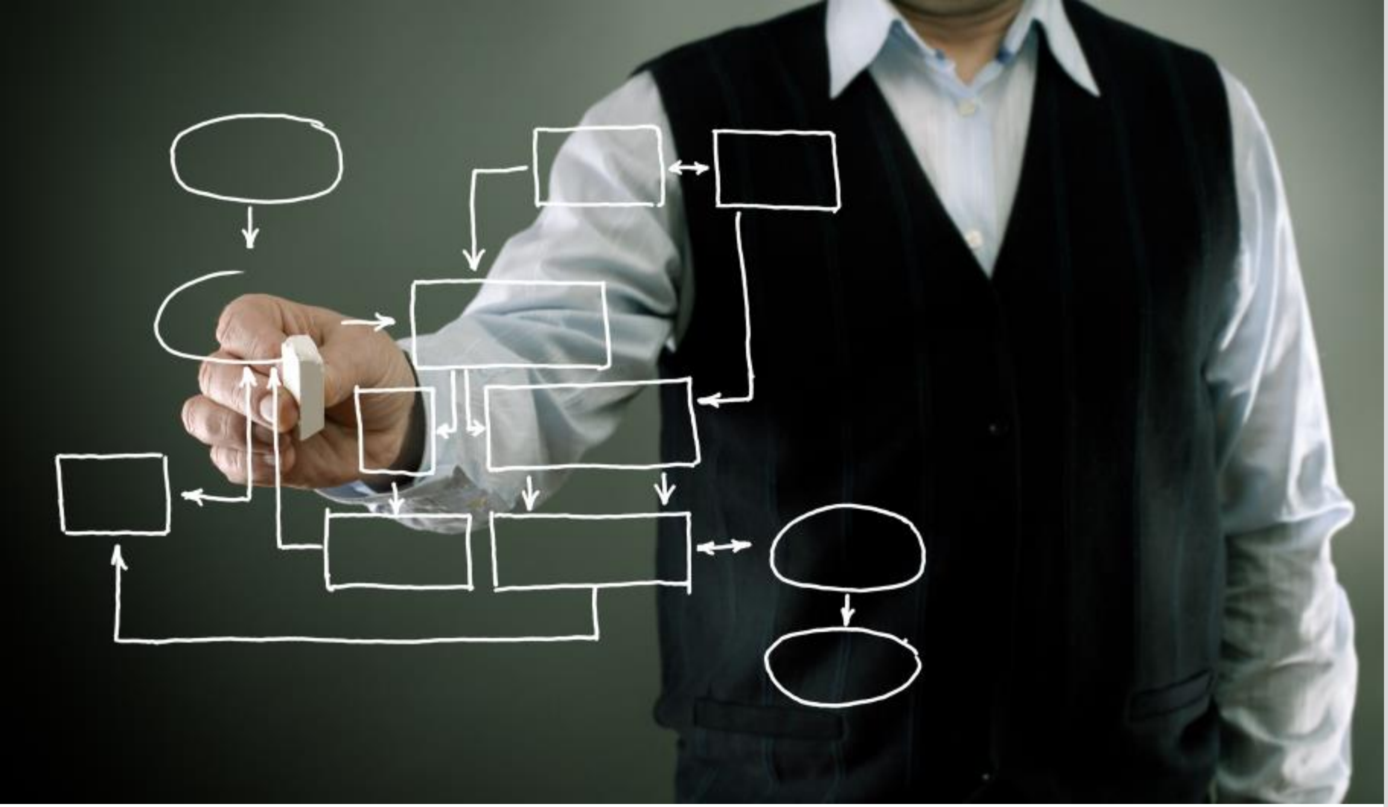


Instance-based learning

HEAL

HEURISTIC AND EVOLUTIONARY ALGORITHMS LABORATORY

Contact:
Dr. Michael Affenzeller
FH OOE - School of Informatics,
Communications and Media
Heuristic and Evolutionary
Algorithms Lab (HEAL)
Softwarepark 11, A-4232
Hagenberg

e-mail:
michael.affenzeller@fh-hagenberg.at
Web:
http://heal.heuristiclab.com
http://heureka.heuristiclab.com

Instance-based learning

Model-based approach

- Induce a Model h: X → Y from data D
- Use h for new queries

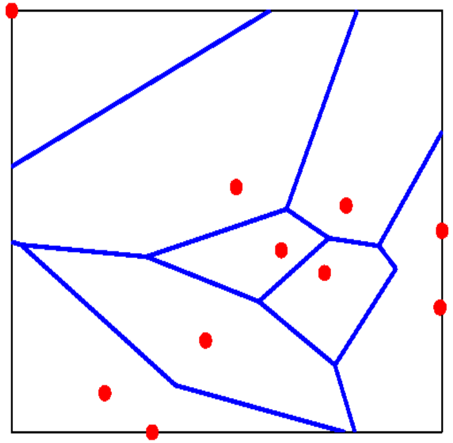
Instance-based approach

- Maintain a set of instances S
- Answer queries with instances of S
- Inference: Learning by remembering examples and 'extrapolate' them

Instance-based learning

Nearest Neighbor (NN) classification:

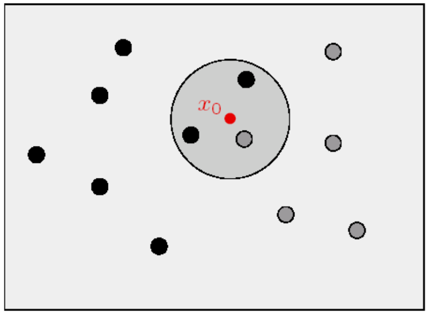
- Estimate the class of x_0 by the use the observation $x \in S$ most similar to x_0
- Assumption: similar objects have similar classes
- Requirement: reasonable similarity measure
- e.g. with Voronoi Diagram:



Instance-based learning

k-Nearest Neighbor (k-NN) classification:

- Learning: save all training examples
- Classification/Prediction of new cases
  - Find the k-nearest (or most similar) neighbors x_1,….x_k in the trainings instances of a new instance (x_0)
  - Prediction:
    - Discrete (classification):
      predict the class, which occurs most frequently in the k-nearest neighbors (voting)
    - Continuous (numerical prediction):
      return the arithmetic mean of the k-nearest neighbors as prediction
      - » Possible simplification: weighted sums (inverse of distance to x_0)



Instance-based learning

Advantages:

- Very simple Method
- No assumption about the model class necessary
- Training is very fast (lazy learning)
- Flexible (decision limits)
- Good statistic properties (only with idealistic assumptions)

Instance-based learning

Disadvantages:

- Slow at query time
- Danger of overfitting is high (with small k)
- “Curse of Dimensionality“:
  required number of observations increases exponentially with the dimension of X
- Easily fooled by irrelevant attributes (no feature selection)

Instance-based learning

Extension:

- Editing strategies (maintaining the set of instances)
- Method to adapt the similarity measure, locale measures
- Method to find the ‘optimal k‘
- Efficient search for the next neighbors

Goal of editing strategies:

- Save only few examples (efficiency)
- Save only examples, which can be classified ‘well’