Decision Trees with Numeric Tests

Universiteit Leiden

Industrial-strength algorithms

- For an algorithm to be useful in a wide range of real-world applications it must:
  - Permit numeric attributes
  - Allow missing values
  - Be robust in the presence of noise
- Basic schemes need to be extended to fulfill these requirements

C4.5 History

- ID3, CHAID – 1960s
- C4.5 innovations (Quinlan):
  - permit numeric attributes
  - deal sensibly with missing values
  - pruning to deal with for noisy data

- C4.5 - one of best-known and most widely-used learning algorithms
  - Last research version: C4.8, implemented in Weka as J4.8 (Java)
  - Commercial successor: C5.0 (available from Rulequest)

Numeric attributes

- Standard method: binary splits
  - E.g. temp < 45
- Unlike nominal attributes,
  every attribute has many possible split points
- Solution is straightforward extension:
  - Evaluate info gain (or other measure)
    for every possible split point of attribute
  - Choose “best” split point
  - Info gain for best split point is info gain for attribute
- Computationally more demanding

Example

- Split on temperature attribute:

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

  - E.g. temperature < 71.5: yes/4, no/2
    temperature ≥ 71.5: yes/5, no/3

  - Info([4,2],[5,3])
    = 6/14 info([4,2]) + 8/14 info([5,3])
    = 0.939 bits

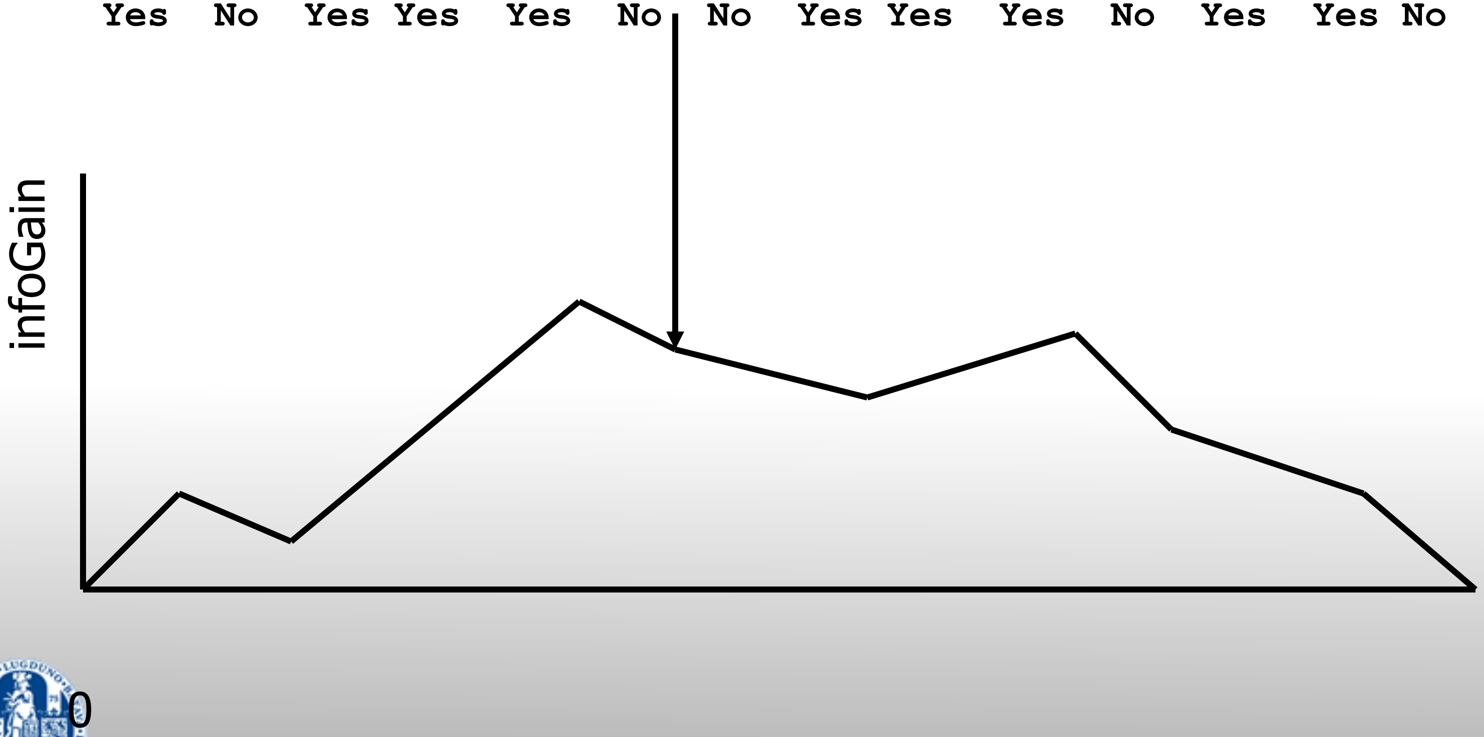
- Place split points halfway between values

- Can evaluate all split points in one pass!

Example

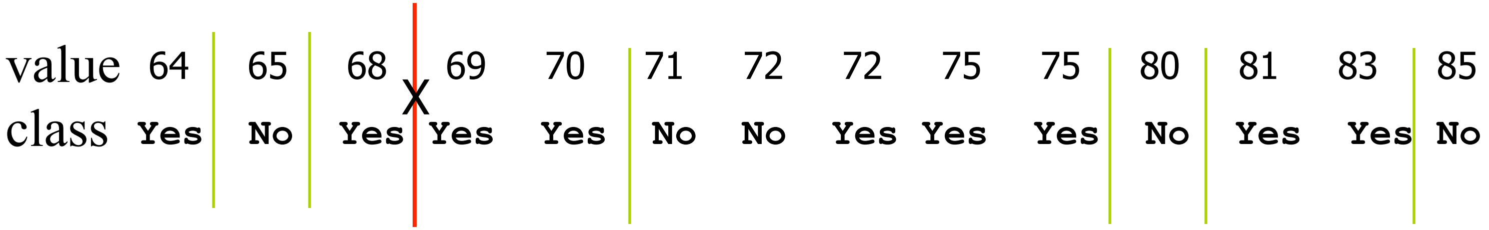
- Split on temperature attribute:

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No



Speeding up

- Entropy only needs to be evaluated between points of different classes (Fayyad & Irani, 1992)

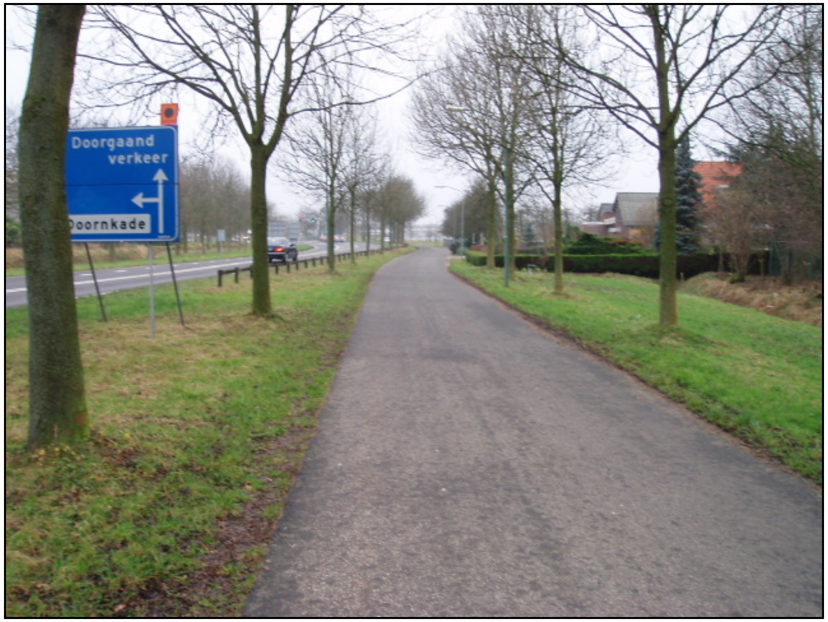


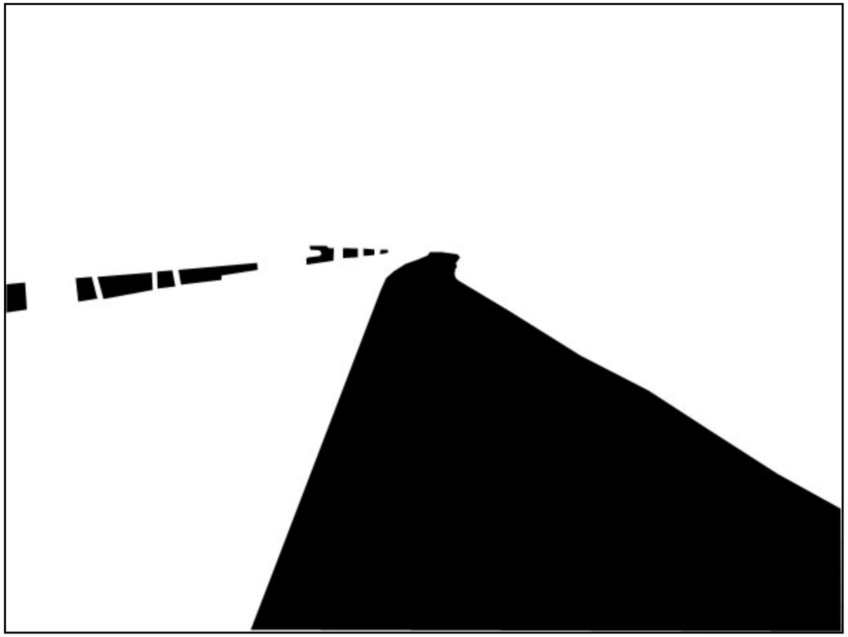
value	64	65	68	69	70	71	72	72	75	75	80	81	83	85
class	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

Potential optimal breakpoints

Breakpoints between values of the same class cannot be optimal

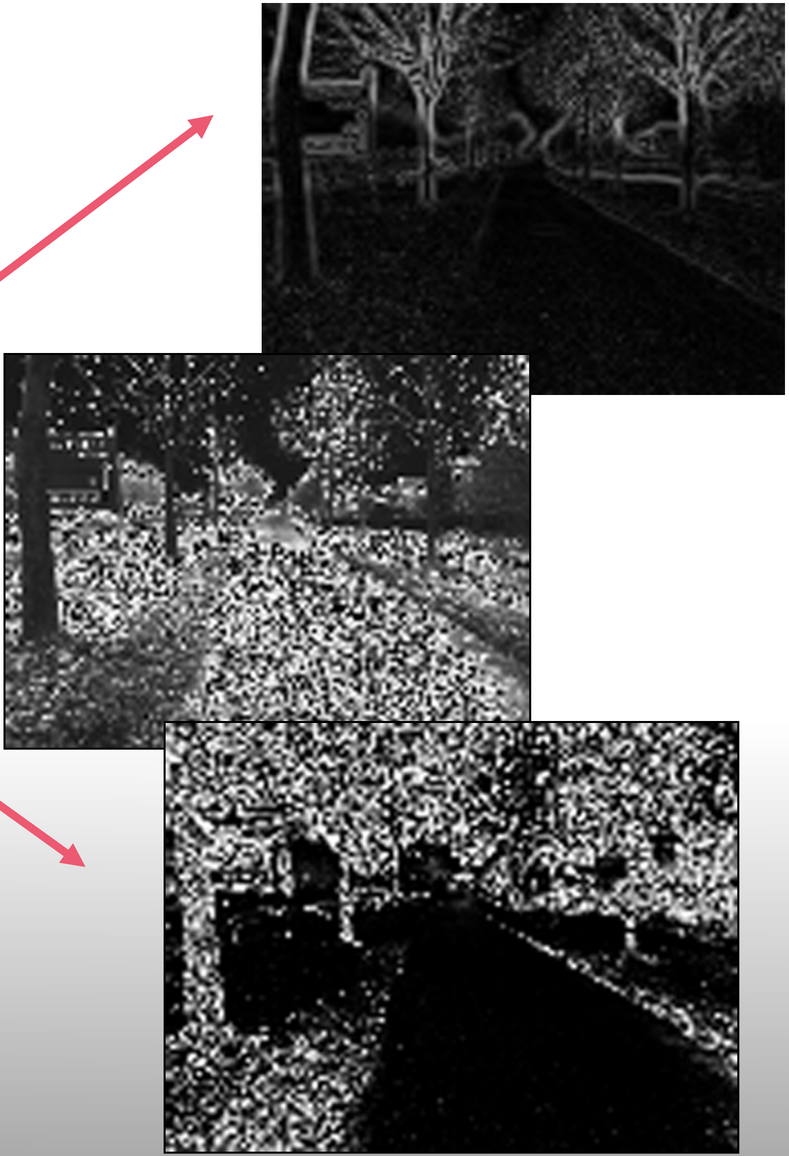
Application: Computer Vision 1





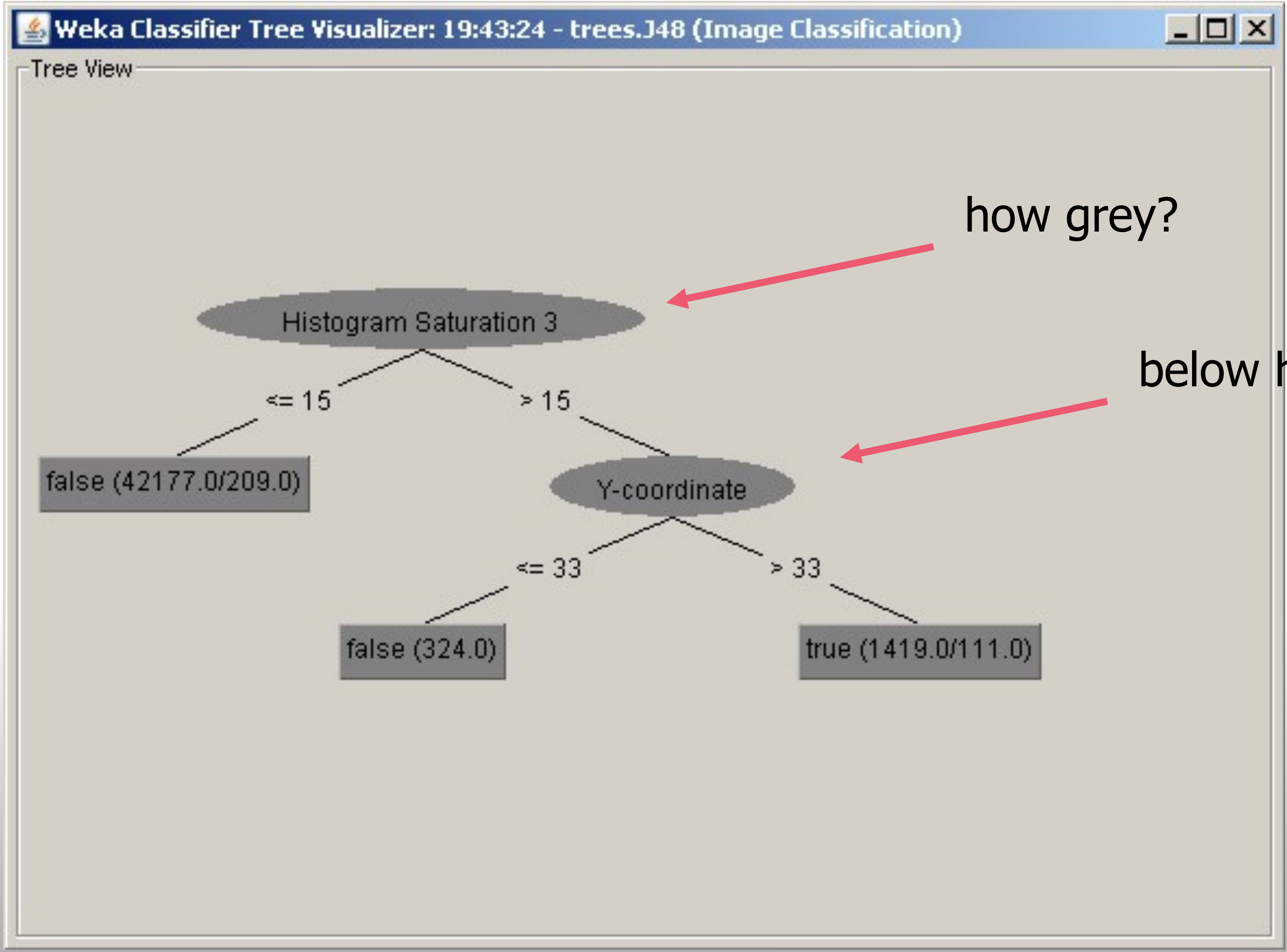
Application: Computer Vision 2

feature extraction



- color (RGB, hue, saturation)
- edge, orientation
- texture
- XY coordinates
- 3D information

Application: Computer Vision 3

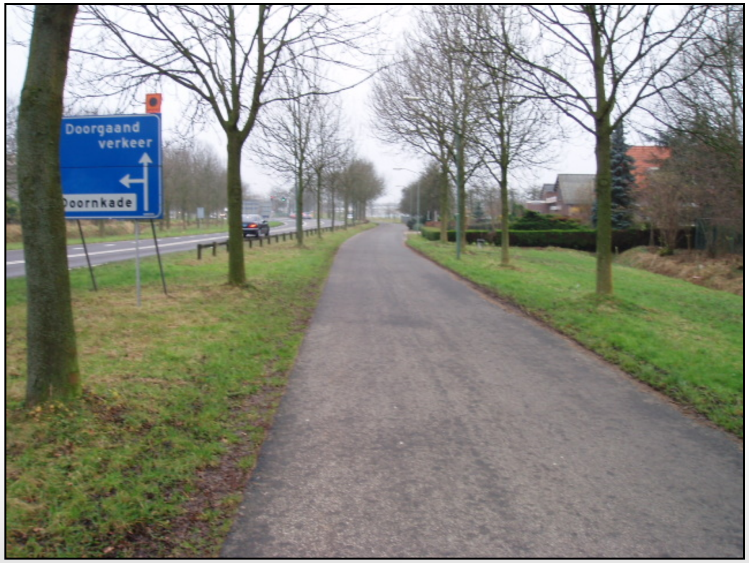


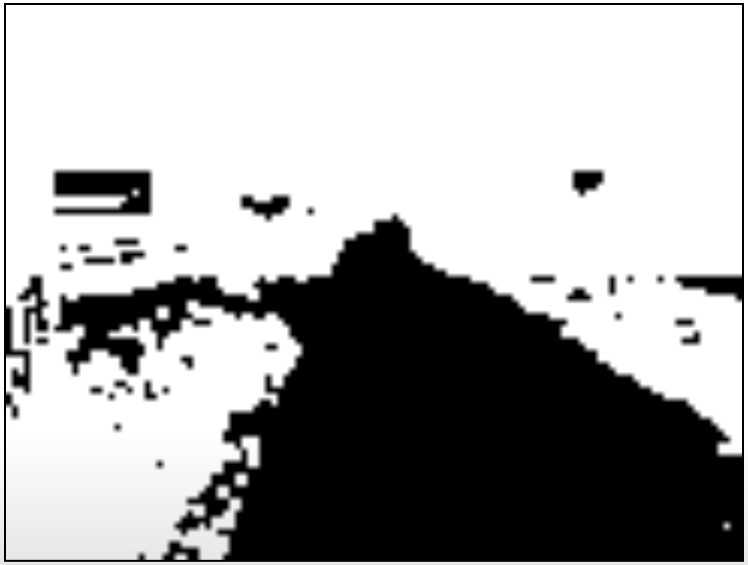
how grey?

below horizon?

Application: Computer Vision 4

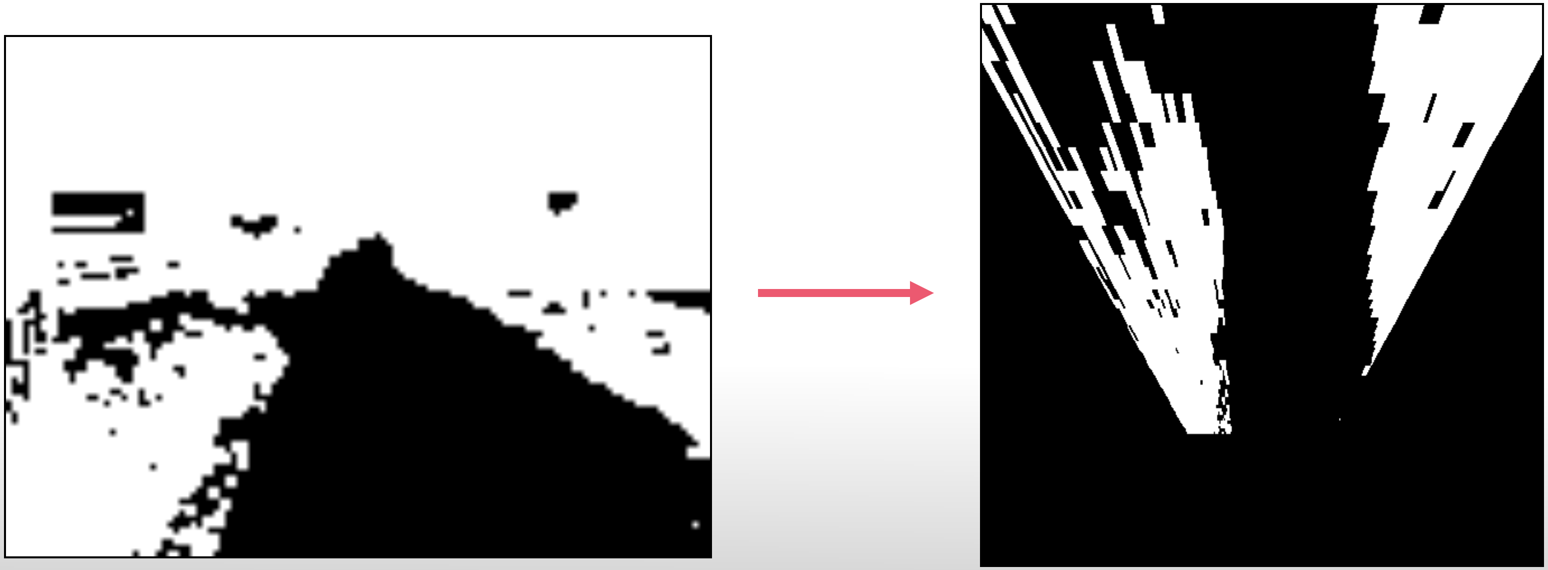
prediction





Application: Computer Vision 4

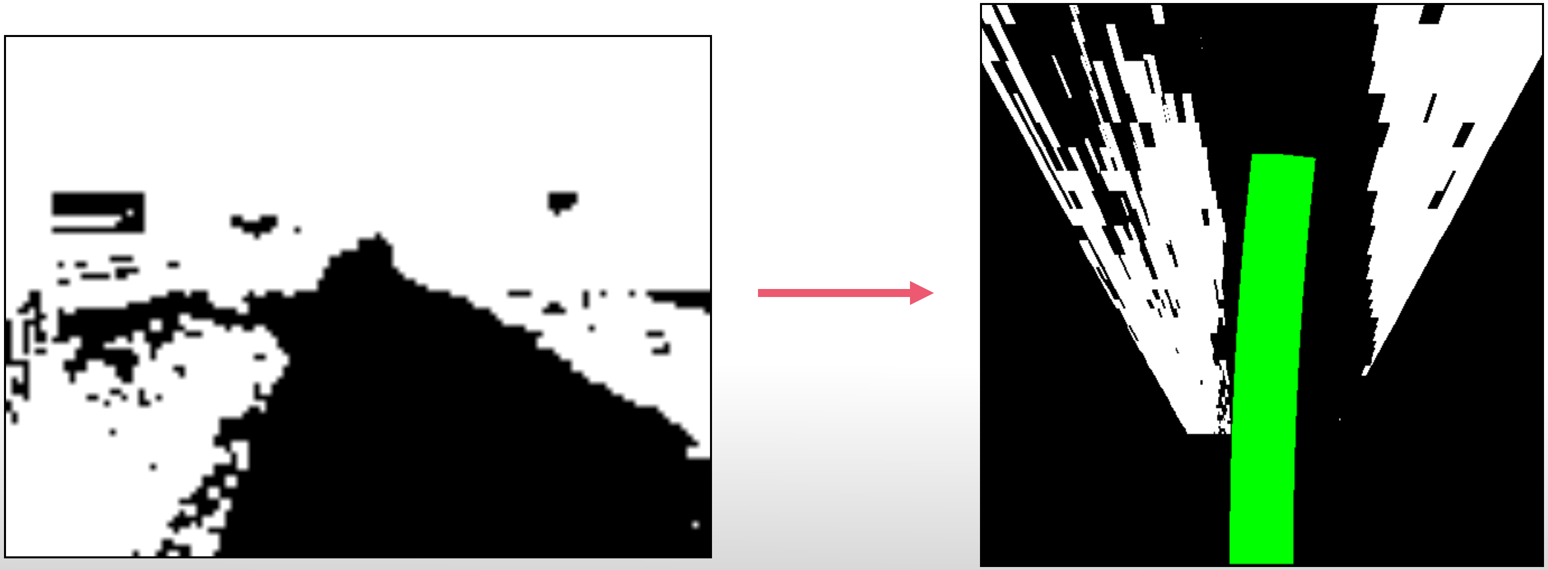
inverse perspective



Application: Computer Vision 5

inverse perspective

path planning

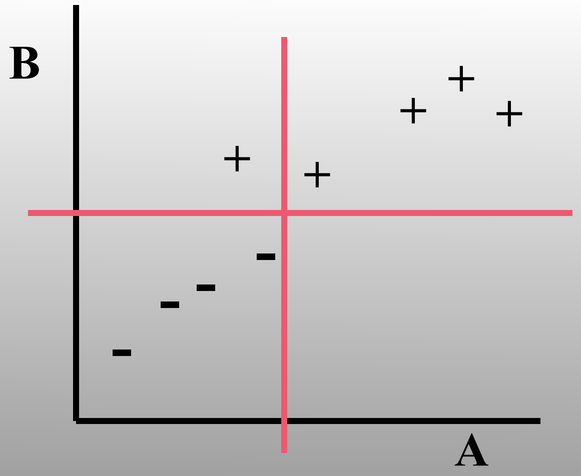


Quiz 1

Q: If an attribute A has high info gain, does it always appear in a decision tree?

A: No.

If it is highly correlated with another attribute B, and infoGain(B) > infoGain(A), then B will appear in the tree, and further splitting on A will not be useful.



Quiz 2

Q: Can an attribute appear more than once in a decision tree?

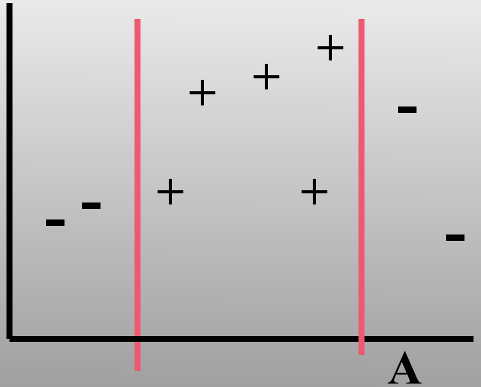
A: Yes.

If a test is not at the root of the tree, it can appear in different branches.

Q: And on a single path in the tree (from root to leaf)?

A: Yes.

Numeric attributes can appear more than once, but only with very different numeric conditions.



Quiz 3

Q: If an attribute A has infoGain(A)=0, can it ever appear in a decision tree?

A: Yes.

1. All attributes may have zero info gain.
2. info gain often changes when splitting on another attribute.

the XOR problem: