Problem Solving by

Case-Based Reasoning

PART 1

Joint Lecture

„Artificial Intelligence“ and „Machine Learning“

Sommersemester 2010

11.05.2010

Dr. Thomas Gabel

Machine Learning Lab

Agenda

P A R T 1 *[AI&ML]*

1. Introduction to CBR
2. Knowledge and Case Representation
3. Similarity
4. Similarity-Based Retrieval

P A R T 2 *[only ML]*

5. Solution Adaptation
6. Learning in Case-Based Reasoning
7. Applications
8. References

1. INTRODUCTION

What is Case-Based Reasoning?

Case-Based Reasoning is …

- an approach to model the way humans think
- an approach to build intelligent systems

Central Ideas:

- store experiences made → as cases
- solving a new problem do the following
  - recall similar experiences (made in the past) from memory
  - reuse that experience in the context of the new situation (reuse it partially, completely or modified)
  - new experience obtained this way is stored to memory again

Classification of CBR (I)

- sub-discipline of Artificial Intelligence
- belongs to Machine Learning methods
  - learning process is based on analogy
    ➔ not on deduction or induction
  - best classified as supervised learning
    (recall the distinction between supervised, unsupervised and reinforcement learning methods typically made in Machine Learning)
  - learning happens in a memory-based manner
    - in a model-based approach training, data is used in order to create a model
    - in the domain considered, the model learnt can then be used, for example, for prediction or classification purposes
    - most „work“ (calculations) are done when building the model
    - by contrast: in a memory-based approach, *most* calculations are done at the time of application, i.e. when doing prediction or classification
      - „most“, but not „all“ → the necessary calculations at application time can be supported and prepared by creation of suitable data structures (e.g. kd-trees) at storing time
    - therefore, memory-based approaches are sometimes also called „lazy learning“

Classification of CBR (II)

- one of the few commercially/industrially really successful AI methods
  - customer support, help-desk systems: diagnosis and therapy of customer‘s problems, medical diagnosis
  - product recommendation and configuration: e-commerce
  - textual CBR: text classification, judicial applications (in particular in the countries where common law (not civil law) is applied) [like USA, UK, India, Australia, many others]
- applicability also in ill-structured and bad understood application domains

Case-Based Reasoning and Cases

- „Case-Based Reasoning is […] reasoning by remembering.“ Leake, 1996
- „A case-based reasoner solves new problems by adapting solutions that were used to solve old problems.“ Riesbeck & Schank, 1989
- „Case-Based Reasoning is a recent approach to problem solving and learning […]“ Aamodt & Plaza, 1994
- „Case-Based Reasoning is both […] the ways people use cases to solve problems and the ways we can make machines use them.“ Kolodner, 1993

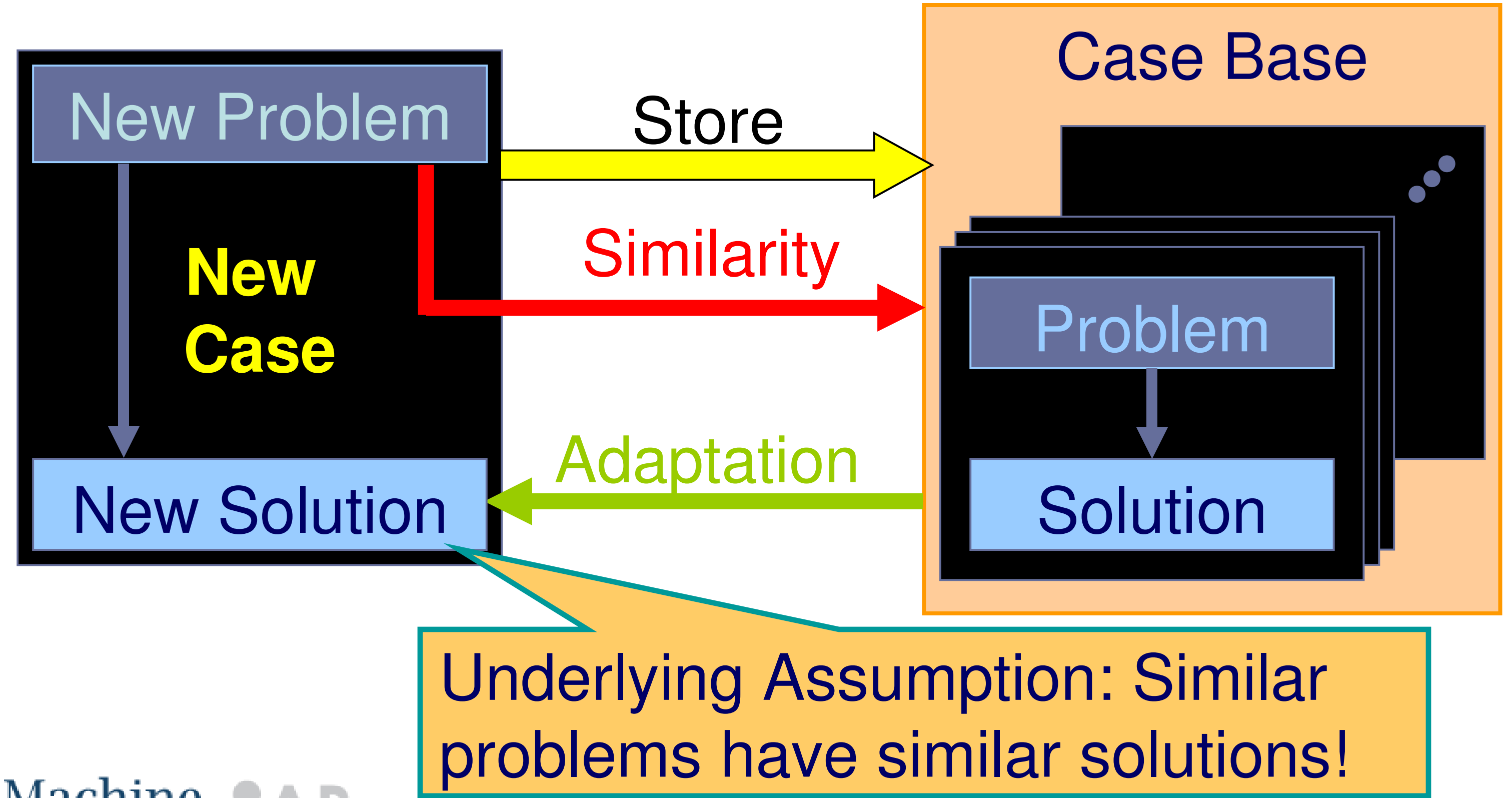
What is a Case?

a) Cognitive Science View:
Cases are abstractions of events, that are limited in space and time; they represent episodic knowledge.

b) Technical View:
A case is a desciption of a problem situation (that actually occurred) together with certain experiences that could be obtained during processing and solving the problem.

Simplified CBR Model

Solve new problems by selecting cases used for similar problems and by (eventually) adapting the belonging solution.



When is CBR of Relevance?

- When a domain theory does not exist, but example cases are easy to find.
- When an expert in the domain is not available, is too expensive, or is incapable of articulating verbally his performance, but example cases are easy to find.
- When it is difficult to specify domain rules, but example cases are easy to find.
- When cases with similar solutions have similar problem descriptions.
  - i.e. there exists a similarity metric for problem descriptions and a corresponding set of adaptation rules
- When a case base already exists.

Contents of a Case

Mandatory

- problem part
- solution part

Optionally

- context (e.g. justifications)
- pointer to other relevant cases
- solution quality assessment
- steps of the solution

➔ The main difficulty arises, when the actual situation is not identical to the previous one. Then, inexactness is involved.

➔ A main feature of CBR techniques is that they allow inexact / approximate reasoning in a controlled manner.

A Simple Example Scenario: Call Center (I)

- Task: Technical Diagnosis of Car Faults
  - symptoms are observed (e.g. engine does not start) and values are measured (e.g. battery voltage = 6.3V)
  - goal: find the cause for the failure (e.g. battery empty) and a repair strategy (e.g. change battery)
- Case-Based Diagnosis
  - a case describes a diagnostic situation and contains
    - a description of the symptoms
    - description of the failure and the cause
    - description of a repair strategy
  - store a collection of cases in the case base
  - find a case similar to the current problem and reuse the repair strategy

A Simple Example Scenario: An Example Case (II)

- A case describes a particular diagnostic situation.
- A case records several features and their specific values occurred in that situation.

➔ A case is not a general rule.

Attributes (Features)

Values

Case 1

Problem (Symptoms):

- Problem : front light does not work
- Car : VW Golf IV, 1.6l
- Year : 1998
- Battery Voltage : 13.6V
- State of Lights : ok
- State of Light Switch : ok

Solution:

- Diagnosis : front light fuse defect
- Repair : repair front light fuse

String

Symbolic Data Type {Polo, Golf,...}

Numeric Data Type $[0;50]\subset R$

A Simple Example Scenario: Solving a New Diagnostic Problem (III)

Case Base with Two Cases

➔ each case describes one particular situation

➔ all cases are independent of one another

Case 1

Problem (Symptoms):

- Problem : front light does not work
- Car : VW Golf IV, 1.6l
- Year : 1998
- Battery Voltage : 13.6V
- State of Light : ok
- State of Light Switch : ok

Solution:

- Diagnosis : front light fuse defect
- Repair : repair front light fuse

Case 2

Problem (Symptoms):

- Problem : front light does not work
- Car : Audi A4
- Year : 2002
- Battery Voltage : 12.9V
- State of Light : surface damaged
- State of Light Switch : ok

Solution:

- Diagnosis : bulb defect
- Repair : replace front light

A New Problem (Query) Has to Be Solved

➔ we make several observations in the current situation

➔ observations define a new problem

➔ not all attribute values have to be known

➔ Note: The new problem is a ``case" without solution part.

Problem (Symptoms):

- Problem : break light does not work
- Car : Audi 80
- Year : 1990
- Battery Voltage : 12.6V
- State of Lights : ok
- State of Light Switch :

Compare the new problem with each case and select the most similar one!

➔ **CASE 1**

Questions:

When are two cases similar?

How to rank the cases according to their similarity?

How to reuse the solution of the corresponding case?

Note:

Similarity is the most important concept in CBR.

Similarity may be assessed based on the similarity of each feature, while the importance of different features may vary (feature weighting).

A Simple Example Scenario: Reuse and Retain (IV)

- Reuse
  - adapt the solution
  - how do differences in the problem affect the solution

Case 1

Problem (Symptoms):
- Problem : **front light** does not work
- …

Solution:
- Diagnosis : **front light** fuse defect
- Repair : repair **front light** fuse

Problem (Symptoms):
- Problem : **break light** does not work
- Car : Audi 80
- Year : 1990
- Battery Voltage : 12.6V
- State of Lights : ok
- State of Light Switch :

New Solution:
- Diagnosis : **break light** fuse defect
- Repair : repair **break light** fuse

- Retain
  - if diagnosis is correct: store new case
  - add case to case base

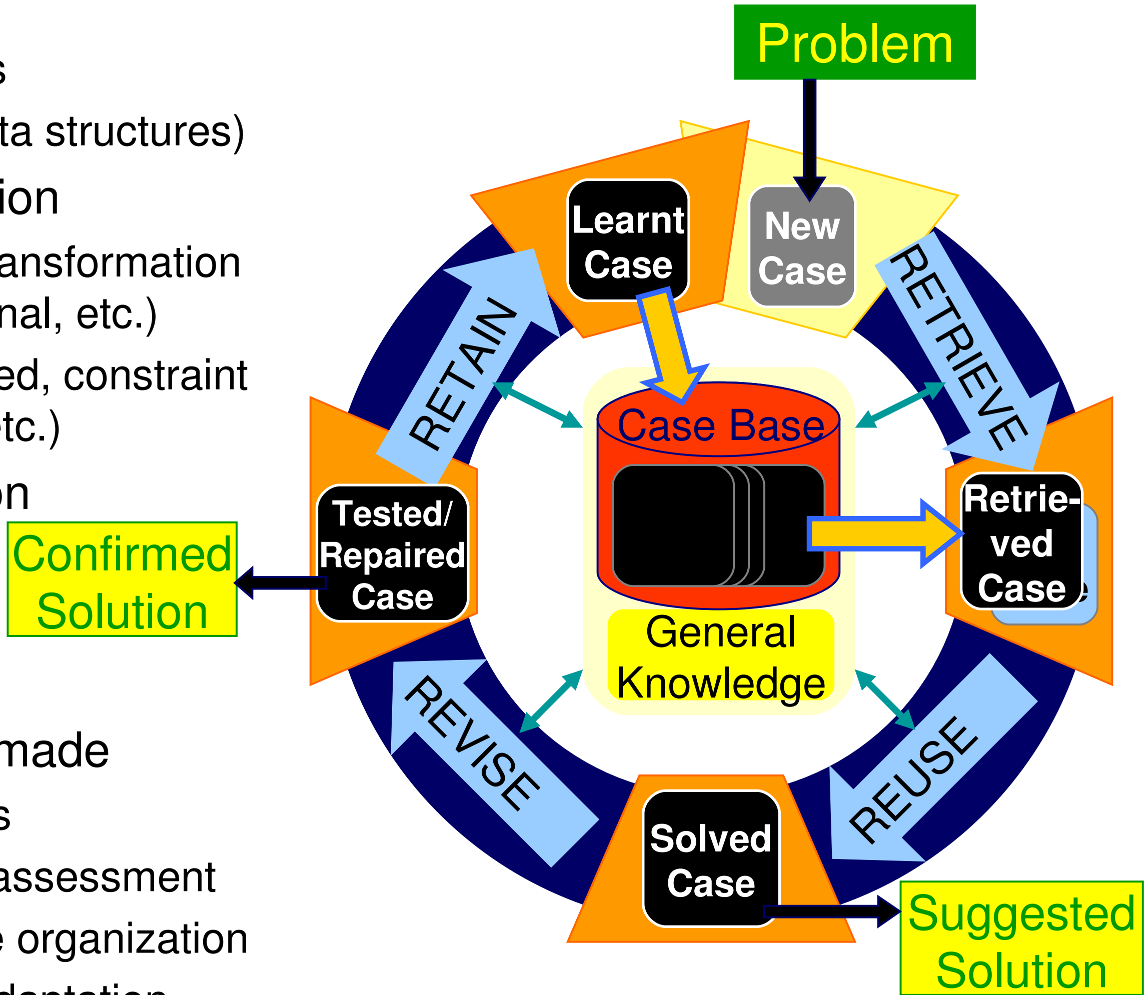
Case 3

Problem (Symptoms):
- Problem : break light does not work
- Car : Audi A80
- Year : 1990
- Battery Voltage : 12.6V
- State of Light : ok
- State of Light Switch :

Solution:
- Diagnosis : break light fuse defect
- Repair : replace break light fuse

CBR Cycle (R4, [Aamodt&Plaza, 1994])

- retrieve: find most similar case(s)
  - similarity measures
  - explanation-based methods
  - case-base organisation (data structures)
- reuse: transform/adapt solution
  - different types of solution transformation (none, interactive, derivational, etc.)
  - different methods (rule-based, constraint satisfaction, model-based etc.)
- revise: verify/improve solution
  - no verification
  - verification by simulation
  - verification in the real world
- retain: keep the experience made
  - learn new cases
  - learn similarity assessment
  - learn case base organization
  - learn solution adaptation



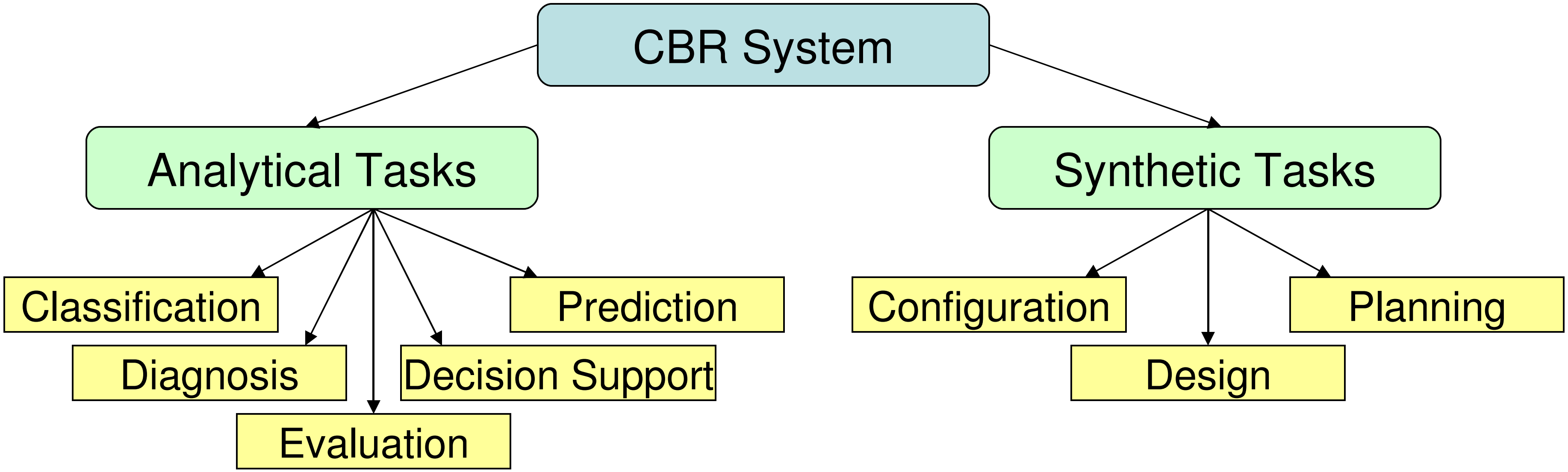
Advantages of CBR (I)

- Avoidance of High Knowledge Acquisition Effort
  - case knowledge is usually easily acquirable
  - not much general knowledge required
- Simpler Maintenance of the Knowledge in the System
  - maintenance by adding/removing cases from the case base
  - cases are independent of one another and easily interpretable (even for non-experts)
- Facilitation of Intelligent Retrieval (compared to data-base systems)
  - DBMS often give too few/many results

Advantages of CBR (II)

- High Quality of Solutions for Poorly Understood Domains
  - case-based systems can be made to retain only ``good" experience in memory
  - if only little adaptation is necessary for reuse, this will not impair the solution's quality too much
- High User Acceptance
  - provided solution corresponds to actual experience
    ➔ may increase trust in the solution
  - selected case and solution adaptation can be explicitly presented to the user
  - problems of rule-based / neural network-based systems
    - black boxes
    - inference process is not traceable or hidden
    - provided solutions are difficult to explain

Typical Application Fields (I)



- Remarks concerning synthetic tasks:
  - main focus is on composing a complex solution from simpler components ➔ focus is often on solution adaptation
  - configuration: e-commerce scenario → product configuration (e.g. personal computers)
  - design: reuse of construction plans in civil engineering
  - planning: production planning

Typical Application Fields (II)

- Remarks concerning analytical tasks:
  - main focus is on analysing a given situation
  - classification (assign objects to a class $K_i \in \{K_1,\ldots,K_n\}$)
    → e.g. recognition of sponges
  - diagnosis (classification + verification + therapy)
    → e.g. fault diagnosis in Airbus engines
  - evaluation/regression (like classification, but assignment of real-valued assessments):
    → e.g. credit risk assessment
  - decision support (search for specific information relevant for decision-making)
    → e.g. web-based product catalogues, like online travel agencies
  - prediction (like classification + time dependency)
    → e.g. prediction of the probability of failure of a machine's part

CBR for Classification (I)

- A classifier for a set M is a mapping f:M→I (where I is a finite index set).
  ➔ A case-based classifier is given by a case base, a similarity measure and the principle of the nearest neighbour.
- **Definition:** Given a case base CB, a similarity measure sim and an object (problem) q∈M, we call c=(p,s)∈CB the *Nearest Neighbour* to q, if: for all (p‘,s‘)∈CB it holds sim(q,p)≥ sim(q,p‘).
- **Definition:** In *Nearest-Neighbour Classification* each new object (query) q∈M is assigned the class s∈I of q‘s nearest neighbour in CB, i.e. when

$$NN = (p_{NN}, s_{NN}) = \arg\max_{c \in CB} sim(q,c)$$

then q is assumed to belong to class s_{NN}.

CBR for Classification (II)

- Note: The classifier defined by the pair (CB,sim) is not unique, if there is more than one nearest neighbour.
- Extension to k-NN Classification:
  - The k most similar neighbours of q are considered. Typically, a majority voting is applied to determine the class of the query q.
  - Formally:
    Let $NN_k(q)=\{((p_1,s_1),(p_2,s_2),\ldots,(p_k,s_k))\}$ denote the set of k nearest neighbors of q. If we denote by $$n_i = \sum_{l=1}^{k} I(s_l = i)$$
    the frequency of class label i within the k nearest neighbor, then q is assumed to belong to class $s = \arg\max_{i \in I} n_i$

k-Nearest Neighbor Classification

- Demonstration video on the k-NN classifier

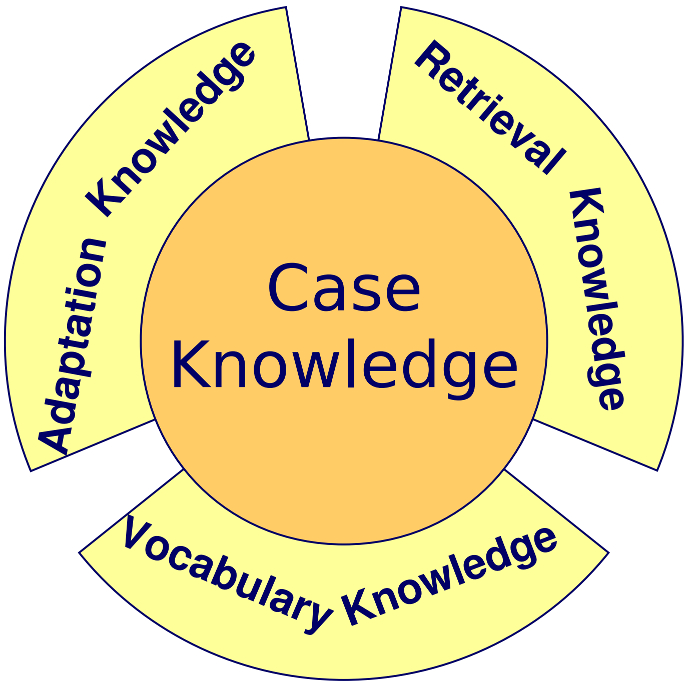
(c) Antal van den Bosch,
Tilburg University

2. KNOWLEDGE AND CASE REPRESENTATION

What forms of knowledge are parts of a CBR system?
How can cases be represented?

Knowledge Container Model [Richter, 1989]

- „In order to solve problems, one needs knowledge.“
- Knowledge of a CBR System
  - vocabulary: knowledge representation
  - retrieval: similarity assessment (measures)
  - solution transformation: rules
  - cases
- Knowledge Management
  - as the environment may change, maintenance of the containers‘ contents over the lifetime of the CBR system is crucial to guarantee its continued usability



Case Contents

Problem / Situation Information

➔ must cover all the information that is necessary to decide if this case is applicable for a new situation

- target of the problem
- constraints
- characteristics

➔ new situation = query

Solution

➔ contains all the information that describes a solution to the problem sufficiently accurately

- ➔ solution itself
- ➔ justifications
- ➔ possible alternative solutions
- ➔ steps that were tried, but failed

Solution Evaluation

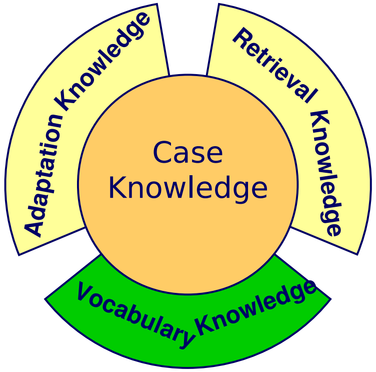
➔ feedback from the real world

- ➔ How good was the solution for the problem?

Case Representation Formalisms (I)

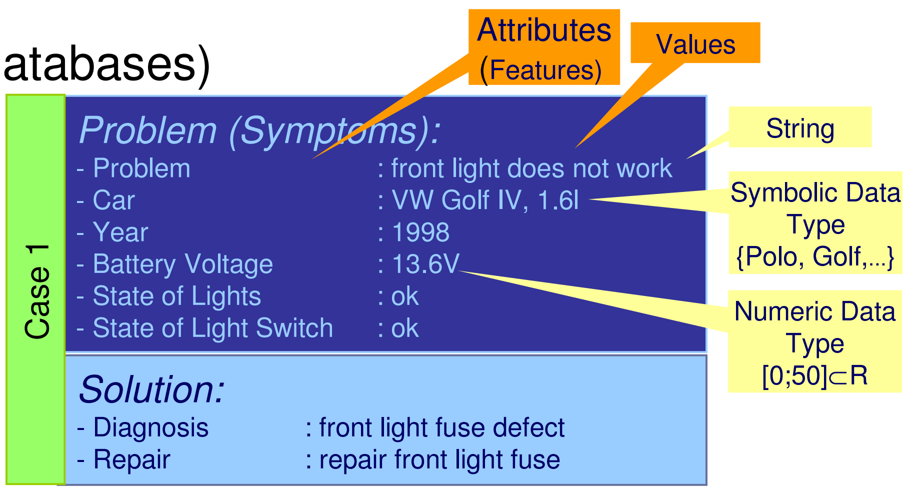
Attribute-Value Based Case Representation

- Case (problem and solution) is represented by pairs of attributes and belonging values.
  - e.g.: price = 9.95€
- Set of attributes $\{A_1,\ldots,A_n\}$ is (in general) fixed for all cases.
- To each attribute A_i there is an associated domain D_i and for each attribute's value it holds $a_i \in D_i$, e.g.
  - numerical attributes (Integer or Real or subsets of those)
  - symbolic attributes (finite domains, $D_i=\{d_1,\ldots,d_m\}$)
  - textual attributes (strings)
- Note:
  - Choice of attributes and corresponding domains to represent cases represents general knowledge: vocabulary knowledge.
  - Choice of domains is mainly influenced by the requirements for similarity computation and solution adaptation.



Case Representation Formalisms (II)

- Choice of attributes
  - must allow for the decision whether a case and a new situation are similar
  - should avoid redundancies
  - should represent independent properties of a case
- Disadvantages
  - no structural or relational information is representable
  - no ordering information (e.g. sequence of actions) is representable
- Advantages
  - straightforward representation
  - easy to understand and implement
  - cases are easy to store (usage of databases)
  - efficient retrieval
- Example
  - recall the example from the Introduction



Case Representation Formalisms (III)

Object-Oriented Case Representation

- Refinement and more structured extension of attribute-value based representation
- Compositing of related attributes to object descriptions; each object is described by a fixed set of attributes ➔ Case = Set of Objects

Graph- and Tree-Based Representation

- e.g. suited for atomic/molecule structures or electrical circuit designs

First-Order-Based Case Representation

- problems and solutions are represented as sets of Grundatome (variable-free)

Hierarchical Case Representation

- each case is represented on several levels of abstraction

Generalised Cases

- each case describes sets of cases at once, which are highly similar to one another
  ➔ smaller case bases, simplified case/solution adaptation

3. SIMILARITY

When is a new problem (query) similar to a case's problem part?
What forms of similarity measures are suitable?

Meaning of Similarity

- Similarity is the central notion in Case-Based Reasoning.
- Similarity is always considered between problems (not solutions of cases).
- Selection of cases during the ``Retrieve“ phase is based on the similarity of cases to a given query.

→ Observation I: There is no universal similarity; similarity always relates to a certain purpose.
  - e.g. two cars can be similar if they have the same max speed or cost approximately the same → different aspects of similarity

→ Observation II: Similarity is not necessarily transitive.
  - e.g. 10€ are similar to 12€, 12€ are similar to 14€ … 100€ are similar to 102€. *But:* 10€ are not similar to 102€ → property of ``small numeric difference“ is intransitive

→ Observation III: Similarity does not have to be symmetric.

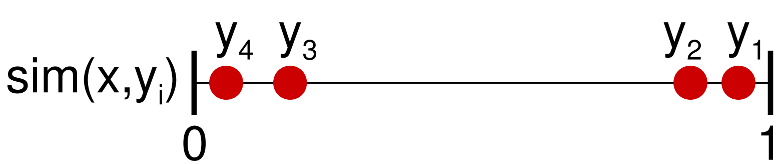
Machine Learning LAB

Similarity and Utility

- Purpose of Similarity: Selection of solutions that can be easily transferred / adapted to the problem at hand.
- Similarity = Utility for Solving a (new) Problem
- *Note:*
  - Utility is an a-posteriori criterion: In general, the utilitiy (of a case) can be estimated after having solved the problem.
  - Similarity concerning problem situations is an a-priori criterion: Similarity must be estimated before solving the problem.
- **Goal:** Similarity must approximate utility as accurately as possible.

Similarity Measures

- *Idee:* Numerical modelling of similarity, capturing the degree of similarity
- **Definition:** A ***Similarity Measure*** on a set M is a real-valued function sim: $M^2 \rightarrow [0,1]$.
  We say that sim is
  - reflexive iff. $\forall x \in M: sim(x,x) = 1$
  - symmetric iff. $\forall x,y \in M: sim(x,y) = sim(y,x)$
- Beyond ordinal information, similarity measures allow for a quantitative statement on the degree of similarity.
- **Definition:** Each similarity measure induces a ***similarity relation*** R_{sim} as
  $R_{sim}(x,y,u,v)$ iff. $sim(x,y) \geq sim(u,v)$
  $y \geq_z x$ iff. $sim(z,y) \geq sim(z,x)$



Distance Measures

- **Definition:** A ***Distance Measure*** on a set M is a real-valued function d: $M^2 \rightarrow \Re^+_0$.
  We say that sim is
  - reflexive iff. $\forall x \in M$: $d(x,x) = 0$
  - symmetric iff. $\forall x,y \in M$: $d(x,y) = d(y,x)$
- **Definition:** A distance measure d on a set M is a ***Metric*** and (M,d) a ***Metric Space*** if
  $\forall x,y \in M$: $d(x,y) = 0 \Rightarrow x=y$
  $\forall x,y,z \in M$: $d(x,y) + d(y,z) \geq d(x,z)$
- **Definition:** Each distance measure induces a ***similarity relation*** R_d as
  $\forall x,y,u,v \in M$: $R_d(x,y,u,v)$ iff. $d(x,y) \leq d(u,v)$
  $\forall x,y,z \in M$: $y \geq_z x$ iff. $d(z,x) \leq d(z,y)$

Machine Learning LAB

Relation Between Distance and Similarity Measures

- **Definition:** A similarity measure sim and a distance measure d are called ***Compatible*** if and only if
  $$\forall x,y,u,v \in M: R_{sim}(x,y,u,v) \leftrightarrow R_d(x,y,u,v)$$
- ***Lemma*** *(Measure Transformation):* If there is a bijective, order-reversing mapping $f: \Re_0^+ \rightarrow [0,1]$ with
  $$f(0) = 1$$
  $$f(\ d(x,y)\ ) = sim(x,y)$$
  then sim and d are compatible.
- *Note:* A transformation function f can be employed to construct a compatible pendant for a given sim or d, respectively.
- *Examples:*
  - $f(x) = 1 - x/(x+1)$
  - $f(x) = 1 - x/x_{max}$

Examplary Similarity Measures (I)

(for attribute-value based case representations)

Hamming Distance

$$H(x,y) = n - \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} (1-x_i)(1-y_i)$$

- for binary features
- $x=(x_1,\ldots,x_n),\ x_i \in \{0,1\}$
- $H(x,y) \in \{0,\ldots,n\}$
- H(x,y) is the number of attributes with differing values
- H is a distance measure: $H(x,x)=0$, $H(x,y)=H(y,x)$
- $H((x_1,\ldots,x_n), (y_1,\ldots,y_n)) = H((1-x_1,\ldots,1-x_n), (1-y_1,\ldots,1-y_n))$

SMC (Simple Matching Coefficient)

$$SMC(x,y) = 1 - \frac{n-(a+d)}{n} = \frac{a+d}{n} = 1 - \frac{b+c}{n}$$

- $a=\sum x_i y_i$, $b=\sum x_i(1-y_i)$, $c = \sum (1-x_i)y_i$, $d = \sum (1-x_i)(1-y_i)$
  ➔ $n=a+b+c+d$ ➔ $H(x,y) = b+c = n-(a+d)$
- transformation of the Hamming distance into a compatible similarity measure by $f(d)=1-d/d_{max}$ yields the simple matching coefficient

Examplary Similarity Measures (II)

(for attribute-value based case representations)

SMC (Simple Matching Coefficient, ctd.)

- is not restricted to binary features (cf. previous slide)
- usable for
  - nominal discrete variables without natural ordering (e.g. colours)
  - ordinal discrete variables with natural ordering (e.g. school grades)

$$SMC(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = y_j)$$

Machine Learning LAB

Examplary Similarity Measures (II)

(for attribute-value based case representations)

Note: Similarity measures for other case representations (e.g. object-oriented, graph-based, etc.) are not considered in this lecture; see literature references.

Measures for Real-Valued Attributes

- $x_i, y_i \in \Re$ for all i
- generalisations of the Hamming distance
  - city-block metric d_1

$$d_1(x,y) = \sum_{i=1}^{n} | x_i - y_i |$$

  - Euclidean distance d_2

$$d_2(x,y) = \| x - y \| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  - weighted Euclidean distance d_{2w}
    - counteract different spreads of different variables
    - weights w_i must be specified

$$d_{2W}(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$$

  - p-norm d_p

Examplary Similarity Measures (III)

(for attribute-value based case representations)

Measures for Sparsely Filled Cases

- In some domains, the value „0“ is dominating which should be taken into consideration by a distance / similarity measure.
- Example:
  - A case describes a customer. Each attribute describes how many times the customer has bought a specific product. There are 1000 different products, hence a case comprises 1000 attributes.
  - Customer A and B have bought one product each, but different ones.
    ➔ Their Euclidean distance is $\sqrt{2}$.
  - Customer C and D have bought 100 different products each, 95 of them are identical.
    ➔ Their Euclidean distance is $\sqrt{10}$.
  - Thus, A and B are more similar than C and D.
    ➔ This is counterintuitive.

Examplary Similarity Measures (IV)

(for attribute-value based case representations)

Measures for Sparsely Filled Cases

- If zeros are dominating, a generalization of the SMC is applied.
- Let n denote the number of attributes and f the number of attributes, in which both cases are equally zero. Then, we define

$$SMC_{00}(x,y) = \frac{\left(\sum_{i=1}^{n} I(x_i = y_j)\right) - f}{n-f}$$

- In the example from the previous slide:
  - $SMC_{00}(A,B) = (998-998) / (1000-998) = 0$
  - $SMC_{00}(C,D) = (990-895) / (1000-895) = 95/105 = 0.905$

Examplary Similarity Measures (IV)

(for attribute-value based case representations)

Measures for Sparsely Filled Cases

- It may be desired to consider two customers similar, even if they have bought the same products differently often.
- **Definition:** The Cosine Similarity Measure is based upon the inner product and is defined as

$$\cos(x, y) = \frac{x^T y}{\| x \| \cdot \| y \|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- In the example we get cos(A,B)=0 and cos(C,D)=95/100=0.95.
- Pearson Correlation: Like cosine measure, but with averages of x and y subtracted, i.e. pearson(x,y)=cos(x-mean(x),y-mean(y)).

Note: Up to now, a single, global similarity measure was used. No differentiation with respect to the individual attributes was made.

Local-Global Principle

- case description by n attributes $A_1,\ldots,A_n$
- each attribute has a certain type T_i (e.g. numeric)
- **Local Similarity**
  - a seperate similarity function is used for each attribute:

    $$sim_{A_i}: T_i \times T_i \rightarrow [0,1]$$

  - local measures are depending on the respective type T_i of the attribute A_i
- **Global Similarity**
  - $sim(x,y) = sim((x_1,\ldots,x_n),(y_1,\ldots,y_n))$
    $= F(sim_{A_1}(x_1, y_1),\ldots, sim_{A_n}(x_n, y_n))$
  - $F:[0,1]^n \rightarrow [0,1]$ ➔ **Amalgamation Function**
  - requirements on F:
    - F is monotonous in each of its arguments
    - $F(0,\ldots,0) = 0$ and $F(1,\ldots,1) = 1$

Very frequently used in practice!

Examples:

- Weighted Average

$$F(s_1,\ldots,s_n)= \sum_{i=1}^{n} w_i s_i$$

- Maximum

$$F(s_1,\ldots,s_n) = \max\{s_1,\ldots,s_n\}$$

- k-Minimum

$F(s_1,\ldots,s_n) = s_{i_k}$ with $s_{i1} \leq s_{i2} \leq \ldots \leq s_{in}$

- etc.

Machine Learning LAB

Local Similarity Measures (I)

(for unordered symbolic and integer/real-valued attribute types)

Similarity Tables

- for attributes with symbolic type $T_A=\{v_1,\ldots,v_k\}$
- sim table/matrix $sim_A(x,y)=s[x,y]$
- example: attribute ``RAM-Type“ with T_A={SD,DDR,RD}

q \ c	SD	DDR	RD
SD	1.0	0.9	0.75
DDR	0.5	1.0	0.75
RD	0.25	0.5	1.0

RAM-Type

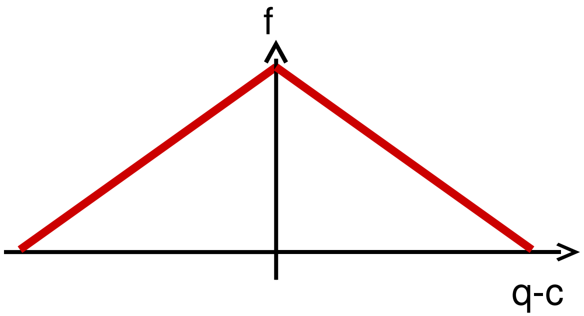
- reflexive similarity measure iff. diagonal elements s[k,k]=1
- symmetric similarity measure iff. $s=s^T$

Difference-Based Similarity Functions

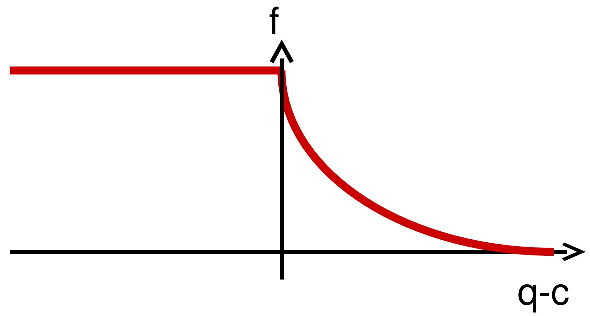
- for attributes with numeric type (e.g. integer or real-valued)
- similarity is based on the numerical difference between case and query value
  - linearly scaled domains: $sim_A(x,y)=f(x-y)$
  - exponentially scaled domains: $sim_A(x,y)=f(\log(x)-\log(y))$
- typical requirements on f
  - $f:\Re\rightarrow[0,1]$ or $f:Z\rightarrow[0,1]$
  - f(0)=1 (reflexivity)
  - f(x) is monotonously falling/increasing
- examples: next slide

Local Similarity Measures (II)

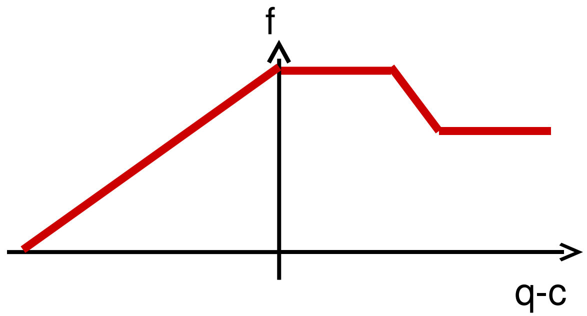
(for unordered symbolic and integer/real-valued attribute types)



- symmetric
- simple distance



- asymmetric
- x = query q, y = case c
- query is minimal requirement
- e.g. minimal horse power of a car required by a customer



- asymmetric
- x = query q, y = case c
- query depicts maximal value
- e.g. maximal price of a product
- background knowledge included (for q-c>0)

Local Similarity for Other Types

➔ *not considered here*

- ordered symbolic data types
  - e.g. T_A={small,average,tall}
- taxonomic data types
  - elements of T_A can be arranged within a taxonomical (tree) structure
  - e.g. attribute to describe the types of CPUs

4. SIMILARITY-BASED RETRIEVAL

How to retrieve a query's nearest neighbour(s)?

Sequential Retrieval

- Retrieval Task
  - Input
    - case base $CB=\{c_1,\ldots,c_n\}$
    - similarity measure sim
    - query (new problem) q
  - Output
    1. most similar case c_i

       or
    2. m most similar cases $\{c_{i_1},\ldots,c_{i_m}\}$

       or
    3. all cases $\{c_{i_1},\ldots,c_{i_t}\}$ which have at least a similarity of sim_{min} to q
  - Main Problem: Efficiency
  - Question: How can the case base be organised in such a way to support an efficient retrieval?

- Sequential Retrieval
  - iterates over all $c \in CB$ and calculates sim(c,q)
  - returns the most similar / m most similar cases to q
  - complexity: O(n)
  - Advantages
    - easy to implement
    - no index structures to maintain
    - usability of arbitrary similarity measures
  - Drawbacks
    - problematic for large case bases
    - effort independent of query
    - effort independent of m

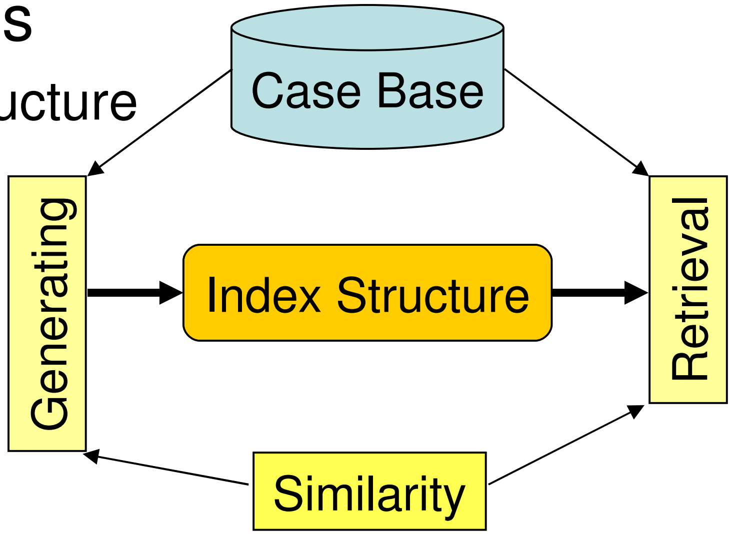
Two-Stage Retrieval

- Idea: **MAC/FAC** (many are called, few are chosen)
  1. preselection of possible solution candidates $M_q \subseteq CB$, where $M_q = \{ c \in CB \mid fac(q,c) \}$
  2. use sequential retrieval on M_q

  ➔ ***Problem:*** Finding of an adequate predicate fac
- Examples for predicate fac
  – partial equality: fac(q,c) iff. q and c are identical w.r.t. at least one attribute
  – local similarity: fac(q,c) iff. q and c are sufficiently similar w.r.t. to each attribute
  – partial local similarity: fac(q,c) iff. q and c are sufficiently similar w.r.t. to at least one attribute
- Advantage: good performance $|M_q|$ if is small
- Drawbacks
  – retrieval errors may occur → α -error: A case c that is sufficiently similar to q w.r.t. sim is not found (because not considered during preselection).
  ➔ completeness of retrieval is not guaranteed
  – determination of an adequate predicate for preselection is usually difficult

Case Retrieval with kd-Trees

- Index-Oriented Retrieval Procedures
  - preprocessing: generating an index structure
  - retrieval: exploit the index structure to efficiently access the cases



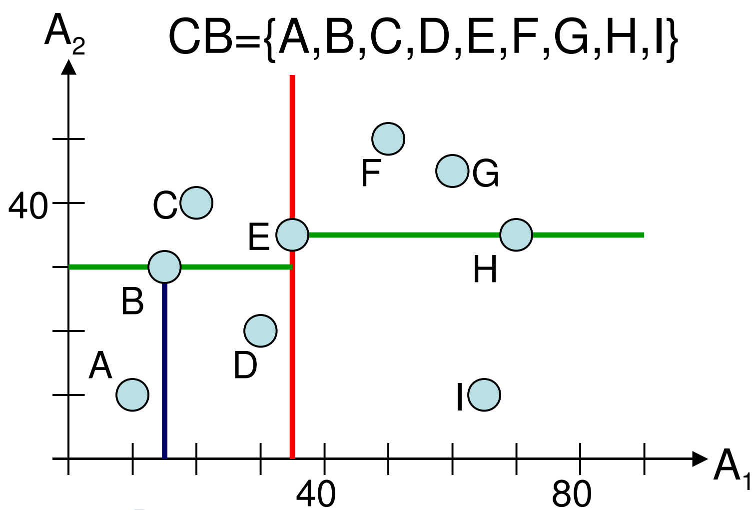
- Possible Index Structur: kd-Tree
  - A ***kd-Tree*** is a k-dimensional binary search tree to support an efficient search over data sets.
  - Idee: partitioning of the data (here: the case base) into small intervals.
  - ordering within a binary tree (similar to a decision tree)
  - during retrieval
    - stepping through the tree from root to the leaves
    - backtracking is possible (unlike in decision trees)

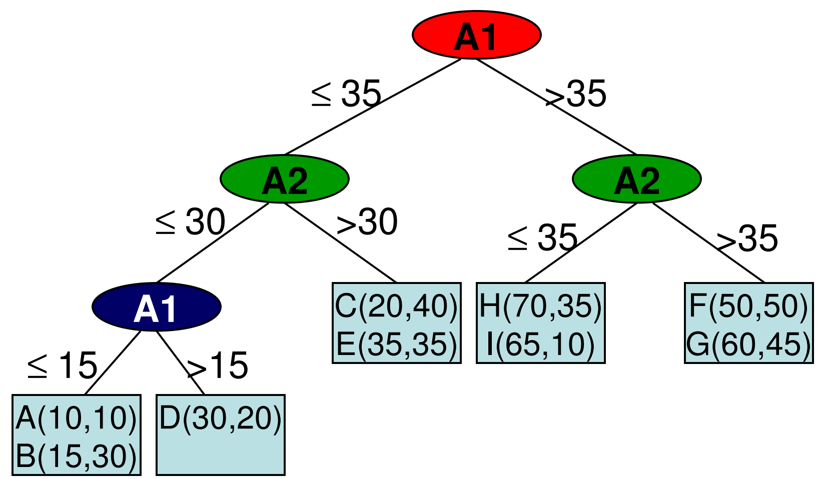
Definition of kd-Trees

- Input
  - k ordered domains $T_1,\ldots,T_k$ for attributes $A_1,\ldots A_k$
  - case base $CB \subseteq T_1 \times \ldots \times T_k$
  - parameter b (bucket size)
- **Definition:** A *kd-Tree T(CB) for case base CB* is a binary tree, that is defined as
  - if $|CB| \leq b$: T(CB) is a leaf of the tree (called bucket), denoted CB
  - if $|CB| > b$: T(CB) is a tree whose
    - root is denoted with an attribute A_i and a value $v_i \in T_i$ and
    - two sub-trees $T_{\leq}(CB_{\leq})$ and $T_{>}(CB_{>})$ are kd-trees, too, with
      - $CB_{\leq} := \{ (x_1,\ldots x_k) \in CB \mid x_i \leq v_i \}$ and
      - $CB_{>} := \{ (x_1,\ldots x_k) \in CB \mid x_i > v_i \}$

Properties of kd-Trees

- kd-tree partitions the case base
  - root represents the entire case base
  - a leaf (bucket) represents a subset of the case base that does not have to be further partitioned
  - at each inner node the case base is partitioned, being divided on the basis of some specific value of an attribute
- Example:





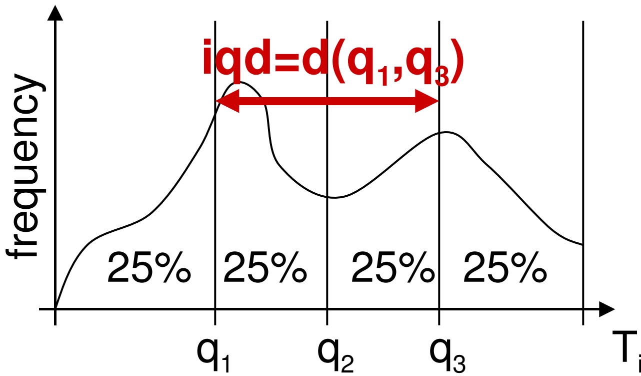
Generating kd-Trees

- Algorithm

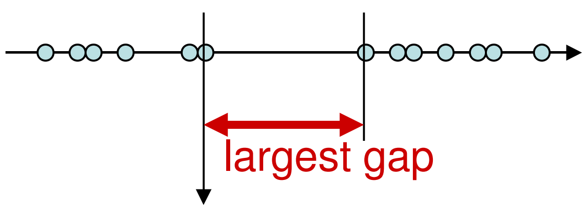
```
PROCEDURE CreateTree(CB): kd-Tree
if |CB|<b
then
  return leaf node marked with case base CB
else
  A_i := choose_attribute(CB)
  v_i := chose_split_value(CB,A_i)
  return
    tree whose root is marked with A_i and v_i
    and which has sub-trees
      CreateTree( {(x_1,…,x_k)∈CB | x_i<=v_i} )
      CreateTree( {(x_1,…,x_k)∈CB | x_i>v_i } )
```

Attribute Selection and Splitting Values

- various methods usable for attribute selection
  - entropy-based
  - inter-quartile distance
    ➔ choose the attribute with the biggest inter-quartile distance **iqd**



- determination of splitting values
  - median splitting: choose median as splitting value
  - maximum splitting: search for the ``largest gap“



Retrieval Algorithm Using kd-Trees

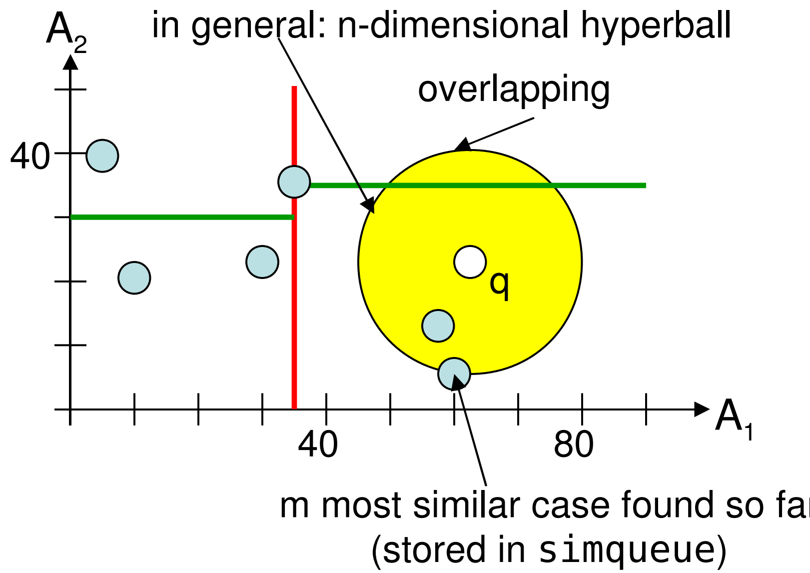
- Algorithm
  - simqueue is a global data structure holding the m most similar cases as well as their corresponding similarities with respect to q

```
PROCEDURE Retrieve(K: kd-Tree, q: Query)
if K is leaf node
then
  forall c∈CB_K do
    if sim(q,c) > simqueue[m].sim
    then insert c into simqueue
else
  A_i := the attribute K is marked with
  v_i := the splitting value K is marked with
  q_i := the attribute A_i's value of q
  if q_i<=v_i
  then
    Retrieve(K_≤,q)
    if BOB-Test is fulfilled then Retrieve(K_>)
  else
    Retrieve(K_>,q)
    if BOB-Test is fulfilled then Retrieve(K_≤)
  if BWB-Test is fulfilled
  then terminate retrieval returning simqueue
  else return
```

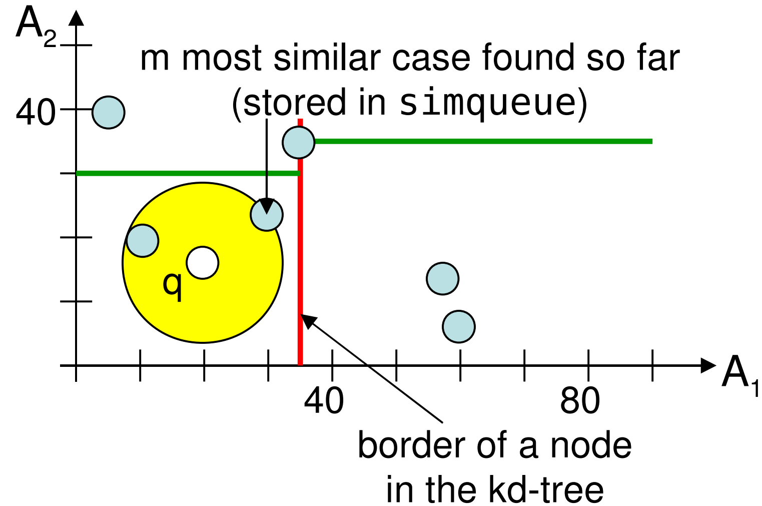
Machine Learning LAB

BOB- and BWB-Tests

- BOB-Test:
  Can there be – in the neighbouring sub-tree – any more similar cases (to query q) than the m most similar cases already found?



- BWB-Test:
  Is it guaranteed that there is no case in a neighbouring sub-tree which is more similar to the query q than the m-most similar case found so far?



Discussion of kd-Tree Retrieval (I)

- Restriction
  - Retrieval using kd-trees guarantees finding the m nearest neighbours, if the similarity measure used fulfills the following condition:
    Compatibility with ordering and monotony:
    $\forall x_1,\ldots,x_n$ and x_i', x_i'' if $x_i <_i x_i' <_i x_i''$
    then $sim((x_1,\ldots,x_n), (x_1,\ldots,x_i',\ldots x_n))$
    $\geq sim((x_1,\ldots,x_n), (x_1,\ldots,x_i'',\ldots,x_n))$
- Advantages
  - efficient retrieval
    - significant savings in lower dimensions
    - due to the tree structure at least $O(\log_2 n)$ operations (comparisons) must be made
    - best case: the similarity between the query and only one case must be calculated
  - effort depends on the number m of most similar cases to find
  - incremental extension of the kd-tree is possible

Discussion of kd-Tree Retrieval (II)

- Drawbacks
  - higher costs for building up the index structure (kd-tree)
  - restrictions implied by kd-trees
    - usability for ordered domains only
    - unknown attribute values are difficult to handle
    - only for monotonous similarity measures that are compatible with the ordering of the respective attribute‘s domain
  - dimensionality of the problem is critical
    - in higher dimensions, often the similarity to very many (or even all – like in linear retrieval) cases must be calculated
    - reason: in higher dimensions, there is the tendency that a query has nearly the same similarity to very many cases; thus the BOB test has to be applied more frequently
- Further Developments
  - R-Tree (Guttman et al.), R*-Tree (Kriegel et al.)

Other Retrieval Methods

- Several further advanced retrieval approaches
  - high efficiency
  - general usability depends on problem setting (e.g. case modelling)
- Examples
  - Case Retrieval Nets [Burkhardt&Lenz]
  - Retrieval with ``Fish and Shrink“ [Schaaf, 1996]
  - Case Retrieval on Top of Relational Databases Utilising SQL [Schumacher, 2000]

1 2 3 4 5 6 7 8

Part 1 (covered today)

Part 2: Outlook

1.Introduction
(What is CBR?)
2.Knowledge and Case Representation
(What knowledge is in a CBR system? How can cases be represented?)
3.Similarity
(When is a new problem similar to an old one? What types of similarity measure may be used?)
4.Similarity-Based Retrieval
(How to retrieve a query's nearest neighbors?)

5. SOLUTION ADAPTATION
How to adapt existing solutions to be applicable for the problem at hand?
6. LEARNING IN CASE-BASED REASONING
Where are the explicit links between CBR and Machine Learning?
7. APPLICATIONS AND TOOLS
Is CBR actually employed in practice? Are there tools available I may use for trying out some of the things introduced in this talk?
8. REFERENCES
Where can I find more about CBR?

Machine Learning LAB

Thanks!

Questions?

tgabel@informatik.uni-freiburg.de