http://goo.gl/rP1KSW

# A Peek into the Black Box: Exploring Classifiers by Randomization


Andreas Henelius
Kai Puolamäki

Finnish Institute of Occupational Health



Henrik Boström
Lars Asker
Panagiotis Papapetrou

Stockholm University


Finnish Institute of Occupational Health

ECML PKDD 2014 - 18 September 2014

$$y = f(x)$$

Classifier

Class

{ healthy, ill } → $$y = f(x)$$

(body temperature, gender, ...)

Vector of attributes

Classifier

Class
{ healthy, ill }

$$y = f(x)$$

(body temperature, gender, …)
Vector of attributes

Properties of a good classifier include:

- high accuracy
- interpretability (e.g., *why is a person healthy?*)

# Peek into the Black Box

$$y = f(x)$$

- **Black box** classifier: the form of $f$ is impossible to interpret
- Even if we can understand the parameters of $f$, we may still not understand how the classifier uses the data *(example later!)*

# Assumption

$$y = f(x)$$

- We don't know the form of $f$
- We can test the classifier $f$ with data of our choosing

- **Idea and problem formulation**
- The GoldenEye algorithm
- Experiments
- Concluding remarks

| Class | *A* | *B* | *C* | *D* |
|---|---|---|---|---|
| **1** | 1 | 0 | 1 | 1 |
| **1** | 1 | 0 | 1 | 0 |
| **1** | 0 | 1 | 1 | 1 |
| **1** | 0 | 1 | 1 | 0 |
| **1** | 0 | 0 | 1 | 1 |
| **1** | 0 | 0 | 1 | 0 |
| **1** | 1 | 1 | 1 | 1 |
| **1** | 1 | 1 | 1 | 0 |
| **1** | 1 | 0 | 0 | 1 |
| **1** | 1 | 0 | 0 | 0 |
| **1** | 0 | 1 | 0 | 1 |
| **1** | 0 | 1 | 0 | 0 |
| **0** | 1 | 1 | 0 | 1 |
| **0** | 1 | 1 | 0 | 0 |
| **0** | 0 | 0 | 0 | 1 |
| **0** | 0 | 0 | 0 | 0 |

$$\text{Class} = (A \oplus B) \vee C$$

| Class = ( | A XOR | B ) OR | C | D |
|---|---|---|---|---|
| **1** | 1 | 0 | 1 | 1 |
| **1** | 1 | 0 | 1 | 0 |
| **1** | 0 | 1 | 1 | 1 |
| **1** | 0 | 1 | 1 | 0 |
| **1** | 0 | 0 | 1 | 1 |
| **1** | 0 | 0 | 1 | 0 |
| **1** | 1 | 1 | 1 | 1 |
| **1** | 1 | 1 | 1 | 0 |
| **1** | 1 | 0 | 0 | 1 |
| **1** | 1 | 0 | 0 | 0 |
| **1** | 0 | 1 | 0 | 1 |
| **1** | 0 | 1 | 0 | 0 |
| **0** | 1 | 1 | 0 | 1 |
| **0** | 1 | 1 | 0 | 0 |
| **0** | 0 | 0 | 0 | 1 |
| **0** | 0 | 0 | 0 | 0 |

Training a black box classifier…

Spoiler:

$$f(x) = (A \oplus B) \vee C$$

$$y = f(x)$$

| Class | y | *A* | *B* | *C* | *D* |
|---|---|---|---|---|---|
| **1** | **1** | 1 | 0 | 1 | 1 |
| **1** | **1** | 1 | 0 | 1 | 0 |
| **1** | **1** | 0 | 1 | 1 | 1 |
| **1** | **1** | 0 | 1 | 1 | 0 |
| **1** | **1** | 0 | 0 | 1 | 1 |
| **1** | **1** | 0 | 0 | 1 | 0 |
| **1** | **1** | 1 | 1 | 1 | 1 |
| **1** | **1** | 1 | 1 | 1 | 0 |
| **1** | **1** | 1 | 0 | 0 | 1 |
| **1** | **1** | 1 | 0 | 0 | 0 |
| **1** | **1** | 0 | 1 | 0 | 1 |
| **1** | **1** | 0 | 1 | 0 | 0 |
| **0** | **0** | 1 | 1 | 0 | 1 |
| **0** | **0** | 1 | 1 | 0 | 0 |
| **0** | **0** | 0 | 0 | 0 | 1 |
| **0** | **0** | 0 | 0 | 0 | 0 |

In this case accuracy = 100%

$$y^* = f(x^*)$$

| y | y* | *A* | *B* | *C* | *D* |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | 0 |
| **1** | ***1*** | 1 | 0 | 1 | 1 |
| **1** | ***1*** | 0 | 1 | 1 | 0 |
| **1** | ***1*** | 0 | 1 | 1 | 1 |
| **1** | ***1*** | 0 | 0 | 1 | 0 |
| **1** | ***1*** | 0 | 0 | 1 | 1 |
| **1** | ***1*** | 1 | 1 | 1 | 0 |
| **1** | ***1*** | 1 | 1 | 1 | 1 |
| **1** | ***1*** | 1 | 0 | 0 | 0 |
| **1** | ***1*** | 1 | 0 | 0 | 1 |
| **1** | ***1*** | 0 | 1 | 0 | 0 |
| **1** | ***1*** | 0 | 1 | 0 | 1 |
| **0** | ***0*** | 1 | 1 | 0 | 0 |
| **0** | ***0*** | 1 | 1 | 0 | 1 |
| **0** | ***0*** | 0 | 0 | 0 | 0 |
| **0** | ***0*** | 0 | 0 | 0 | 1 |

# Randomization 1

$$y^* = f(x^*)$$

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | *1* |
| **1** | ***1*** | 1 | 0 | 1 | *1* |
| **1** | ***1*** | 0 | 1 | 1 | *0* |
| **1** | ***1*** | 0 | 1 | 1 | *1* |
| **1** | ***1*** | 0 | 0 | 1 | *1* |
| **1** | ***1*** | 0 | 0 | 1 | *0* |
| **1** | ***1*** | 1 | 1 | 1 | *1* |
| **1** | ***1*** | 1 | 1 | 1 | *0* |
| **1** | ***1*** | 1 | 0 | 0 | *0* |
| **1** | ***1*** | 1 | 0 | 0 | *1* |
| **1** | ***1*** | 0 | 1 | 0 | *0* |
| **1** | ***1*** | 0 | 1 | 0 | *0* |
| **0** | ***0*** | 1 | 1 | 0 | *1* |
| **0** | ***0*** | 1 | 1 | 0 | *0* |
| **0** | ***0*** | 0 | 0 | 0 | *0* |
| **0** | ***0*** | 0 | 0 | 0 | *1* |

# Randomization 2

$$y^* = f(x^*)$$

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | *0* |
| **1** | ***1*** | 1 | 0 | 1 | *0* |
| **1** | ***1*** | 0 | 1 | 1 | *0* |
| **1** | ***1*** | 0 | 1 | 1 | *0* |
| **1** | ***1*** | 0 | 0 | 1 | *0* |
| **1** | ***1*** | 0 | 0 | 1 | *1* |
| **1** | ***1*** | 1 | 1 | 1 | *1* |
| **1** | ***1*** | 1 | 1 | 1 | *1* |
| **1** | ***1*** | 1 | 0 | 0 | *1* |
| **1** | ***1*** | 1 | 0 | 0 | *1* |
| **1** | ***1*** | 0 | 1 | 0 | *0* |
| **1** | ***1*** | 0 | 1 | 0 | *0* |
| **0** | ***0*** | 1 | 1 | 0 | *1* |
| **0** | ***0*** | 1 | 1 | 0 | *1* |
| **0** | ***0*** | 0 | 0 | 0 | *1* |
| **0** | ***0*** | 0 | 0 | 0 | *0* |

$$y^* = f(x^*)$$

# Randomization 3

| **y** | ***y**** | ***A*** | ***B*** | ***C*** | ***D*** |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | *1* |
| **1** | ***1*** | 1 | 0 | 1 | *1* |
| **1** | ***1*** | 0 | 1 | 1 | *1* |
| **1** | ***1*** | 0 | 1 | 1 | *0* |
| **1** | ***1*** | 0 | 0 | 1 | *0* |
| **1** | ***1*** | 0 | 0 | 1 | *1* |
| **1** | ***1*** | 1 | 1 | 1 | *0* |
| **1** | ***1*** | 1 | 1 | 1 | *0* |
| **1** | ***1*** | 1 | 0 | 0 | *1* |
| **1** | ***1*** | 1 | 0 | 0 | *1* |
| **1** | ***1*** | 0 | 1 | 0 | *0* |
| **1** | ***1*** | 0 | 1 | 0 | *1* |
| **0** | ***0*** | 1 | 1 | 0 | *0* |
| **0** | ***0*** | 1 | 1 | 0 | *0* |
| **0** | ***0*** | 0 | 0 | 0 | *1* |
| **0** | ***0*** | 0 | 0 | 0 | *0* |

$$\text{fidelity} = \#(y = y^*)/N = 1$$

Finnish Institute of Occupational Health

Kai Puolamäki

$$y^* = f(x^*)$$

| y | y* | *A* | *B* | *C* | *D* |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | 0 |
| **1** | ***1*** | 1 | 0 | 1 | 1 |
| **1** | ***1*** | 0 | 1 | 1 | 0 |
| **1** | ***1*** | 0 | 1 | 1 | 1 |
| **1** | ***1*** | 0 | 0 | 1 | 0 |
| **1** | ***1*** | 0 | 0 | 1 | 1 |
| **1** | ***1*** | 1 | 1 | 1 | 0 |
| **1** | ***1*** | 1 | 1 | 1 | 1 |
| **1** | ***1*** | 1 | 0 | 0 | 0 |
| **1** | ***1*** | 1 | 0 | 0 | 1 |
| **1** | ***1*** | 0 | 1 | 0 | 0 |
| **1** | ***1*** | 0 | 1 | 0 | 1 |
| **0** | ***0*** | 1 | 1 | 0 | 0 |
| **0** | ***0*** | 1 | 1 | 0 | 1 |
| **0** | ***0*** | 0 | 0 | 0 | 0 |
| **0** | ***0*** | 0 | 0 | 0 | 1 |

$$y^* = f(x^*)$$

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | *0* | 1 |
| **1** | ***1*** | 1 | 0 | *1* | 0 |
| **1** | ***1*** | 0 | 1 | *0* | 1 |
| **1** | ***1*** | 0 | 1 | *1* | 0 |
| **1** | ***0*** | 0 | 0 | *0* | 1 |
| **1** | ***0*** | 0 | 0 | *0* | 0 |
| **1** | ***1*** | 1 | 1 | *1* | 1 |
| **1** | ***0*** | 1 | 1 | *0* | 0 |
| **1** | ***1*** | 1 | 0 | *0* | 1 |
| **1** | ***1*** | 1 | 0 | *1* | 0 |
| **1** | ***1*** | 0 | 1 | *1* | 1 |
| **1** | ***1*** | 0 | 1 | *0* | 0 |
| **0** | ***0*** | 1 | 1 | *0* | 1 |
| **0** | ***1*** | 1 | 1 | *1* | 0 |
| **0** | ***1*** | 0 | 0 | *1* | 1 |
| **0** | ***1*** | 0 | 0 | *1* | 0 |

$$\text{fidelity} = \#(y = y^*)/N = 0.63$$

$$y^* = f(x^*)$$

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | 0 |
| **1** | ***1*** | 1 | 0 | 1 | 1 |
| **1** | ***1*** | 0 | 1 | 1 | 0 |
| **1** | ***1*** | 0 | 1 | 1 | 1 |
| **1** | ***1*** | 0 | 0 | 1 | 0 |
| **1** | ***1*** | 0 | 0 | 1 | 1 |
| **1** | ***1*** | 1 | 1 | 1 | 0 |
| **1** | ***1*** | 1 | 1 | 1 | 1 |
| **1** | ***1*** | 1 | 0 | 0 | 0 |
| **1** | ***1*** | 1 | 0 | 0 | 1 |
| **1** | ***1*** | 0 | 1 | 0 | 0 |
| **1** | ***1*** | 0 | 1 | 0 | 1 |
| **0** | ***0*** | 1 | 1 | 0 | 0 |
| **0** | ***0*** | 1 | 1 | 0 | 1 |
| **0** | ***0*** | 0 | 0 | 0 | 0 |
| **0** | ***0*** | 0 | 0 | 0 | 1 |

# Within-class randomization

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | *1* | 1 |
| **1** | ***1*** | 1 | 0 | *1* | 0 |
| **1** | ***1*** | 0 | 1 | *0* | 1 |
| **1** | ***1*** | 0 | 1 | *0* | 0 |
| **1** | ***1*** | 0 | 0 | *1* | 1 |
| **1** | ***1*** | 0 | 0 | *1* | 0 |
| **1** | ***0*** | 1 | 1 | *0* | 1 |
| **1** | ***1*** | 1 | 1 | *1* | 0 |
| **1** | ***1*** | 1 | 0 | *0* | 1 |
| **1** | ***1*** | 1 | 0 | *0* | 0 |
| **1** | ***1*** | 0 | 1 | *1* | 1 |
| **1** | ***1*** | 0 | 1 | *0* | 0 |
| **0** | ***0*** | 1 | 1 | *0* | 1 |
| **0** | ***0*** | 1 | 1 | *0* | 0 |
| **0** | ***0*** | 0 | 0 | *0* | 1 |
| **0** | ***0*** | 0 | 0 | *0* | 0 |

$$\text{fidelity} = \#(y = y^*)/N = 0.94$$

$$Pr(A,B,C,D \mid y) \approx Pr(C \mid y) \times Pr(A,B,D \mid y)$$

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | *1* | 1 |
| **1** | ***1*** | 1 | 0 | *1* | 0 |
| **1** | ***1*** | 0 | 1 | *0* | 1 |
| **1** | ***1*** | 0 | 1 | *0* | 0 |
| **1** | ***1*** | 0 | 0 | *1* | 1 |
| **1** | ***1*** | 0 | 0 | *1* | 0 |
| **1** | ***0*** | 1 | 1 | *0* | 1 |
| **1** | ***1*** | 1 | 1 | *1* | 0 |
| **1** | ***1*** | 1 | 0 | *0* | 1 |
| **1** | ***1*** | 1 | 0 | *0* | 0 |
| **1** | ***1*** | 0 | 1 | *1* | 1 |
| **1** | ***1*** | 0 | 1 | *0* | 0 |
| **0** | ***0*** | 1 | 1 | *0* | 1 |
| **0** | ***0*** | 1 | 1 | *0* | 0 |
| **0** | ***0*** | 0 | 0 | *0* | 1 |
| **0** | ***0*** | 0 | 0 | *0* | 0 |

$$\text{fidelity} = \#(y = y^*)/N = 0.94$$

Finnish Institute of
Occupational Health

Kai Puolamäki

$$y^* = f(x^*)$$

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| **1** | ***1*** | 1 | 0 | 1 | 0 |
| **1** | ***1*** | 1 | 0 | 1 | 1 |
| **1** | ***1*** | 0 | 1 | 1 | 0 |
| **1** | ***1*** | 0 | 1 | 1 | 1 |
| **1** | ***1*** | 0 | 0 | 1 | 0 |
| **1** | ***1*** | 0 | 0 | 1 | 1 |
| **1** | ***1*** | 1 | 1 | 1 | 0 |
| **1** | ***1*** | 1 | 1 | 1 | 1 |
| **1** | ***1*** | 1 | 0 | 0 | 0 |
| **1** | ***1*** | 1 | 0 | 0 | 1 |
| **1** | ***1*** | 0 | 1 | 0 | 0 |
| **1** | ***1*** | 0 | 1 | 0 | 1 |
| **0** | ***0*** | 1 | 1 | 0 | 0 |
| **0** | ***0*** | 1 | 1 | 0 | 1 |
| **0** | ***0*** | 0 | 0 | 0 | 0 |
| **0** | ***0*** | 0 | 0 | 0 | 1 |

# 2 independent within-class randomizations

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | 0 |

$$\text{fidelity} = \#(y = y^*)/N = 0.75$$

# 2 joint within-class randomizations

| **y** | ***y**** | ***A*** | ***B*** | ***C*** | ***D*** |
|---|---|---|---|---|---|
| **1** | ***1*** | *0* | *1* | 1 | 1 |
| **1** | ***1*** | *0* | *1* | 1 | 0 |
| **1** | ***1*** | *0* | *1* | 1 | 1 |
| **1** | ***1*** | *1* | *0* | 1 | 0 |
| **1** | ***1*** | *0* | *0* | 1 | 1 |
| **1** | ***1*** | *1* | *0* | 1 | 0 |
| **1** | ***1*** | *1* | *1* | 1 | 1 |
| **1** | ***1*** | *1* | *1* | 1 | 0 |
| **1** | ***1*** | *1* | *0* | 0 | 1 |
| **1** | ***1*** | *1* | *0* | 0 | 0 |
| **1** | ***1*** | *0* | *1* | 0 | 1 |
| **1** | ***0*** | *0* | *0* | 0 | 0 |
| **0** | ***0*** | *0* | *0* | 0 | 1 |
| **0** | ***0*** | *1* | *1* | 0 | 0 |
| **0** | ***0*** | *0* | *0* | 0 | 1 |
| **0** | ***0*** | *1* | *1* | 0 | 0 |

$$\text{fidelity} = \#(y = y^*)/N = 0.94$$

Finnish Institute of Occupational Health

Kai Puolamäki

# Grouping of attributes

- *D* neither used nor needed
- *C* used and needed
- *C* independent of other variables
- *A* and *B* both important, must occur together

$$\{\{A, B\}, \{C\}\}$$

$$f(x) = (A \oplus B) \vee C$$

Finnish Institute of Occupational Health
Kai Puolamäki

The grouping { { *A* , *B* }, { *C* } } means that

- *A* and *B* randomized together within-class
- *C* is randomized within-class
- *D* is fully randomized

| y | y* | A | B | C | D |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 | 1 |

# Problem formulations

***Optimal k-grouping of attributes.***
Given a dataset, a classifier, and a constant *k*, find a grouping of attributes of size *k* such that the fidelity is maximized.

$$\{\{A, B\}, \{C\}, \{D\}\}$$

# Problem formulations

***Optimal k-grouping of attributes.***
Given a dataset, a classifier, and a constant *k*, find a grouping of attributes of size *k* such that the fidelity is maximized.

$$\{\{A,B\},\{C\},\{D\}\}$$

***Optimal pruning of singleton attributes.***

$$\{\{A,B\},\{C\}\}$$

- Idea and problem formulation
- **The GoldenEye algorithm**
- Experiments
- Concluding remarks

# The GoldenEye algorithm

- Finds a grouping of attributes
- Greedy iterative top-down algorithm
- GoldenEye can find the optimal solution, if *monotonicity* holds (breaking groups appearing in "optimal solution" decreases fidelity)

# The GoldenEye algorithm

$\{ \{ A, B, C, D \} \}$

fidelity = 1

# The GoldenEye algorithm

{ { *A*, *B*, *C*, *D* } }
fidelity = 1

{ { *B*, *C*, *D* }, { *A* } }
fidelity = 0.75

# The GoldenEye algorithm

$\{\{A, B, C, D\}\}$

fidelity = 1

$\{\{B, C, D\}, \{A\}\}$

fidelity = 0.75

$\{\{A, B, C\}, \{B\}\}$

fidelity = 0.75

# The GoldenEye algorithm

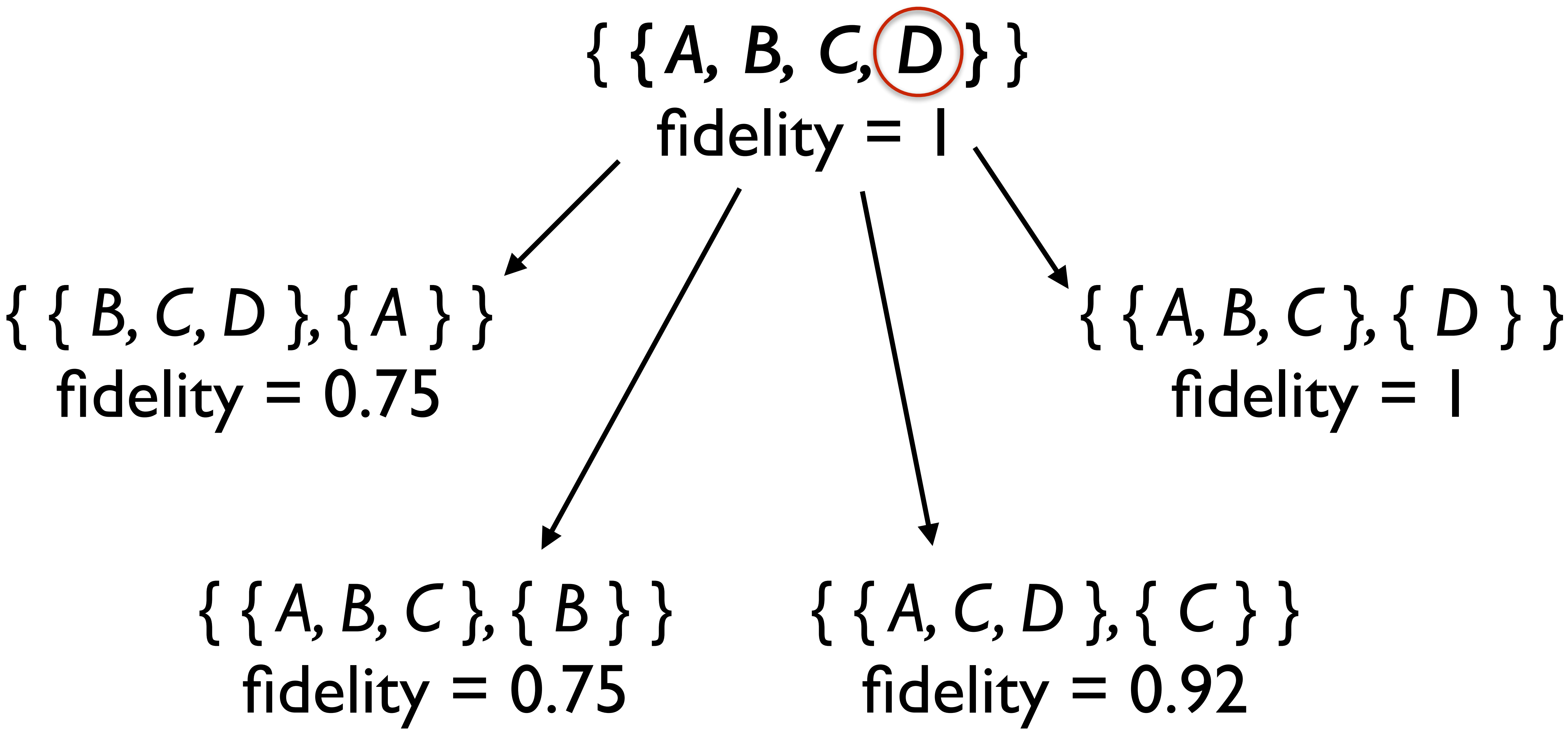
{ { *A*, *B*, *C*, *D* } }
fidelity = 1

{ { *B*, *C*, *D* }, { *A* } }
fidelity = 0.75

{ { *A*, *B*, *C* }, { *B* } }
fidelity = 0.75

{ { *A*, *C*, *D* }, { *C* } }
fidelity = 0.92

# The GoldenEye algorithm




Finnish Institute of Occupational Health



Kai Puolamäki

# The GoldenEye algorithm

$\{ \{ A, B, C, D \} \}$
fidelity = 1

$\{ \{ B, C, D \}, \{ A \} \}$
fidelity = 0.75

$\{ \{ A, B, C \}, \{ B \} \}$
fidelity = 0.75

$\{ \{ A, C, D \}, \{ C \} \}$
fidelity = 0.92

$\{ \{ A, B, C \}, \{ D \} \}$
fidelity = 1

# The GoldenEye algorithm

$\{ \{ A, B, C \}, \{ D \} \}$

fidelity = 1

# The GoldenEye algorithm

{ { *A*, *B*, *C* }, { *D* } }

fidelity = 1

{ { *B*, *C* }, { *A* }, { *D* } }

fidelity = 0.75

# The GoldenEye algorithm

$\{\, \{ A, B, C \}, \{ D \} \,\}$

fidelity = 1

$\{\, \{ B, C \}, \{ A \}, \{ D \} \,\}$

fidelity = 0.75

$\{\, \{ A, C \}, \{ B \}, \{ D \} \,\}$

fidelity = 0.75

# The GoldenEye algorithm

$\{\{A, B, C\}, \{D\}\}$

fidelity = 1

$\{\{B, C\}, \{A\}, \{D\}\}$

fidelity = 0.75

$\{\{A, B\}, \{C\}, \{D\}\}$

fidelity = 0.92

$\{\{A, C\}, \{B\}, \{D\}\}$

fidelity = 0.75

# The GoldenEye algorithm

$\{\{A, B, C\}, \{D\}\}$
fidelity = 1

$\{\{B, C\}, \{A\}, \{D\}\}$
fidelity = 0.75

$\{\{A, B\}, \{C\}, \{D\}\}$
fidelity = 0.92

$\{\{A, C\}, \{B\}, \{D\}\}$
fidelity = 0.75

# The GoldenEye algorithm

$\{ \{ A, B \}, \{ C \}, \{ D \} \}$

fidelity = 0.92

# The GoldenEye algorithm

$\{ \{ A, B \}, \{ C \}, \{ D \} \}$
fidelity = 0.92

$\{ \{ A \}, \{ B \}, \{ C \}, \{ D \} \}$
fidelity = 0.75

# The GoldenEye algorithm

{ { *A*, *B* }, { *C* }, { *D* } }
fidelity = 0.92

~~{ { *A* }, { *B* }, { *C* }, { *D* } }
fidelity = 0.75~~

# The GoldenEye algorithm

{ { *A*, *B* }, { *C* }, { *D* } }
fidelity = 0.92

~~{ { *A* }, { *B* }, { *C* }, { *D* } }
fidelity = 0.75~~

Output { *A*, *B* }

# The GoldenEye algorithm

{ { *C*, *D* }, { *A* }, { *B* } }

fidelity = 0.75

Finnish Institute of Occupational Health

# The GoldenEye algorithm

{ { *C*, *D* }, { *A* }, { *B* } }
fidelity = 0.75

{ { *C* }, { *D* }, { *A* }, { *B* } }
fidelity = 0.75

# The GoldenEye algorithm

{ { *C*, *D* }, { *A* }, { *B* } }
fidelity = 0.75

{ { *C* }, { *D* }, { *A* }, { *B* } }
fidelity = 0.75

→

Output
{ *C* } and { *D* }

# The GoldenEye algorithm

{ { *C*, *D* }, { *A* }, { *B* } }
fidelity = 0.75

{ { *C* }, { *D* }, { *A* }, { *B* } }
fidelity = 0.75

Output
{ *C* } and { *D* }

Final result:
{ { *A*, *B* }, { *C* }, { *D* } }

# The GoldenEye algorithm

Finally, unnecessary singletons are pruned.

# The GoldenEye algorithm

Finally, unnecessary singletons are pruned.

Randomising $D$ fully does not reduce fidelity, hence singleton $D$ can be pruned. $\{\{A, B\}, \{C\}\}$

# The GoldenEye algorithm

Finally, unnecessary singletons are pruned.

Randomising *D* fully does not reduce fidelity, hence singleton *D* can be pruned. { { *A* , *B* }, { *C* } }

Randomising C fully reduces fidelity too much, hence singleton C can't be pruned.

**Final output { { *A* , *B* }, { *C* } }.**

# The GoldenEye algorithm

- Efficient implementation using random sampling and permutations
- Easily parallelizable
- 0 to 2 parameters
- Running time:
  - Constant in number of data items
  - Quadratic in number of attributes

- Idea and problem formulation
- The GoldenEye algorithm
- **Experiments**
- Concluding remarks

# Experiments

- 26 data sets (synthetic and UCI)
- 15 commonly used classifiers

# Previous toy data with noise

| | Acc | | *A* | *B* | *C* | *D* |
|---|---|---|---|---|---|---|
| "Correct" | | *{ { A, B }, { C } }* | X | X | o | . |
| DecisionStump | 0.74 | *{ }* | . | . | . | . |
| OneR | 0.74 | *{ }* | . | . | . | . |
| SMO | 0.74 | *{ }* | . | . | . | . |
| naiveBayes | 0.72 | *{ { C } }* | . | . | o | . |
| AdaBoostM1 | 0.69 | *{ { A, B, C, D} }* | X | X | X | X |
| Logistic | 0.69 | *{ { A, B, C, D} }* | X | X | X | X |
| LogitBoost | 0.69 | *{ { A, B, C, D} }* | X | X | X | X |
| Bagging | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| IBk | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| J48 | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| JRip | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| LMT | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| PART | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| SMO radial | 0.91 | *{ { A, B}, { C} }* | X | X | o | . |
| randomForest | 0.90 | *{ { A, B}, { C} }* | X | X | o | . |

# UCI glass data

| | Acc | *mg* | *al* | *ri* | *si* | *na* | *fe* | *ca* | *k* | *ba* |
|---|---|---|---|---|---|---|---|---|---|---|
| OneR | 0.52 | . | o | . | . | . | . | . | . | . |
| JRip | 0.55 | . | o | . | . | . | . | o | o | . |
| SMO | 0.51 | X | X | X | . | . | X | X | o | . |
| J48 | 0.58 | X | X | X | . | X | X | | | o |
| randomForest | 0.73 | X | X | X | X | X | X | X | X | . |
| naiveBayes | 0.52 | X | X | X | X | X | X | X | X | X |
| Bagging | 0.72 | X | X | X | X | X | X | o | . | . |
| PART | 0.63 | X | X | X | X | X | o | . | o | o |
| IBk | 0.69 | X | X | X | X | o | X | o | . | . |
| SMO radial | 0.66 | X | X | X | X | o | X | o | o | . |
| LMT | 0.55 | X | X | X | X | o | o | . | X | o |
| Logistic | 0.56 | X | o | . | X | X | . | X | o | o |
| AdaBoostM1 | 0.47 | o | . | . | . | . | . | . | . | . |
| DecisionStump | 0.47 | o | . | . | . | . | . | . | . | . |
| LogitBoost | 0.65 | o | X | X | X | X | . | X | o | o |

# UCI glass data

| | Acc | *mg* | *al* | *ri* | *si* | *na* | *fe* | *ca* | *k* | *ba* |
|---|---|---|---|---|---|---|---|---|---|---|
| OneR | 0.52 | . | o | . | . | . | . | . | . | . |
| JRip | 0.55 | . | o | . | . | . | . | o | o | . |
| SMO | 0.51 | X | X | X | . | . | X | X | o | . |
| J48 | 0.58 | X | X | X | . | X | X | | | o |
| randomForest | 0.73 | X | X | X | X | X | X | X | X | |
| naiveBayes | 0.52 | X | X | X | X | X | X | X | X | X |
| Bagging | 0.72 | X | X | X | X | X | X | o | . | . |
| PART | 0.63 | X | X | X | X | X | o | . | o | o |
| IBk | 0.69 | X | X | X | X | o | X | o | . | . |
| SMO radial | 0.66 | X | X | X | X | o | X | o | o | . |
| LMT | 0.55 | X | X | X | X | o | o | . | X | o |
| Logistic | 0.56 | X | o | . | X | X | . | X | o | o |
| AdaBoostM1 | 0.47 | o | . | . | . | . | . | . | . | . |
| DecisionStump | 0.47 | o | . | . | . | . | . | . | . | . |
| LogitBoost | 0.65 | o | X | X | X | X | . | X | o | o |

# UCI glass data

| | Acc | *mg* | *al* | *ri* | *si* | *na* | *fe* | *ca* | *k* | *ba* |
|---|---|---|---|---|---|---|---|---|---|---|
| OneR | 0.52 | . | o | . | . | . | . | . | . | . |
| JRip | 0.55 | . | o | . | . | . | . | o | o | . |
| SMO | 0.51 | X | X | X | . | . | X | X | o | . |
| J48 | 0.58 | X | X | X | . | X | X | | | o |
| randomForest | 0.73 | X | X | X | X | X | X | X | X | . |
| naiveBayes | 0.52 | X | X | X | X | X | X | X | X | X |
| Bagging | 0.72 | X | X | X | X | X | X | o | . | . |
| PART | 0.63 | X | X | X | X | X | o | . | o | o |
| IBk | 0.69 | X | X | X | X | o | X | o | . | . |
| SMO radial | 0.66 | X | X | X | X | o | X | o | o | . |
| LMT | 0.55 | X | X | X | X | o | o | . | X | o |
| Logistic | 0.56 | X | o | . | X | X | . | X | o | o |
| AdaBoostM1 | 0.47 | o | . | . | . | . | . | . | . | . |
| DecisionStump | 0.47 | o | . | . | . | . | . | . | . | . |
| LogitBoost | 0.65 | o | X | X | X | X | . | X | o | o |

Finnish Institute of Occupational Health



Kai Puolamäki

# Experiments not shown

- More datasets
- Stability of groupings
- Effects of the parameters to the GoldenEye
- Comparison to attribute selection

- Idea and problem formulation
- The GoldenEye algorithm
- Experiments
- **Concluding remarks**

# Understanding parameters is not enough

- It is not enough to understand the parameters of the classifier
- The structure of data affects classification results
- Example: Naive Bayes binary classifier with 2 binary attributes benefits from correlations!

# Conclusion

- A method based on randomization to find out how a classifier uses the data
  - It is not enough to just to understand the classifier, the structure of the data matters, too!
- Groupings are useful for exploration, maybe to improve classifiers
- Download our GoldenEye R package and come see our poster!

# Conclusion

- A method based on randomization to find out how a classifier uses the data
  - It is not enough to just to understand the classifier, the structure of the data matters, too!
- Groupings are useful for exploration, maybe to improve classifiers
- Download our GoldenEye R package and come see our poster!

Finnish Institute of Occupational Health

Thank you!

Kai Puolamäki

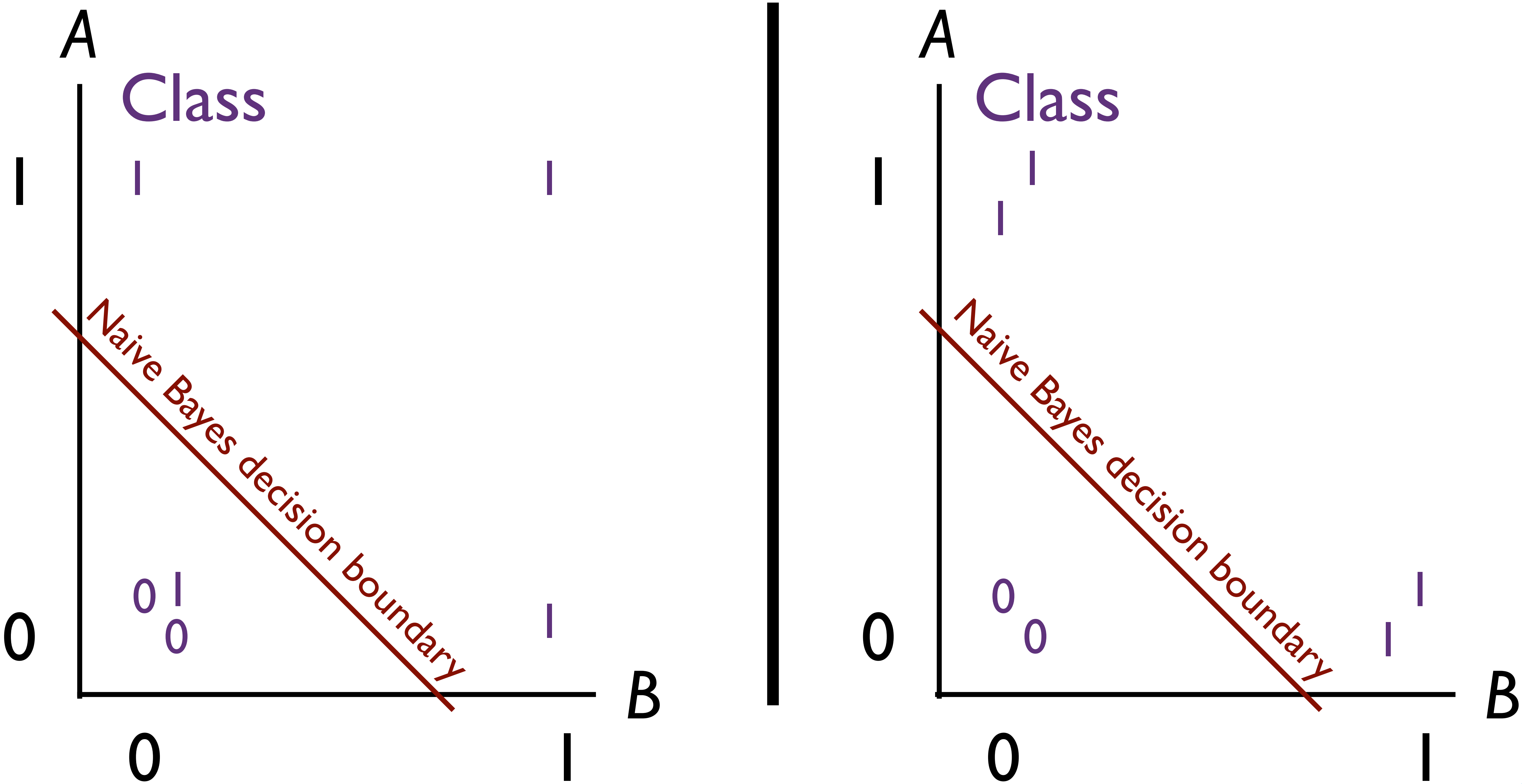
## Data attributes *A* and *B* independent for a given class


## Data attributes *A* and *B* independent for class 0 but correlated for class 1


Finnish Institute of Occupational Health
Kai Puolamäki

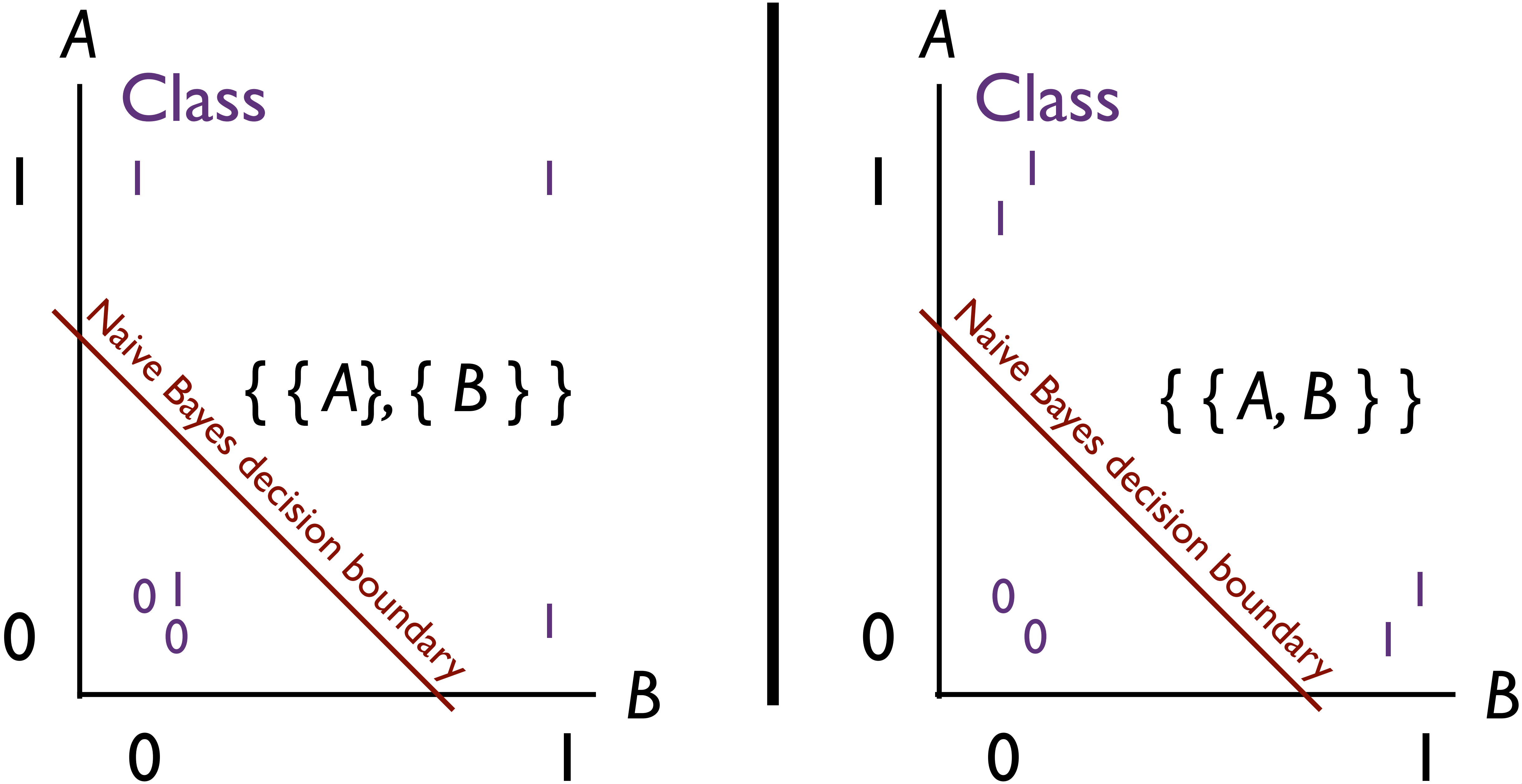
Data attributes *A* and *B* independent for a given class

Data attributes *A* and *B* independent for class 0 but correlated for class 1

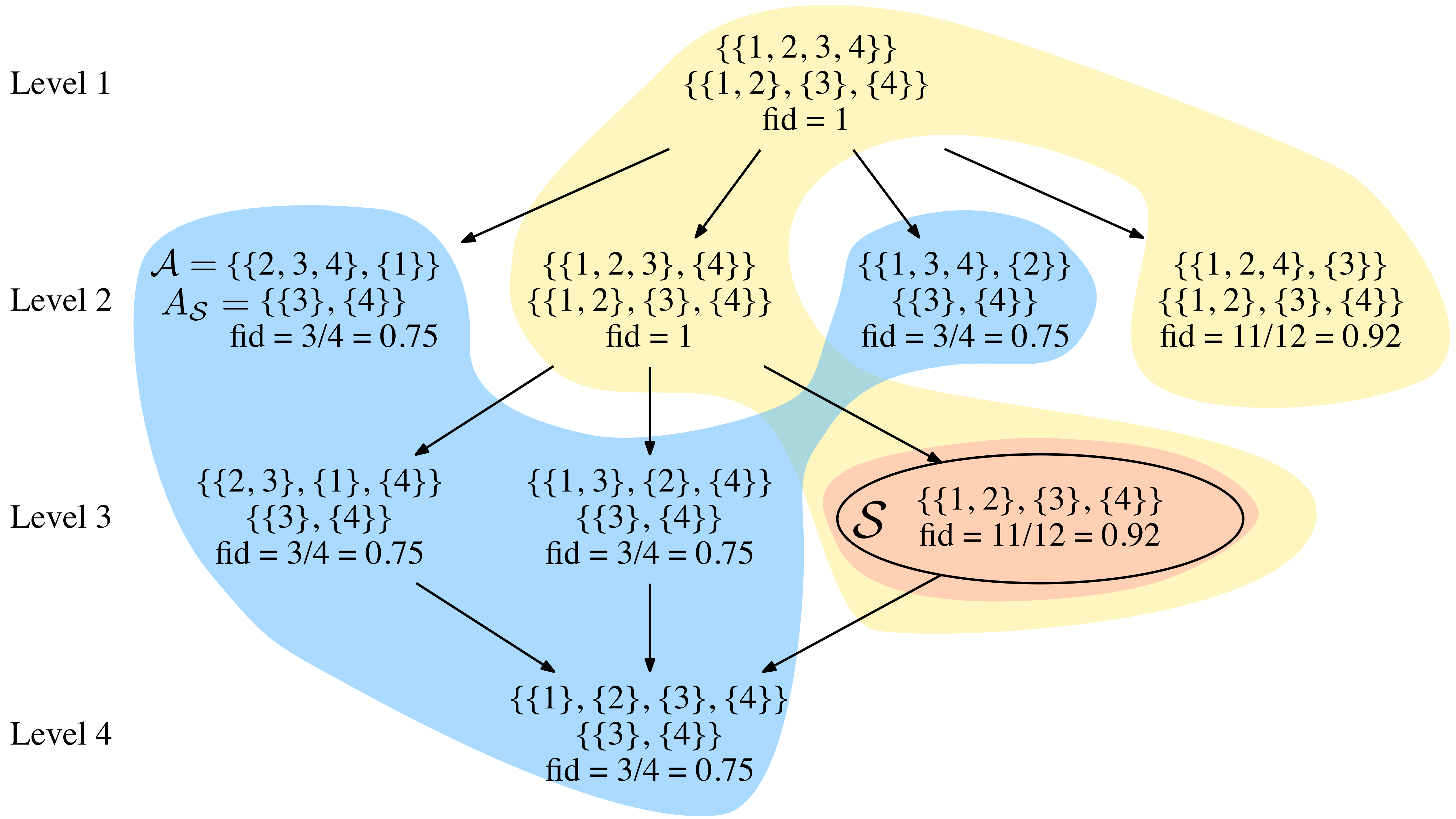



$$f(x) = I(A + B \geq 1/2)$$

- Independent within-class randomization impacts classifier performance more if attributes are correlated
- Interpretation: Naive Bayes uses correlations in data (also see Domingos and Pazzani 1997)




$$f(x) = I(A + B \geq 1/2)$$


Level 1
Level 2
Level 3
Level 4
{{1, 2, 3, 4}}
{{1, 2}, {3}, {4}}
fid = 1
$\mathcal{A} = \{\{2,3,4\},\{1\}\}$
$A_{\mathcal{S}} = \{\{3\},\{4\}\}$
fid = 3/4 = 0.75
{{1, 2, 3}, {4}}
{{1, 2}, {3}, {4}}
fid = 1
{{1, 3, 4}, {2}}
{{3}, {4}}
fid = 3/4 = 0.75
{{1, 2, 4}, {3}}
{{1, 2}, {3}, {4}}
fid = 11/12 = 0.92
{{2, 3}, {1}, {4}}
{{3}, {4}}
fid = 3/4 = 0.75
{{1, 3}, {2}, {4}}
{{3}, {4}}
fid = 3/4 = 0.75
$\mathcal{S}$
{{1, 2}, {3}, {4}}
fid = 11/12 = 0.92
{{1}, {2}, {3}, {4}}
{{3}, {4}}
fid = 3/4 = 0.75