

Beyond Candidate Elimination

MSC. BUI QUOC KHANH
KHANHBQ@HANU.EDU.VN

Classifying by using the VS

- There is 1 most specific hypothesis $h_1 = YZ$ and 1 most general hypothesis $h_2 = Z$
- No conjunction in between S_2 and G_2 , so as $VS = \{h_1, h_2\}$

Training set

X	Y	Z	W	C
0	1	1	0	1
1	1	1	1	1
0	1	0	0	0

Classifying by using the VS

Seen
instances

Unseen
instances

X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	?	0	0
0	1	0	1	?	0	0
0	0	0	1	?	0	0
0	1	1	1	?	1	1
1	0	0	0	?	0	0
1	0	0	1	?	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	?	0	0
1	1	0	1	?	0	0
1	1	1	0	?	1	1
0	0	1	0	?	0	1

Over the training set the class label according to YZ and Z coincides with the target function

Classifying by using the VS

Seen instances

Unseen instances

The class is unknown

The class according to YZ

The class according to Z

X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	?	0	0
0	1	0	1	?	0	0
0	0	0	1	?	0	0
0	1	1	1	?	1	1
1	0	0	0	?	0	0
1	0	0	1	?	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	?	0	0
1	1	0	1	?	0	0
1	1	1	0	?	1	1
0	0	1	0	?	0	1

Classifying by using the VS

Seen
instances

Unseen
instances


X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	?	0	0
0	1	0	1	?	0	0
0	0	0	1	?	0	0
0	1	1	1	?	1	1
1	0	0	0	?	0	0
1	0	0	1	?	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	?	0	0
1	1	0	1	?	0	0
1	1	1	0	?	1	1
0	0	1	0	?	0	1

What is the true target function?

If we assume that it is a conjunction, then it is either YZ or Z

Otherwise, it is one of the $2^{13} - 2$ remaining possible functions (not conjunctions)

Classifying by using the VS



X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	?	0	0
0	1	0	1	?	0	0
0	0	0	1	?	0	0
0	1	1	1	?	1	1
1	0	0	0	?	0	0
1	0	0	1	?	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	?	0	0
1	1	0	1	?	0	0
1	1	1	0	?	1	1
0	0	1	0	?	0	1

Assume that the target function is a conjunction, i.e., either YZ or Z

Since $Z \geq YZ$ then
 $YZ=1 \rightarrow Z=1$

The two models agree

Classifying by using the VS

Seen instances

Unseen instances

X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	?	0	0
0	1	0	1	?	0	0
0	0	0	1	?	0	0
0	1	1	1	1	1	1
1	0	0	0	?	0	0
1	0	0	1	?	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	?	0	0
1	1	0	1	?	0	0
1	1	1	0	1	1	1
0	0	1	0	?	0	1

Assume that the target function is a conjunction, i.e., either YZ or Z

Since $Z \geq YZ$ then
 $YZ=1 \rightarrow Z=1$

The two models agree

Hence, $C=1$

If an instance is classified positive by the most specific model YZ then it is classified positive with confidence 100%

Classifying by using the VS

Seen instances

Unseen instances

X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	0	0	0
0	1	0	1	0	0	0
0	0	0	1	0	0	0
0	1	1	1	1	1	1
1	0	0	0	0	0	0
1	0	0	1	0	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	0	0	0
1	1	0	1	0	0	0
1	1	1	0	1	1	1
0	0	1	0	?	0	1

Assume that the target function is a conjunction, i.e., either YZ or Z

Since $Z \geq YZ$ then
 $Z=0 \rightarrow YZ=0$

The two models agree

Hence, $C=0$

If an instance is classified negative by the most general model Z then it is classified negative with confidence 100%

Classifying by using the VS

Seen
instances

Unseen
instances

X	Y	Z	W	C	C(YZ)	C(Z)
0	1	1	0	1	1	1
1	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	1	1	?	0	1
0	0	0	0	0	0	0
0	1	0	1	0	0	0
0	0	0	1	0	0	0
0	1	1	1	1	1	1
1	0	0	0	0	0	0
1	0	0	1	0	0	0
1	0	1	0	?	0	1
1	0	1	1	?	0	1
1	1	0	0	0	0	0
1	1	0	1	0	0	0
1	1	1	0	1	1	1
0	0	1	0	?	0	1

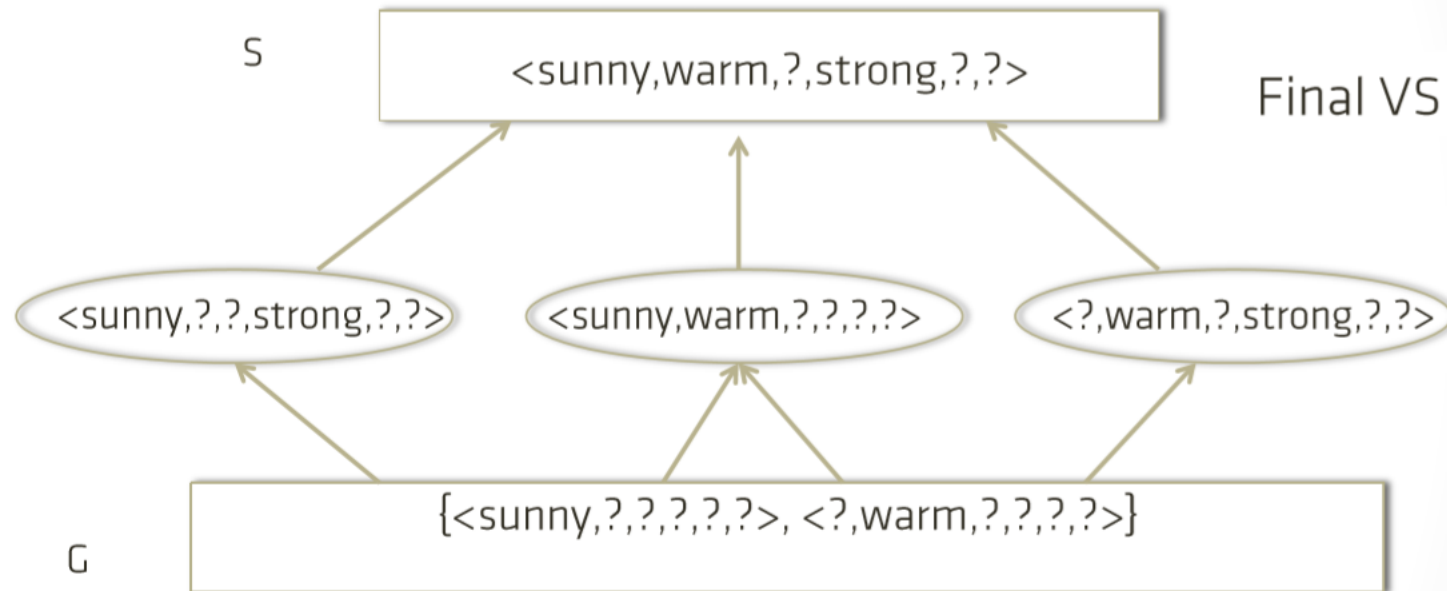
What about the other instances?

They can be classified as class 1 or 0 with the same probability

Classifying by using the VS

- Let's assume that the target function is a conjunction – either Z or YZ
 - A new instance e classified as positive by the most specific hypothesis YZ is classified positively by the most general hypothesis Z as well. Hence, e is classified positive with confidence 100%
 - A new instance e classified negatively by the most general hypotheses Z is classified as negative by the most specific hypothesis YZ as well. Then, e is classified negatively with confidence 100%
 - A new instance e classified positive by Z and negative by YZ will be classified as positive (or negative) with confidence 50%

Classifying by using the VS the EnjoySport example



All models (hypotheses compatible with the examples)

Classifying by using the VS the EnjoySport example

Unseen instances

Instance	Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
A	Sunny	Warm	Normal	Strong	Cool	Change	?
B	Rainy	Cold	Normal	Light	Warm	Same	?
C	Sunny	Warm	Normal	Light	Warm	Same	?
D	Sunny	Cold	Normal	Strong	Warm	Same	?

MODELS

- **<sunny,warm,?,strong,?,?>**
- <sunny,?,?,strong,?,?>
- <sunny,warm,?,?,?,?>
- <?,warm,?,strong,?,?>
- **<sunny,?,?,?,?,?>**
- **<?,warm,?,?,?,?,?>**

A: Yes 100% (it is classified by S)

B: No 100% (it is not classified by G)

C: classified by 3 classifiers of the VS out of 6 (50% confidence)

D: classified positive by 2 classifiers out of 6 (33% confidence)

Beyond Candidate Elimination - KhanhBQ

Classifying by using the VS the EnjoySport example

Unseen instances

Instance	Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
A	Sunny	Warm	Normal	Strong	Cool	Change	Yes (100%)
B	Rainy	Cold	Normal	Light	Warm	Same	No (100%)
C	Sunny	Warm	Normal	Light	Warm	Same	yes/ (50%)
D	Sunny	Cold	Normal	Strong	Warm	Same	No (66%)

MODELS

- **<sunny,warm,?,strong,?,?>**
- <sunny,?,?,strong,?,?>
- <sunny,warm,?,?,?,?>
- <?,warm,?,strong,?,?>
- **<sunny,?,?,?,?,?>**
- **<?,warm,?,?,?,?,?>**

A: Yes 100% (it is classified by S)

B: No 100% (it is not classified by G)

C: classified by 3 classifiers of the VS out of 6 (50% confidence)

D: classified positive by 2 classifiers out of 6 (33% confidence)

Beyond Candidate Elimination - KhanhBQ

Classifying by using the VS

- If a new instance x satisfies the most specific hypothesis in VS, then it satisfies all hypotheses in VS. Hence, x is classified with confidence 100% (as if the target function were known)
- If a new instance x does not satisfy any of the most general hypotheses in VS then it is classified as negative with confidence 100%
- If an instance x is classified positive by some of the hypotheses in VS we have less confidence in classifying it as positive than case 1 above

The non-determinism of the classification task

- Using only the training data, one can learn a number of hypotheses compatible with the training set (i.e., $h(x)=f(x)$, for each x)
- Classification uses data about the past to foresee the future. But what is the hypothesis that correctly predicts the future? No one knows!
- Induction is not truth-preserving, non-deterministic
- Inductive task: predict what is the next number in the series 1, 4, 9, 16, ?
- If the generator polynomial (prediction function) is n^2 , then the number is $n=25$
- If it is $(-5n^4 + 50n^3 - 151n^2 + 25n - 120)/24$, then $n=20$

Inductive learning assumption

- ***The Inductive Learning Assumption (ILA)***: Any hypothesis found to represent the target function well over a sufficiently large set of training examples will also approximate the target function well over unobserved examples

Beyond conjunctions

DNF hypotheses

- What is the probability that a concept is representable by a conjunction?
- Consider a training set S with n boolean attributes. There are
 - $x = 2^n$ possible instances
 - $y = 3^n$ conjunctive hypotheses constructible over S ($\langle ?, 0, \dots, 1 \rangle$, $\langle 0, ?, \dots, ? \rangle$, ...)
 - $z = 2^x$ concepts that can be defined over S (we are assuming binary classification)
- The probability is then y/z – with merely 5 attributes it is equal to $243/4.294.967.296$ --- very small!

Classifying by using the VS

X	Y	Z	W	C
0	1	1	0	?
1	1	1	1	?
0	1	0	0	?
0	0	1	1	?
0	0	0	0	?
0	1	0	1	?
0	0	0	1	?
0	1	1	1	?
1	0	0	0	?
1	0	0	1	?
1	0	1	0	?
1	0	1	1	?
1	1	0	0	?
1	1	0	1	?
1	1	1	0	?
0	0	1	0	?

2^{16} possible target functions

3^4 possible conjunctive functions

Prob = 0,001

Beyond conjunctions

DNF hypotheses

X	Y	Z	C
0	0	0	0
0	1	0	1
1	1	1	0
1	0	0	1

A disjunctive normal form (DNF) hypothesis is a disjunction of conjunctions of attribute constraints

DNF hypothesis space is **complete**

$$h = \langle 0, 1, 0 \rangle + \langle 1, 1, 1 \rangle = \bar{X}Y\bar{Z} + XYZ$$

$h = \bar{X}Y\bar{Z} + X\bar{Y}\bar{Z}$ is a model

Beyond conjunctions

DNF hypotheses

X	Y	Z	C
0	0	0	0
0	1	0	1
1	1	1	0
1	0	0	1

No conjunctive model
Use DNF

Find-S $\Rightarrow h = \bar{X}Y\bar{Z} + X\bar{Y}\bar{Z}$

Model: the disjunction of the observed
positive examples

Problem: any new instance is classified as negative

$h(1,0,1) = 0, h(0,0,1) = 0, \dots$

Useless, no generalization capability!

Using DNF hypotheses

X	Y	Z	C
0	0	0	0
0	1	0	1
1	1	1	0
1	0	0	1

$$\text{Cand-Elim} \rightarrow S = \{ \bar{X}\bar{Y}\bar{Z} + \bar{X}\bar{Y}Z \}$$

$$G = \{ \text{not}(\bar{X}\bar{Y}\bar{Z} + \bar{X}\bar{Y}Z) \}$$

Using DNF with CE produces trivial results:

S boundary the disjunction of the observed positive examples

G boundary the negation of the disjunction of the observed negative examples

Using DNF hypotheses

S classifies as positive all and only the positive examples of S

G classifies as positive whatever is not a negative example in S

$$S = \{ \bar{X}\bar{Y}\bar{Z} + \bar{X}\bar{Y}Z \}$$

$$G = \{ \text{not}(\bar{X}\bar{Y}\bar{Z} + \bar{X}\bar{Y}Z) \}$$

An unseen instance is classified as positive by G (50% confidence)

Any new instance is classified with confidence 50% - useless!!

Using DNF hypotheses

- The only instances classified with confidence 100% are those in the training set
- The others are classified with confidence 50%!!

The Futility of D.NF Learning

- If the DNF space is used, CE induces trivial classifiers (no generalization capabilities)
- $S = \{s\}$, with s = disjunction of positive examples
- $G = \{g\}$, with g = Negated disjunction of negative examples
- Only training examples will be unambiguously classified

Beyond Candidate Elimination

- Conjunctive hypothesis space too narrow –no conjunctive models may exist
- DNF space complete, but FindS and CE generate only trivial DNF models (no generalization capabilities)

Beyond CE algorithm

A	B	C	D	class
0	0	0	1	1
0	1	0	1	1
0	1	0	0	1
1	0	0	0	0
1	0	1	0	1
1	1	1	1	1

- CE outcome: Trivial DNF

$\bar{A}\bar{B}\bar{C}D \vee \bar{A}B\bar{C}D \vee \bar{A}B\bar{C}\bar{D} \vee \bar{A}B\bar{C}D \vee A\bar{B}\bar{C}D$
(S boundary induced by CE)

not ($\bar{A}\bar{B}\bar{C}\bar{D}$) (G boundary)

- zero errors over the TS
- no generalization capability (every new instance is classified as negative)

Algorithms that learn non-trivial DNF models needed!!

- Example - $\bar{A} \vee C$
 - zero errors over the TS (model)
 - more generalization capability (e.g., 0110 and 0111 will be classified positive)

Approximate Hypotheses

X	Y	Z	C
0	0	0	0
0	1	0	1
1	0	0	1

- DNF by CE: $S = \{\bar{X}\bar{Y}\bar{Z} \vee \bar{X}\bar{Y}Z\}$, $G = \text{not } (\bar{X} \bar{Y} \bar{Z})$ – useless
- No other DNF model exists

Approximate hypotheses:

- X: 1 error
- Y: 1 error
- not Z: 1 error
- not X: 2 errors
- not Y: 2 errors
- ...

Approximate Solutions

- Non-trivial models may not exist, so approximate solutions are needed
- Even when non-trivial models exist, approximate solutions are preferred because of the overfitting problem

Overfitting

- Real-world training data may be noisy (instances erroneously classified)
- Thus, a hypothesis that exactly fits the training data may be wrong and have bad generalization capabilities
- Overfitting occurs when the model is too tailored over the training data so as it may reflect its contingent properties rather than its structural properties

Overfitting due to noise

Mammal training data - $h = \langle \text{warm, yes, yes, ?} \rangle$ is a model

Name	Body Temp	Gives birth	4-legged	hibernates	mammal
Porcupine	warm	Yes	Yes	Yes	yes
cat	warm	Yes	Yes	no	yes
bat	warm	yes	no	yes	no
whale	warm	yes	No	No	No
salamander	cold	No	Yes	Yes	No
k. dragon	cold	No	Yes	No	No
Python	cold	No	No	Yes	No
Salmon	cold	No	No	No	No
Eagle	warm	No	No	No	no
Guppy	cold	Yes	No	No	no

Overfitting due to noise

Mammal training data - $h = \langle \text{warm, yes, yes, ?} \rangle$ is a model

Name	Body Temp	Gives birth	4-legged	hibernates	mammal
Porcupine	warm	Yes	Yes	Yes	yes
cat	warm	Yes	Yes	no	yes
bat	warm	yes	no	yes	no
whale	warm	yes	No	No	No
salamander	cold	No	Yes	Yes	no
k. dragon	cold	No	Yes	No	no
Python	cold	No	No	Yes	no
Salmon	cold	No	No	No	no
Eagle	warm	No	No	No	no
Guppy	cold	Yes	No	No	no

2 misclassified examples

Overfitting due to noise

Name	Body Temp	Gives birth	4-legged	hibernates	mammal	Predicted label
Human	warm	Yes	No	No	yes	no
pigeon	warm	No	No	No	no	no
elephant	warm	yes	Yes	No	yes	yes
Leopard seal	cold	yes	No	No	no	no
turtle	cold	No	Yes	No	no	no
penguin	cold	No	no	No	no	no
eel	cold	No	No	No	no	no
dolphin	warm	yes	No	No	yes	no

- Elephant is correctly classified as mammal
- Both humans and dolphins are misclassified – they are not 4-legged
- All other instances are correctly classified
- 25% test errors

Overfitting due to noise

Mammal training data - $h_1 = \langle \text{warm, yes, ?, ?} \rangle$ NOT a model

Name	Body Temp	Gives birth	4-legged	hibernates	mammal
Porcupine	warm	Yes	Yes	Yes	yes
cat	warm	Yes	Yes	no	yes
bat	warm	yes	no	yes	no
whale	warm	yes	No	No	No
salamander	cold	No	Yes	Yes	No
k. dragon	cold	No	Yes	No	No
Python	cold	No	No	Yes	No
Salmon	cold	No	No	No	No
Eagle	warm	No	No	No	no
Guppy	cold	Yes	No	No	no

Overfitting due to noise

Unseen instances

$h = \langle \text{warm, yes, ?, ?} \rangle$ - 0 test errors

Name	Body Temp	Gives birth	4-legged	hibernates	mammal	Predicted label
Human	warm	Yes	No	No	yes	yes
pigeon	warm	No	No	No	no	no
elephant	warm	yes	Yes	No	yes	yes
Leopard seal	cold	yes	No	No	no	no
turtle	cold	No	Yes	No	no	no
penguin	cold	No	no	No	no	no
eel	cold	No	No	No	no	no
dolphin	warm	yes	No	No	yes	yes

Overfitting due to noise

- $h = \langle \text{warm, yes, yes, ?} \rangle$
 - compatible with training data
 - 0% training errors, 25% test errors
- $h_1 = \langle \text{warm, yes, ?, ?} \rangle$
 - NOT compatible with training data
 - 20% training errors, 0% test errors
- h_1 (not a model) performs better than h (model) over unseen data!

Overfitting due to lack of examples

Training set for classifying mammals

name	Body temp	Gives birth	Four-legged	hibernates	mammal
salamander	cold	no	yes	yes	no
Guppy	cold	yes	no	no	no
Eagle	warm	No	No	No	no
Poorwill	warm	No	No	Yes	no
Platypus	warm	No	Yes	Yes	yes

- The only representative mammal is the platypus!!!
- $h = \langle \text{warm, no, yes, yes} \rangle$ is a model
- According to h both humans and elephants are NOT mammals!

Overfitting

- **Definition (Mitchell):** a hypothesis h is said to overfit the training data if there is another hypothesis h' such that h has a smaller error over the training examples, but h' has a smaller error over the unseen examples

Inductive learning assumption

- Real-world training data are noisy
- So, **approximate** solutions are in general preferred
- Search problem: Find $h \in H$ that **well approximates** the training set
- ***The Inductive Learning Assumption (ILA)***: Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples.

Conclusions

- Find-S and CE algorithms
 - Conjunctive Space: models may not exist
 - DNF space: trivial solutions (no generalization capabilities)
 - Not robust to overfitting
- Algorithms generating non-trivial DNF models are needed
- Approximate solutions are preferred – the overfitting problem
- Need to devise learning algorithms capable of inducing models that well approximate the training data and are robust to overfitting