M_0 \subset M_1 \subset M_2 \subset \ldots \tag{1}$$

- For binary classification, generalization error is defined as

$$E_g = E_t + \frac{\epsilon}{2}\left(1 + \sqrt{1 + \frac{4E_t}{\epsilon}}\right) \tag{2}$$

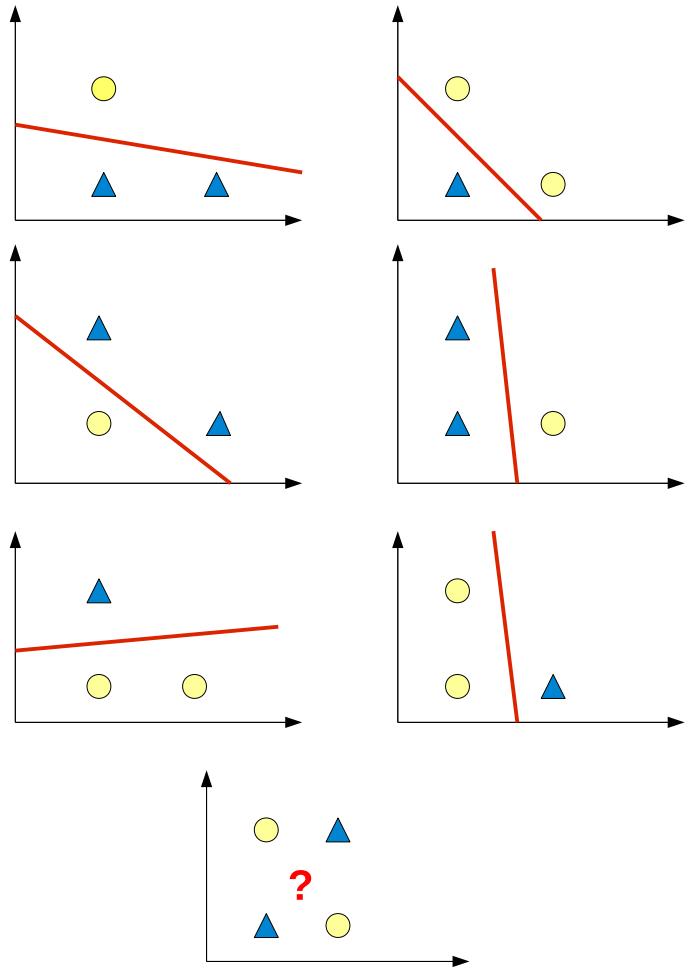
where ϵ is given by the formula

$$\epsilon = a_1 \frac{V[\log(a_2 N/V) + 1] - \log(\nu)}{N} \tag{3}$$

where V represents the VC dimension of the model.

Vapnik-Chervonenkis (VC) dimension

- VC dimension for a class of functions $f(x, \alpha)$, where α denotes the parameter vector, is defined to be the largest number of points that can be shattered by members of $f(x, \alpha)$.
- A set of data points is *shattered* by a class of functions $f(x, \alpha)$ if for each possible class labeling of the points, one can find a member of $f(x, \alpha)$ which perfectly separates them.



Outline

Setup

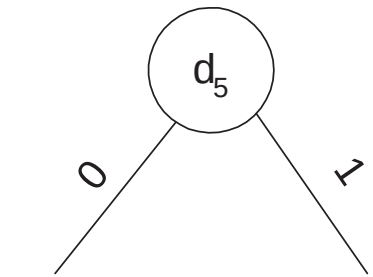
- A sample of m labeled points
  $S = ((\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)}) \in (X \times Y)^m$
- Input space: $X \in \{0,1\}^d$
- Label set: $Y \in \{0,1\}$

Single Node Tree

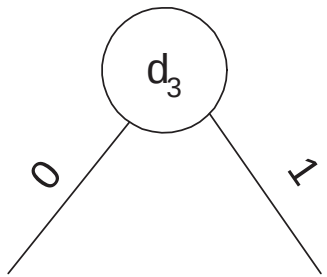
Theorem

The VC-dimension of a univariate decision tree with a single node that classifies d dimensional data is $\lfloor \log_2(d+1) \rfloor + 1$.

	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



$\mathbf{x}^{(3)}$ (0) $\mathbf{x}^{(4)}$ (0) $\mathbf{x}^{(1)}$ (1) $\mathbf{x}^{(2)}$ (1)



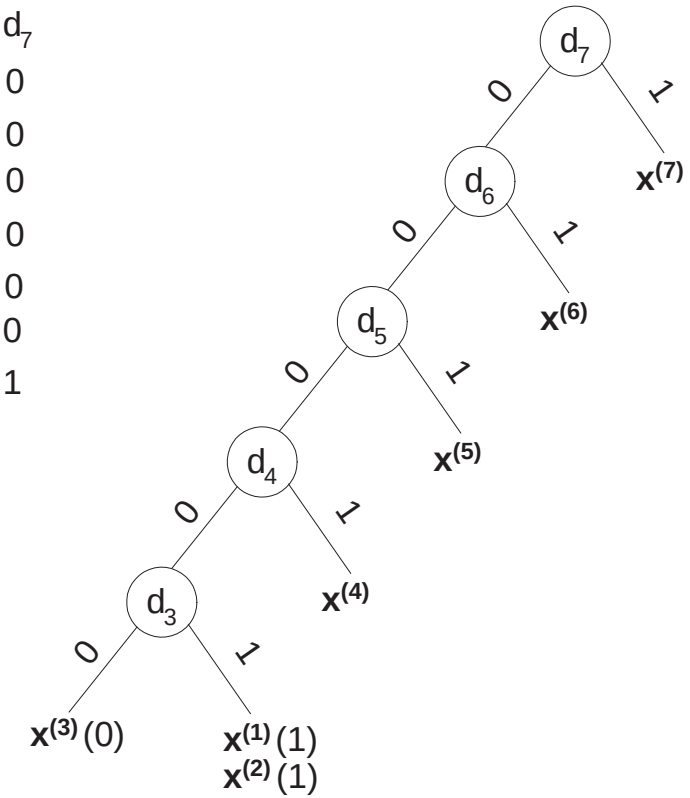
$\mathbf{x}^{(1)}$ (0) $\mathbf{x}^{(2)}$ (0) $\mathbf{x}^{(3)}$ (1)

Degenerate Tree

Theorem

The VC-dimension of a degenerate univariate decision tree with N nodes that classifies d dimensional data is at least $\lfloor \log_2(d - N + 2) \rfloor + N$.

	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

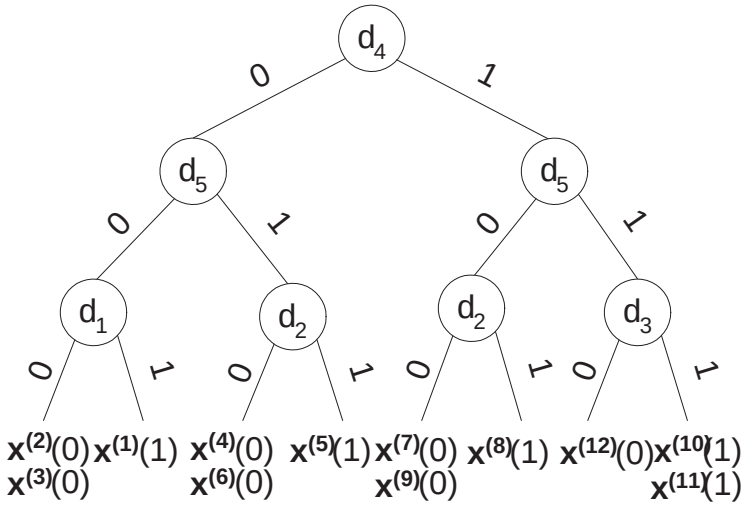


Full Tree

Theorem

The VC-dimension of a full univariate decision tree of height h that classifies d dimensional data is at least $2^{h-1}(\lfloor \log_2(d - h + 2) \rfloor + 1)$.

	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



Lower Bound for VC-Dim of Binary Decision Tree

Theorem

The VC-dimension of a univariate decision tree with binary features that classifies d dimensional data is at least the sum of the VC-dimensions of its left and right subtrees those classifying $d-1$ dimensional data.

```
VCDimension LB(DT, d)
1  if DT is a leaf node
2      return 1
3  if left and right subtrees of DT are leaves
4      return ⌊log2(d + 1)⌋ + 1
5  DT_L = Left subtree of DT
6  DT_R = Right subtree of DT
7  return LB(DT_L, d − 1) + LB(DT_R, d − 1)
```

Outline

Setup

- A sample of m labeled points
  $S = ((\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)}) \in (X \times Y)^m$
- Input space: $X_i \in \{1, 2, \ldots, L_i\}$
- Label set: $Y \in \{0, 1\}$

Single Node Tree

Theorem

The VC-dimension of a single node L-ary decision tree that classifies d dimensional data is $\lfloor \log_2(\sum_{i=1}^{d}(2^{L_i-1}-1)+1) \rfloor + 1$.

Lower Bound for VC-Dim of L-ary Decision Tree

Theorem

The VC-dimension of L-ary decision tree that classifies d dimensional data is at least the sum of the VC-dimensions of its subtrees those classifying $d-1$ dimensional data.

```
VCDimension LB-L-ary(DT, d)
1   if DT is a leaf node
2       return 1
3   if all subtrees of DT are leaves
4       return ⌊log2(∑_{i=1}^{d} (2^{L_i − 1} − 1) + 1)⌋ + 1
5   sum = 0
6   for i = 1 to number of subtrees
7       sum += LB-L-ary(DT_i, d − 1)
8   return sum
```

Outline

Complexity Control Using VC-Dimension Bounds

- SRMPrune
  - Find E_{g1} using Equation 2, where V is the VC-dimension and E_t is the training error of the subtree.
  - Find E_{g2} using Equation 2, where V is the VC-dimension and E_t is the training error of the leaf node.
  - If $E_{g2} < E_{g1}$ prune subtree, otherwise keep it.
- CVPrune
  - Find E_1 the error of the subtree on the validation set.
  - Find E_2 the error of the leaf node on the validation set.
  - If $E_{g2} < E_{g1}$ prune subtree, otherwise keep it.
- NOPrune

Details of Datasets

Set	d	N	n/v
Acceptors	88	3889	88/4
Artificial	10	320	10/2
Donors	13	6246	13/4
Krvskp	36	3196	35/2, 1/3
Monks	6	432	1/2, 3/3, 1/4
Mushroom	22	8124	mixed
Promoters	57	106	57/4
Spect	22	267	22/2
Tictactoe	9	958	9/3
Titanic	3	2201	2/2, 1/4
Vote	16	435	16/2

Complexity Control: Error Rate

Set	NOprune	CVprune	SRMprune
Acceptors	17.1 ± 1.6	$\mathbf{15.5 \pm 2.3}$	15.6 ± 1.6
Artificial	$\mathbf{0.0 \pm 0.0}$	0.5 ± 1.4	$\mathbf{0.0 \pm 0.0}$
Donors	8.0 ± 1.1	7.1 ± 1.1	$\mathbf{6.7 \pm 1.1}$
Krvskp	$\mathbf{0.3 \pm 0.3}$	1.2 ± 0.7	0.6 ± 0.4
Monks	$\mathbf{4.2 \pm 5.9}$	10.0 ± 7.6	$\mathbf{4.2 \pm 5.9}$
Mushroom	$\mathbf{0.0 \pm 0.0}$	0.0 ± 0.1	$\mathbf{0.0 \pm 0.0}$
Promoters	23.6 ± 12.5	24.7 ± 12.9	$\mathbf{20.6 \pm 12.3}$
Spect	25.4 ± 7.9	$\mathbf{20.9 \pm 3.6}$	22.1 ± 7.1
Tictactoe	$\mathbf{14.2 \pm 3.8}$	18.5 ± 4.2	$\mathbf{14.2 \pm 3.8}$
Titanic	$\mathbf{21.0 \pm 1.7}$	21.5 ± 2.1	22.6 ± 2.1
Vote	6.3 ± 3.6	4.4 ± 2.9	$\mathbf{3.9 \pm 3.4}$

Complexity Control: Tree Complexity

Set	NOprune	CVprune	SRMprune
Acceptors	1015 ± 29	$\mathbf{55 \pm 42}$	838 ± 31
Artificial	16 ± 0	$\mathbf{15 \pm 2}$	16 ± 0
Donors	1489 ± 32	$\mathbf{145 \pm 35}$	910 ± 74
Krvskp	138 ± 6	$\mathbf{80 \pm 13}$	122 ± 9
Monks	121 ± 50	$\mathbf{57 \pm 17}$	121 ± 50
Mushroom	43 ± 0	$\mathbf{41 \pm 4}$	43 ± 0
Promoters	48 ± 5	$\mathbf{13 \pm 6}$	39 ± 3
Spect	165 ± 9	$\mathbf{5 \pm 10}$	60 ± 16
Tictactoe	437 ± 31	$\mathbf{123 \pm 25}$	436 ± 31
Titanic	32 ± 1	16 ± 4	$\mathbf{5 \pm 2}$
Vote	89 ± 8	$\mathbf{9 \pm 8}$	23 ± 8

Outline

Summary

- Try to fill the gap in the statistical learning theory, where there is no explicit formula for the VC-dimension of a decision tree.
- Give and prove lower bounds of the VC-dimension of different decision tree structures.
- Prove that the VC-dimension of a univariate decision tree depends on the number of features and the VC-dimension of the subtrees of it (tree structure).

Summary

- VC-dimension bounds are then used in pruning using SRM and compared with cross-validation pruning.
- SRM pruning works well and find trees that are as accurate as CV pruning without the overhead of cross-validation.

Questions?