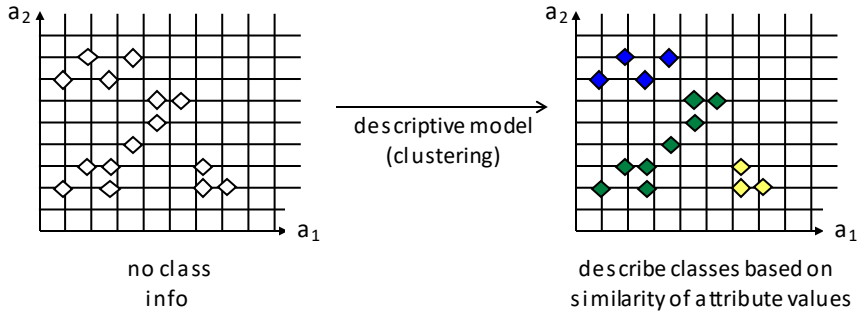
3. CLUSTERING

Clustering and Classification

Given a dataset of *objects* described by *attributes*, build a model that assigns objects to a *class (or label)*



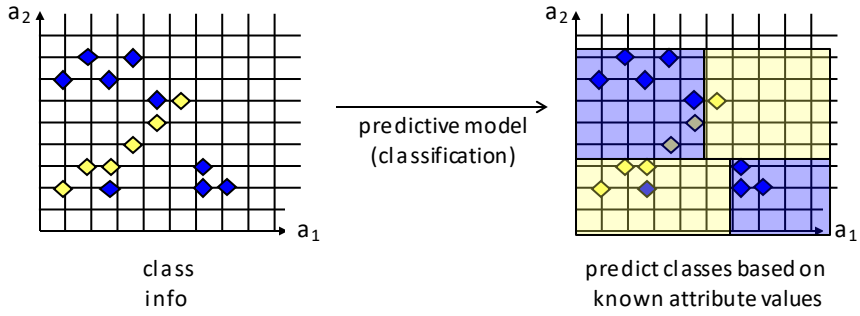
For inferring global models of data collections there exist two types of approaches: descriptive and predictive modeling. We illustrate the difference among them by an example.

We assume that a set of data items (or objects) with two attributes a1 and a2 is given. Assume the global model we are interested in is a classification (or as often said labeling) of the data items.

In descriptive modeling we just know the data items, as indicated by the points in the 2-dimensional grid. A descriptive modeling technique, such as clustering, produces classes, which are not known in advance. For doing this it relies on some criteria that specify when two data items probably belong to the same class. Such a criteria is usually based on a similarity measure.

Clustering and Classification

Given a dataset of *objects* described by *attributes*, build a model that assigns objects to a *class*



A predictive modeling technique, such as classification, starts from a given classification (or labeling) of data items. Using that classification of the dataset the classification method infers conditions on the properties of the data objects, that allow to predict the membership to a specific class. For example, the prediction could be based on a partitioning of the attribute values along each dimension, as shown in the figure on the right. There, first attribute a1 is partitioned into two intervals, and for each of the intervals a different partitioning of the attribute a2 is used to determine the regions corresponding to classes. Misclassifications may occur as seen in the example.

Clustering

Model: partition a set of objects into clusters

- Objects that belong to the same cluster are similar
- Objects that belong to different clusters are dissimilar

Clustering belongs to *unsupervised* machine learning

- Objects do not have class information

Both clustering and classification aim at partitioning a dataset into subsets that have similar characteristics. Different to classification, clustering does not assume any prior knowledge, which are the classes/clusters to be searched for. There exist no class label attributes, that would tell which classes exist. Thus clustering serves in particular for exploratory data analysis with little or no prior knowledge.

One important application of clustering is information retrieval. The basic problem of information retrieval, i.e., finding a set of documents matching a query, can be interpreted as a clustering problem, where the goal is to find two clusters of documents, namely the cluster of relevant ones and the cluster of non-relevant ones. With the tf-idf similarity metrics in fact the tf-measure served to measure intra-cluster similarity for the two document clusters, whereas the idf-measure served to measure inter-cluster dissimilarity of the document clusters.

Usage of Clustering

Information retrieval

- C_1: relevant to query, C_2, C_3: not relevant

Web content classification

- C_1: sport, C_2: politics, C_3: economics …

Customer behaviour analysis

- C_1: diapers-beer buyers, C_2: Fri afternoon buyers …

Image/video processing

- C_1: image with faces, C_2: image with animals, …

Clustering finds wide usage in many applications, where the different types of classes of objects (the labels) are not known in advance. The standard process is to first perform a clustering, and then inspecting the result. Inspection of the result may be performed manually (e.g. a user inspects some images in image clusters, and recognizes that in one cluster all, or most, images contain faces), or features of the objects in the cluster are extracted (e.g. frequent keywords if the objects are documents) and a user inspects the result. For example, if the keywords of documents in a cluster would turn around sports, the user would decide the cluster is on sports and would label it as sports.

Clustering

Problem

Given a database D with n data items described by d attributes

Find

Partition of D into k clusters

such that

Intra-cluster similarity is high

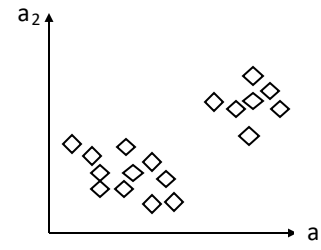
Inter-cluster similarity is low

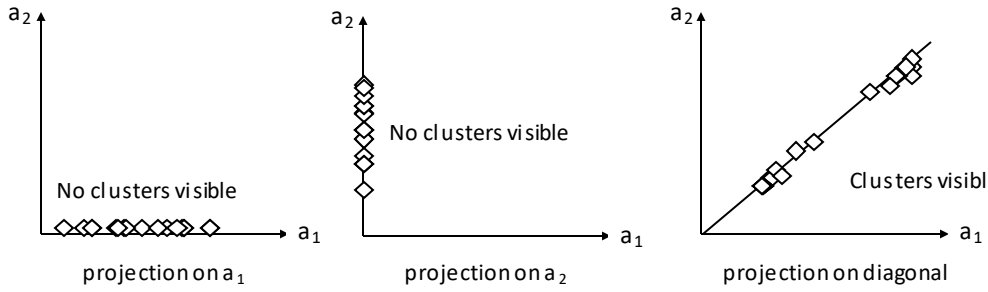
In its simplest formulation the clustering problem can be described in a way analogous to the vector space retrieval model. Given a database of data items that are represented by d-dimensional vectors (feature vectors), partition the database into k clusters. Not all elements need to be assigned to a cluster, then we consider those elements as noise in the data collection. Frequently used similarity measures include Euclidean distance and Manhattan distance.

Characteristics of Clustering Methods

Quantitative

- Scalability: high n
- Dimensionality: high d





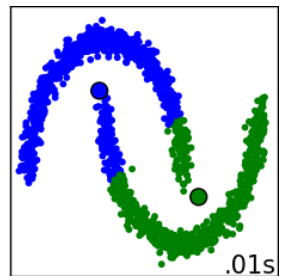
Clearly, clustering methods have to work efficiently for large datasets. Another scalability problem clustering methods have to deal with is dimensionality: in general high-dimensional data (large d) is too sparsely distributed in its high-dimensional space in order to identify meaningful clusters. Therefore the data is often projected into a lower dimensional space. Doing so makes the clustering method sensitive to the choice of dimensions used for projection. The figure illustrates this problem for the case of 2 dimensions:

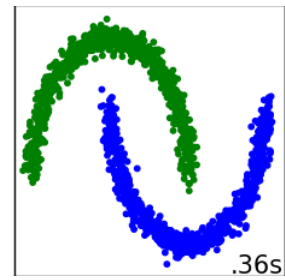
In the 2-dimensional representation we see clearly two clusters. However, if we project either on the a1 or a2 axes, there is no cluster structure visible. Only when we project on the diagonal we will also see the clusters in one dimension. Thus, when projecting into lower dimensions the choice of the subspaces used for projection is crucial. The number of choices for projection dimensions grows on the other hand combinatorial, which makes the problem computationally hard.

Characteristics of Clustering Methods

Qualitative

- Different types of attributes
  - numerical, categorical
- Type of shapes
  - spheres, hyperplanes ...





Qualitative criteria for assessing the performance of a clustering method concern the ability of dealing with continuous as well as categorical attributes, and the type of clusters that can be found. Many clustering methods can detect only very simple geometrical shapes, like spheres, hyperplanes etc.

Characteristics of Clustering Methods

Robustness

- Sensitivity to noise and outliers
- Sensitivity to the processing order

User interaction

- Incorporation of user constraints (e.g., number of clusters, maximal size of clusters)
- Interpretability and usability

Clustering methods can be sensitive both to noisy data and the order of how the records are processed. In both cases it would be undesirable to have a dependency of the clustering result on these aspects which are unrelated to the nature of data in question.

Finally, an important criterion is the ability of how well a clustering method can incorporate user requirements both in terms of information that is provided from the user to the clustering method (in terms of constraints), which can guide the clustering process, and in terms of what information is provided from the method to the user.

Partitioning Methods: Score Function

Model: Partitioning

Given database D of n objects, split D into k sets $C_1,\ldots,C_k$ such that $C_i \cap C_j = \emptyset$ for all $C_i \neq C_j$ and $\bigcup_i C_i = D$

Score function: find C_i that minimises loss function J

$$J = \frac{1}{n}\sum_{i=1}^{k}\sum_{x_j \in C_i} \left\|x_j - \mu_i\right\|^2, \mu_i = \frac{1}{|C_i|}\sum_{x_j \in C_i} x_j$$

μ_i = centroid of C_i

More formally partitioning methods can be described as methods that attempt to optimize a cost function for a set of clusters. We show here the formulation for k-Means.

The cost function minimizes the distances of the cluster elements from the cluster representation, which is the centroid of the cluster.

Partitioning Methods

Optimal algorithm: enumerate all partitions and pick the best (not practical)

Heuristic algorithms:

- **K-means**: cluster is represented by the ***point*** whose *mean* distance with the objects in the cluster is minimal
- **K-medoids**: cluster is represented by the ***object*** whose *mean* distance with the objects in the cluster is minimal
- **K-medians**: cluster is represented by the ***point*** whose *median* distance with all the objects in the cluster is minimal

Since an exhaustive enumeration for finding the optimal partitioning is not practical various heuristic methods have been proposed. On class of algorithms are partitioning methods, that represent clusters be selected points and objects. The difference between using points and objects, is that points are not part of the dataset that is being clusters.

Partitioning methods are one of the most basic and widely used approaches to clustering. Partitioning methods attempt to partition the data set into a predetermined number k of clusters while maximizing the intra-cluster similarity.

Suppose we have a dataset of pictures and we want to cluster them. Which partitioning algorithm seems more appropriate?

A. k-medoids
B. k-medians
C. k-means
D. none of the above

K-means Algorithm

Initialise k random points as **cluster centers**

Assign each object to the "nearest" cluster center

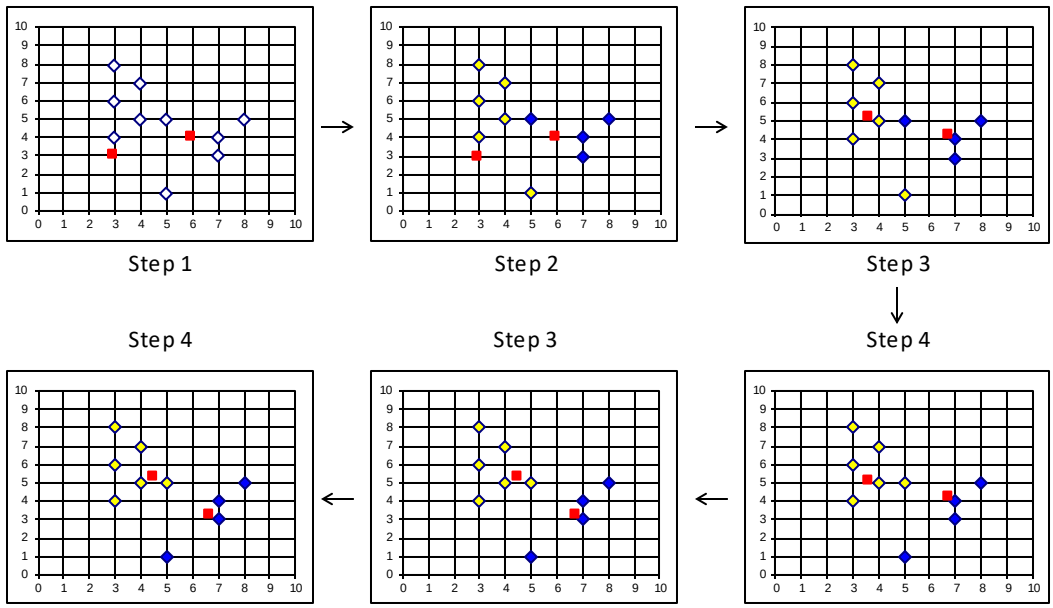
generates a partitioning $C_1,...,C_k$ of D

While partitioning changes

- for each cluster, calculate the centroid of the points and set it as new cluster center
- assign each point to the "nearest" cluster center

In the widely used k-means algorithm clusters are represented by the centroid of their elements. The algorithm starts from some initial partitioning of the data. Initial centroids can be k random poins in the space or k objects (Forgy initialisation). Then, it iteratively reassigns data points to the closests centroids till the algorithm converges. The distances are computed according to a given metrics, e.g. Euclidean distance.

K-means Algorithm: Example



Characteristics of K-means

Advantages

- Efficient: O(tkn)
- n is #objects, k is #clusters, t is #iterations (k,t << n)
- Converges fast

Shortcomings

- Often terminates at a local minimum
- Needs a distance function to compute the mean
- Needs to specify k in advance
- Does not handle noisy data and outliers
- Clusters only have convex shapes

We assess here the properties of k-means following the list of criteria for evaluating clustering methods that we have introduced earlier.

K-means for Categorical Attributes

Needs a distance function to compute the mean:

Matching coefficient for categorical attributes

- a and b are objects with d attributes

$$distance(a, b) = \frac{|\{i \mid a_i \neq b_i\}|}{d}$$

Example:

	domain	location	category
x_1 =	.com	US	sport
x_2 =	.com	CH	econ

distance(x_1, x_2) = 2 / 3

One of the short-comings of k-means is that it relies on the existence of a distance function. Thus, when the attributes are categorical such a distance function needs to be constructed as it is not naturally available. A simple approach to define a distance function in this setting is to count the number of attribute values that do not match (mismatches) and divide by the total number of attributes. Note that this function is a distance function, as it satisfies the properties of a distance function:

$d(x,y) \geq 0$

$d(x,y) = 0$ iff $x = y$

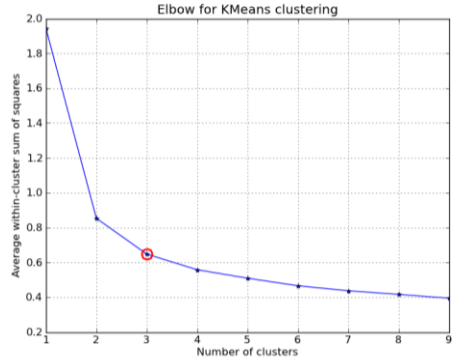
$d(x,y) = d(y,x)$

$d(x,z) \leq d(x,y) + d(y,z)$

Choosing Parameter k for K-means

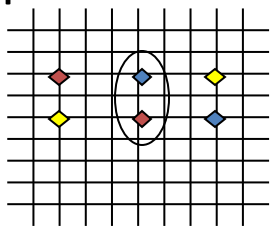
Needs to specify **model parameter** k in advance

- Execute for k = 1, 2, 3 ...
- Plot the score J against k
- Pick k such that J(k) ~ J(k+1)

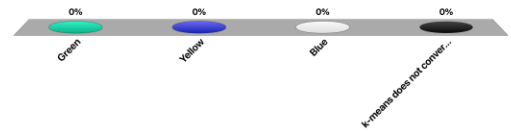


To find a suitable number of k we rely on the following intuition. The higher the value of k, the lower the within-cluster sum of squares. However, creating too many clusters is useless. Thus we want to find the smallest k after which the within-cluster sum of squares decreases "slowly"

What will be the color of the middle points after convergence (k=3)?



A. Green
B. Yellow
C. Blue
D. k-means does not converge

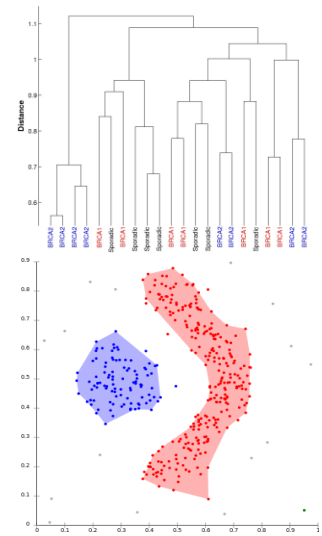


Advanced Clustering

Distance measures for mixed attributes

Other methods

- Density-based clustering
- Hierarchical clustering
- Online incremental clustering



There is a lot of existing work on clustering, including more complex distance functions for mixed (numerical + nominal) attributes and similarity functions that computes "how similar" are two objects (e.g., cosine similarity)

There exists also other methods beyond partitioning. Density-based methods are able to discover clusters of arbitrary shape by checking if the number of objects in the neighbourhood of a given objects is above a certain density threshold. Bottom-up hierarchical methods starts from a number of cluster equal to the number of objects (i.e., each object is a cluster) and then iteratively merge similar clusters. Online incremental methods create clusters in an incremental way, by placing each new incoming object into an existing cluster or into a new one.

Density-based Clustering - DBSCAN

Clustering based on a local, density-based criterion

Properties

- Discovers clusters of arbitrary shape
- Handles noise
- Clusters in one scan
- No predetermined number of clusters
- Model parameters: density parameters

We have identified a number of shortcomings of K-means. Density based clustering addresses a number of those, in particular the need to specify the number of clusters in advance, the possibility to discover non-convex clusters, handling local minima and handling of outliers. Of course, this will not come for free: on the one hand also density-based clustering requires specification of (other) model parameters, and as we will see, the are more costly.

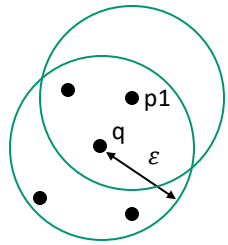
Basic Notions

We assume a distance metric d is given

***ε*- neighborhood**: $N_\varepsilon(q) = \{p | d(p,q) < \varepsilon\}$

A point q is **a core point** if $|N_\varepsilon(q)| \geq \mu$

ε and μ are model parameters



q is a core point with $\mu = 5$
p1 is not a core point with $\mu = 5$

We first introduce the basic notions used in density-based clustering. For each point we can consider its epsilon-neighborhood. If we find within such a neighborhood a certain number of points, we consider such a point as being inside a cluster and call it a core point. The definition of core points also introduces the two model parameters the density-based clustering relies on, epsilon and mu.

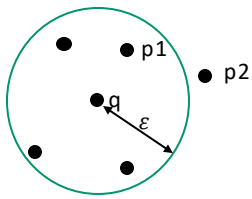
Basic Notions: Directly Density-Reachable

A point p is **directly density-reachable** from q if

$$p \in N_\varepsilon(q) \text{ and } |N_\varepsilon(q)| \geq \mu$$

A point that is directly density-reachable but not a core point is a **border point**

A point that is not directly density-reachable is an **outlier**

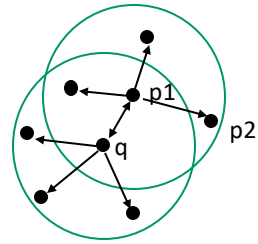


p1 directly density-reachable from q with $\mu = 5$
p2 not directly density-reachable from q with $\mu = 5$
q is not directly density-reachable from p1

Based on the notion of core points, we can next introduce the notion of directly-density reachable. A point is directly-density reachable if it is in the neighborhood of a core point. Since not every point in the neighborhood of a core point is a core point itself, we identify a second type of points, border points. They are directly-density –reachable, informally this means they are part of a cluster, but not core points. Thus they define the border of a cluster. In addition, certain points might not be directly-density reachable at all, these are considered as outliers.

Directly Density-Reachable: Properties

Direct density-reachability induces a **directed graph** on the points

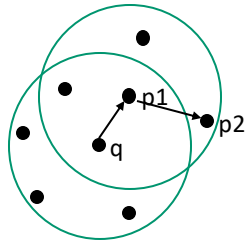


Direct-density reachable induces a directed graph structure on the points. Note that border points are connected with a simple directed link coming from a core point, whereas two core points are connected by bi-directional links.

Basic Notions: Density-Reachable

A point p is **density-reachable** from a point q with ε, μ if there is a chain of points p_1, …, p_n, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i

- Asymmetric relationship



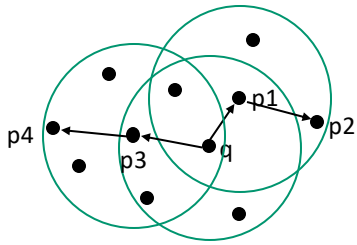
p2 density-reachable from q with $\mu = 5$
p1 and q not density-reachable from p2
q density-reachable from p1

Next we consider the transitive closure of the direct-density reachability, resulting in the notion of density-reachability. A point is density-reachable when it can be reached through a chain of direct density reachability links. Note that this relationship is asymmetric. A border point can be density reachable from a core point, but the inverse is not true, since the border point has no reverse link.

Basic Notions: Density-Connected

A point *p* is **density-connected** to a point *q* with ε, μ if there is a point *r* such that both, *p* and *q* are density-reachable from *r* with ε, μ

– Symmetric relationship

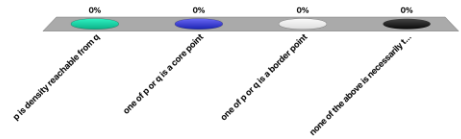


p2 and p4 density-connected with $\mu = 5$

To introduce a symmetric relationships that connects all points within the same cluster, we define the notion of density-connected. Two points are density connected if the can reached both from the some (core) point. This allows then also to connect different border points.

If p and q are density connected, then …

A. p is density reachable from q
B. one of p or q is a core point
C. one of p or q is a border point
D. none of the above is necessarily true



Clusters and Noise

Definition: a **cluster** C satisfies

- **Maximality**: if q in C is a core point, and p is density reachable from q, then also p is in C
- **Connectivity**: any two points in C must be density connected

Properties

- Connectivity implies that a cluster contains at least one core point
- The set of clusters is unique
- Clusters are not necessarily disjoint

Now we have all the notions in order to define what a cluster is in the context of density-based clustering. The basic idea is that density-connected points are considered to belong to the same cluster and that points not density-connected to any other point are considered as noise. To make the definition precise we can state the two properties of maximality and connectivity. With this two properties one can prove that the clusters are uniquely defined,

DBSCAN Algorithm: Initialization

Construct a directed graph G using direct density-reachability

Initialize

- V_{core} = set of core points
- P = set of all points
- set of clusters C = {}

We have now introduced the prerequisites to describe the DBSCAN algorithm for density-based clustering. The algorithm starts by first computing the direct density-reachability graph. This requires to scan the whole database and to compute the neighborhoods of each point. Without any further optimizations this requires each point with every other point, thus a cost that is square in the size of the database. Spatial indexing could be used to reduce this cost. Then three variables are initialized, the set of all core points, the set of all points and the set of clusters found, which initially is empty.

DBSCAN Algorithm: Cluster Construction

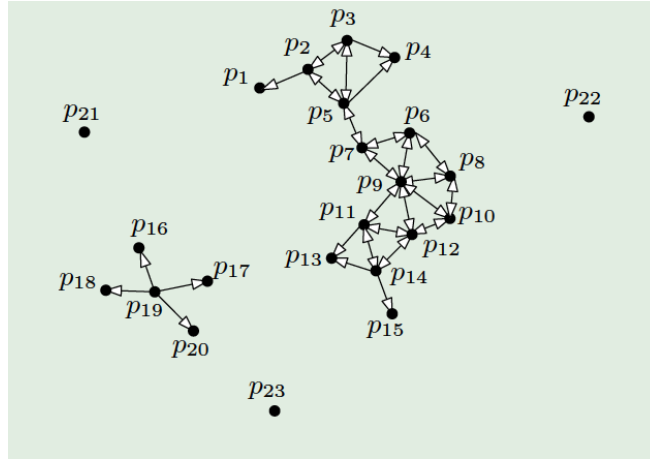
while V_{core} not empty

- select a point p from V_{core} and construct *S(p)*, the set of all points density-reachable from p: breadth-first search on G starting from p
- $C = C \cup \{S(p)\}$
- $P = P \setminus S(p)$
- $V_{core} = V_{core} \setminus S_{core}(p)$, where $S_{core}(p)$ = core points in *S(p)*

Mark remaining points in P as unclustered

For constructing clusters the algorithm packs any core point that has not yet been included in a cluster, and computes all points that are density-reachable from that bound. This set of points can be found be performing a breadth-first search in the directed graph G starting from p and this will return a uniquely defined set of points forming a cluster. After such a cluster has been found the three variables are V_core, C, and P are updated. The new cluster is added to the set of clusters C, the points that have been included in the cluster are removed from P and thus are known to be assigned to a cluster, and the core points contained in the new cluster are removed from V_core and can thus no longer be used as starting point to find a new cluster. Once V_core is empty, we know that no more new clusters can be found. Still, the set P might be non-empty. This means, that those points are outliers that do not belong to any clusters and they are marked as such. Note that the set P is not maintained to indicate the points that need still to be clustered (some points might belong to multiple clusters), but to check which points become never part of a cluster and are finally identified as noise.

Example



Source: Yufei Tao, Chinese University of Hong Kong

First cluster: choose p_2 and compute $S(p_2) = \{p_1,\ldots,p_{15}\}$

Then, $V_{core} = \{p_{19}\}$

Second cluster: $\{p_{16},\ldots,p_{20}\}$

$p_{21},\ldots,p_{23}$ are outliers

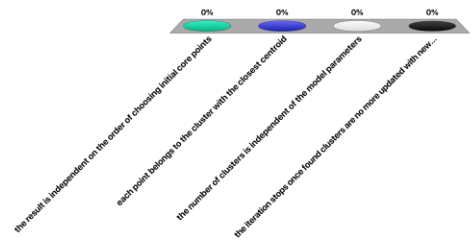
This example illustrates how DBSCAN works. Initially V_core would be {p2, p3, …, p19}. If we select p2 as first core point to construct a cluster we obtain the "big" cluster S(p2). Then the only remaining core point is p19, which gives the second cluster. The remaining points are outliers. The figure also illustrates the difference between core and border points. For example for the big cluster the border points are p1, p4 and p15.

In density-based clustering, which points can belong to multiple clusters?

A. Core points
B. Border points
C. Outliers
D. None

When executing DBSCAN …

A. the result is independent on the order of choosing initial core points
B. each point belongs to the cluster with the closest centroid
C. the number of clusters is independent of the model parameters
D. the iteration stops once found clusters are no more updated with new points



DBSCAN Complexity

Construction of directed graph $O(n^2)$

Construction of clusters $O(n^2)$

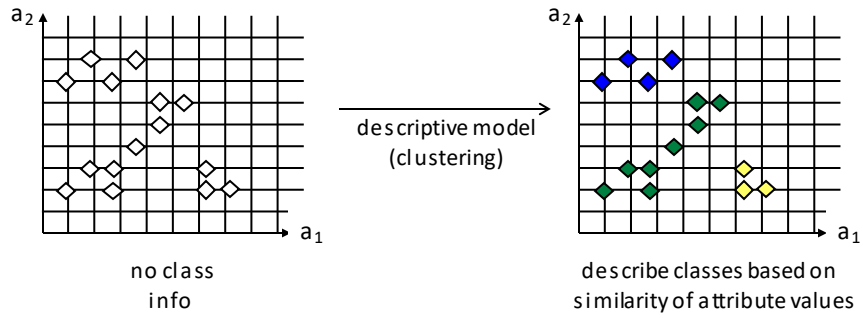
Even in the case of dimension d = 3 the worst case complexity is $\Omega(n^{\frac{4}{3}})$

The complexity of the DBSCAN algorithm is O(n^2) which in practice makes it expensive for larger datasets. It is an open problem in data mining finding better algorithms for density clustering, ideally algorithms with running time O(n log(n)^c) for some constant c. It turns out that even in dimension 3 this cannot be achieved, unless some theoretical problems are solved through breakthroughs. The problem can be reduced to the unit-spherical emptiness checking problem, for which today the best algorithm has complexity O(n^4/3 log^4/3 n)

4. CLASSIFICATION

Clustering and Classification

Given a dataset of *objects* described by *attributes*, build a model that assigns objects to a *class (or label)*



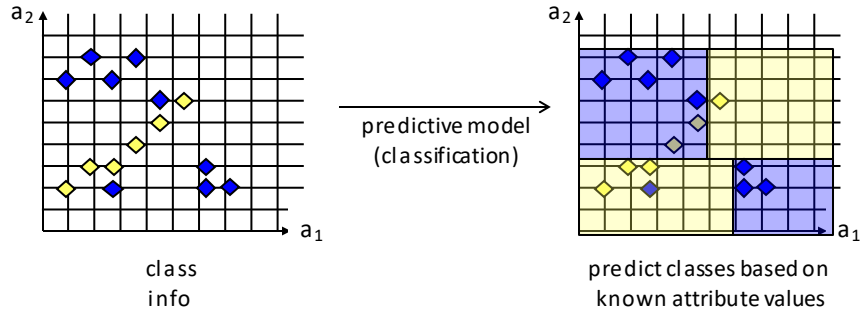
For inferring global models of data collections there exist two types of approaches: descriptive and predictive modeling. We illustrate the difference among them by an example.

We assume that a set of data items (or objects) with two attributes a1 and a2 is given. Assume the global model we are interested in is a classification (or as often said labeling) of the data items.

In descriptive modeling we just know the data items, as indicated by the points in the 2-dimensional grid. A descriptive modeling technique, such as clustering, produces classes, which are not known in advance. For doing this it relies on some criteria that specify when two data items probably belong to the same class. Such a criteria is usually based on a similarity measure.

Clustering and Classification

Given a dataset of *objects* described by *attributes*, build a model that assigns objects to a *class*



A predictive modeling technique, such as classification, starts from a given classification (or labeling) of data items. Using that classification of the dataset the classification method infers conditions on the properties of the data objects, that allow to predict the membership to a specific class. For example, the prediction could be based on a partitioning of the attribute values along each dimension, as shown in the figure on the right. There, first attribute a1 is partitioned into two intervals, and for each of the intervals a different partitioning of the attribute a2 is used to determine the regions corresponding to classes. Misclassifications may occur as seen in the example.

Classification Problem

Input: set of objects with categorical/numerical attributes and one class label

Output: A model that returns the class label given the object attributes

- Model is a function represented as rules, decision trees, formulae

Classification belongs to *supervised* ML

- Objects have class information

Classification creates a **global model**, that is used for predicting the class label of unknown data. Since the classification function is learned from existing data, this approach is also called a supervised learning approach.

Classification is clearly useful in many decision problems, where for a given data item a decision is to be made (which depends on the class to which the data item belongs). Classification is also often called predictive analytics.

Classification: General Approach

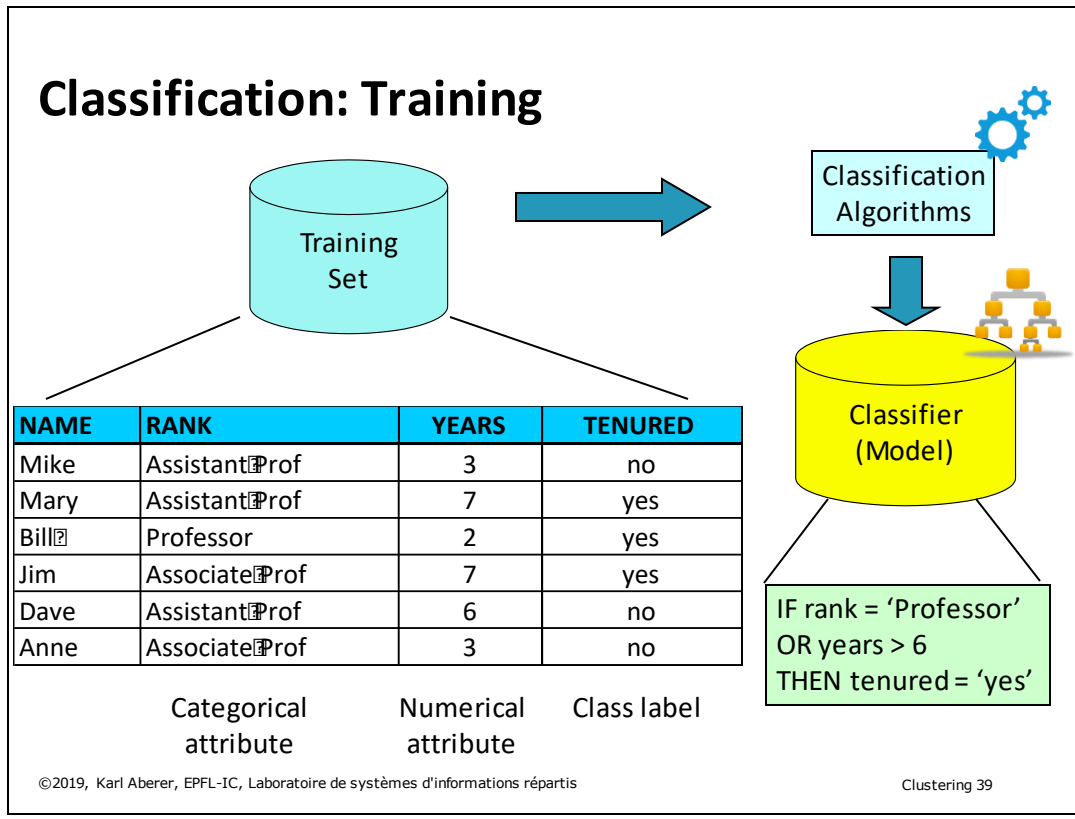
Model is learnt from a set of objects with known labels: **training set**

The quality of the model is evaluated by comparing the predicted class labels with those from a set of objects with known labels: **test set**

- Test set is independent of training set, otherwise over-fitting will occur

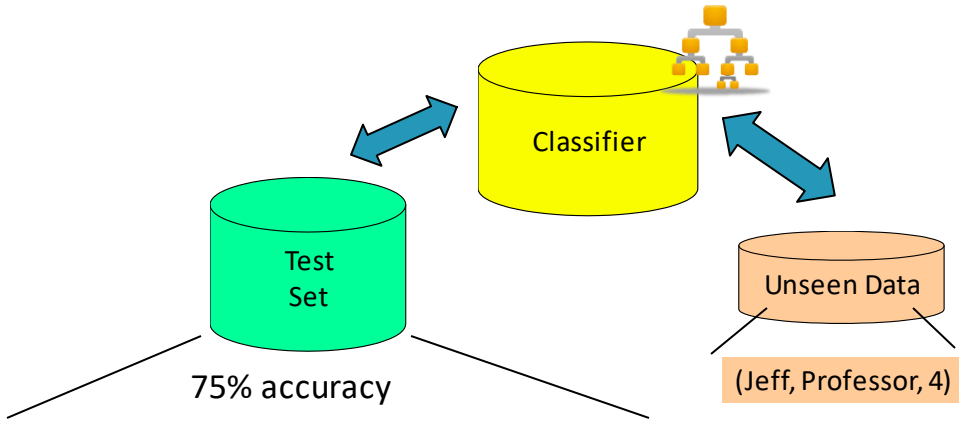
The model is applied to data with unknown labels: **prediction**

In order to test the quality of a model it is tested by using a test set. Assuming that a set of objects with known labels is available, it is split into (typically larger) set that is used for learning model, which is the training set, and a (typically smaller) set of that is used to test the quality of the model, the test set.



In classification the classes are known and given by the class label attributes. In this example, for the given data collection TENURED would be the class label attribute. The goal of classification is to determine rules on the other attributes that allow to predict the class label attribute, as the one shown right on the bottom.

Classification: Model Test and Usage



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Tenured? YES

In order to determine the quality of the rules derived from the training set, the test set is used. We see that the classifier that has been found is correct in 75% of the cases. If rules are of sufficient quality they are used in order to classify data that has not been seen before. Since the reliability of the rule has been evaluated as 75% by testing it against the test set and assuming that the test set is a representative sample of all data, then the accuracy of the rule applied to unseen data should be the same.

Classification: Problem Formulation

Problem

Given a database D with n data items described by d categorical/numerical attributes and one categorical attribute (class label C)

Find

A function f: $X^d \rightarrow C$

rules
decision tree
formula

Such that

classifies *accurately* the items in the *training* set

generalises well for the (unknown) items in the *test* set

We formulate the classification problem formally.

Characteristics of Classification Methods

Predictive accuracy

Speed and scalability

- Time to build the model
- Time to use the model
- In memory vs. on disk processing

Robustness

- Handling noise, outliers and missing values

Interpretability

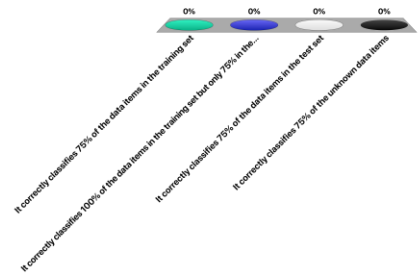
- Understanding the model and its decisions (black box) vs. white box
- Compactness of the model

As for clustering methods, we can also identify for classification methods a range of criteria to assess and compare the properties of different approaches.

Predictive accuracy is the natural main objective to optimize for a classifier. It characterizes how well the classifier performs its job. Often we encounter a trade off between predictive accuracy and the speed and scalability of the method. Methods that achieve high accuracy tend also to be more expensive. As for clustering, also for classification noise and outliers can pose additional problems affecting accuracy. Finally, a very important criterion is the interpretability of the model. In many concrete applications, it is critical that humans are able to understand based on which criteria a classifier takes a decision, e.g. for accountability. Imagine a case, where an assurance policy is refused based on the decision taken by a classifier, and the client would oppose in court. Only with human-interpretable methods the decision could be argued for. However, the most powerful classifiers today tend to produce models that are very hard to interpret for humans, as they represent very complex functions.

If a classifier has 75% accuracy, it means that ...

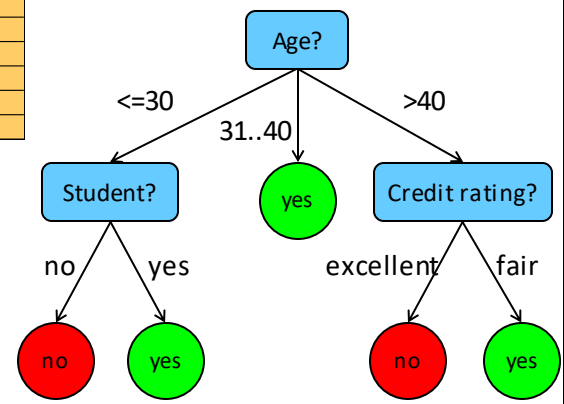
A. It correctly classifies 75% of the data items in the training set
B. It correctly classifies 100% of the data items in the training set but only 75% in the test set
C. It correctly classifies 75% of the data items in the test set
D. It correctly classifies 75% of the unknown data items



Decision Trees

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- Nodes are tests on a single attribute
- Branches are attribute values
- Leaves are marked with class labels



A standard type of classification function is a decision tree. In a basic decision tree, at each level one of the available attributes is used to partition the data set based on the different attribute values. At the leaf level of the decision tree, the values of the class label attribute are found. Thus, for a given data item with unknown class label attribute, by traversing the tree from the root to the leaf according to its data values, its class can be determined by choosing the class label found at the leaf level. Note that in different branches of the tree, different attributes may be used for classification.

A decision tree is constructed in a top-down manner, by recursively splitting the training set using conditions on the attributes. How these conditions are determined is one of the key questions for decision tree induction. After the decision tree construction, it may occur that at the leaf level the granularity is too fine, i.e., many leaves correspond to outliers in the data. Thus, in a second phase such leaves are identified and eliminated.

The key problem in constructing a decision tree is thus to determine the attributes that are used to partition the data set at each level of the decision tree.

Decision Tree Induction: Algorithm

Tree construction (top-down divide-and-conquer strategy)

- At the beginning, all training samples belong to the root
- Examples are partitioned recursively based on a selected "most discriminative" attribute
- Discriminative power determined based on information gain (ID3/C4.5)

Partitioning stops if

- All samples belong to the same class → assign the class label to the leaf
- There are no attributes left → majority voting to assign the class label to the leaf
- There are no samples left

The basic algorithm for decision tree induction proceeds in a greedy manner. First the set of all data objects are associated with the root. Among all attributes one is chosen to partition the set. The criterion that is applied to select the attribute is based on measuring the information gain that can be achieved, or how much uncertainty on the classification of the data objects is removed by the partitioning.

Three conditions can occur such that no further partitions can be performed:

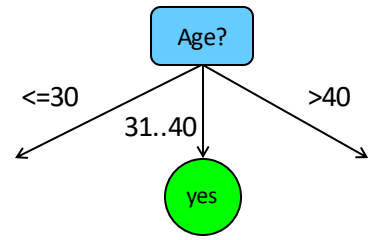
(1) all data objects are in the same class, therefore further splitting makes no sense,

(2) no attributes are left which can be used to split. Still data objects from different classes can be in the leaf, then majority voting is applied.

(3) no data objects are left.

Example: Decision Tree Induction

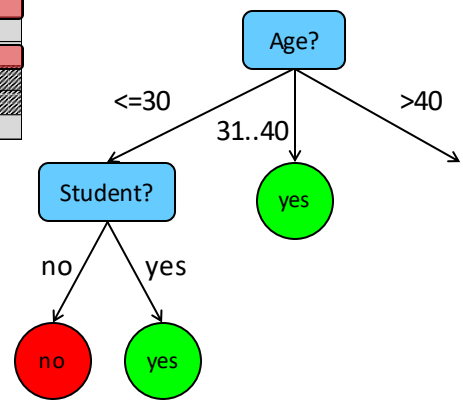
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Based on this approach for attribute selection, we can now illustrate the induction of the decision tree. In a first step, age is chosen for a split. The partition 31..40 contains after the split only instances from one class, the positive class, thus for this branch of the tree the induction terminates.

Example: Decision Tree Induction

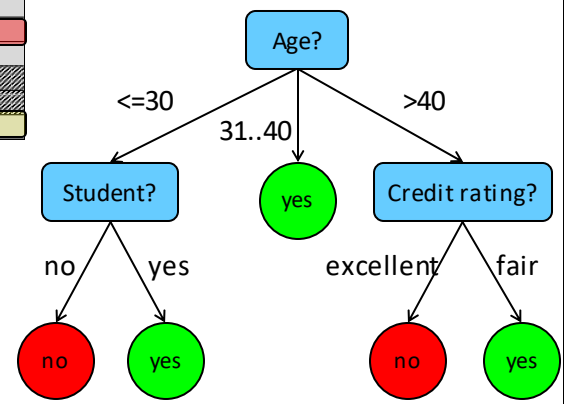
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



For the partition <= 30 we find that the student attribute is the best to be chosen for further splitting. Further splitting makes no more sense, as the two resulting partitions, after splitting by the student attribute, are consisting of either positive or negative instances only.

Example: Decision Tree Induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



Similarly, for the partition >40 we find that credit rating gives the largest information gain. As before, further splitting is no more needed, as the resulting partitions contain only positive respectively negative instances.

Attribute Selection

At a given branch in the tree, the set of samples S to be classified has P positive and N negative instances

The entropy of the set S is

$$H(P, N) = -\frac{P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N}$$

Note

- If P=0 or N=0 H(P, N) = 0 → no uncertainty
- If P=N H(P, N) = 1 → maximal uncertainty

Now we introduce the approach to select the split attributes during the construction of a decision tree. It is shown for the case of binary classification, and generalizes naturally to multiple class labels.

The approach is based on an information-theoretic argument. Assuming that we have a binary category, i.e., two classes **P** and **N** to which objects in S have to be assigned, we can compute the amount of information required to determine the class, by H(P, N), the standard entropy measure, where P and N denote the cardinalities of the classes **P** and **N**. Given an attribute A that can be used for partitioning the data collection, we can calculate the amount of information needed to classify the data after the split according to attribute A has been performed. This value is obtained by calculating H(P, N) for each of the partitions and weighting these values by the probability that a data item belongs to the respective partition.

Attribute Selection: Example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$H_S = H(9, 5) = 0.94$

Age	[<=30]	H(2, 3) = 0.97	Income	[high]	H(2, 2) = 1
Age	[31...40]	H(4, 0) = 0	Income	[med]	H(4, 2) = 0.92
Age	[>40]	H(3, 2) = 0.97	Income	[low]	H(3, 1) = 0.81
Student	[yes]	H(6, 1) = 0.59	Rating	[fair]	H(6, 2) = 0.81
Student	[no]	H(3, 4) = 0.98	Rating	[exc]	H(3, 3) = 1

We illustrate the attribute selection process now for our running example. Initially the data contains P = 9 positive instances and N = 5 negative instances. This results in an entropy of 0.94, i.e. 0.94 bits are required to decide the class of one instance. We compute next the entropies of all partitions that result from splitting all attributes. For example, if we split for Age, we obtain 3 partitions, each with a different distribution of positive and negative instances; and thus with different entropies.

Attribute Selection: Information Gain

Attribute A partitions S into S_1, S_2, … S_v

Entropy of attribute A is

$$H(A) = \sum_{i=1}^{v} \frac{P_i + N_i}{P + N} H(P_i, N_i)$$

The information gain obtained by splitting S using A is

$$Gain(A) = H(P, N) - H(A)$$

Gain(Age) = 0.94 – 0.69 = 0.25 ← split on age

Gain(Income) = 0.94 – 0.91 = 0.03

Gain(Student) = 0.94 – 0.78 = 0.16

Gain(Rating) = 0.94 – 0.89 = 0.05

The information gained by a split can thus be determined as the difference of the amount of information needed for correct classification before and after the split. Thus we calculate the reduction in uncertainty that is obtained by splitting according to attribute A and select among all possible attributes the one that leads to the highest reduction. For our example we can conclude that it is best to split on attribute age.

Attribute Selection: Example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$H_S = H(9, 5) = 0.94$

$H_{Age} = p([<=30]) \cdot H(2, 3) + p([31...40]) \cdot H(4, 0) + p([>40]) \cdot H(3, 2) =$
$= 5/14 \cdot 0.97 + 4/14 \cdot 0 + 5/14 \cdot 0.97 = 0.69$

$H_{Income} = p([high]) \cdot H(2, 2) + p([med]) \cdot H(4, 2) + p([low]) \cdot H(3, 1) =$
$= 4/14 \cdot 1 + 6/14 \cdot 0.92 + 4/14 \cdot 0.81 = 0.91$

$H_{Student} = p([yes]) \cdot H(6, 1) + p([no]) \cdot H(3, 4) = 7/14 \cdot 0.59 + 7/14 \cdot 0.98 = 0.78$

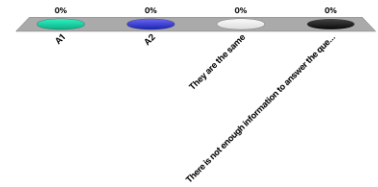
$H_{Rating} = p([fair]) \cdot H(6, 2) + p([exc]) \cdot H(3, 3) = 8/14 \cdot 0.81 + 6/14 \cdot 1 = 0.89$

Next we compute the weighted sum of all entropies of the partitions in a split. The weights correspond to the probability of an instance falling into an element of the partition. Computing this for all attributes shows, that the attribute H results in the lowest entropy, i.e., leaves the lowest remaining uncertainty about the class membership of instances after the split.

Given the distribution of positive and negative samples for attributes A_1 and A_2, which is the best attribute for splitting?

A_1	P	N
a	2	2
b	4	0
A_2	P	N
x	3	1
y	3	1

A. A1
B. A2
C. They are the same
D. There is not enough information to answer the question



Clustering 53

Pruning

The construction phase does not filter out noise → **overfitting**

Pruning strategies

- Stop partitioning a node when large majority of samples is positive or negative, i.e., N/(N+P) or P/(N+P) > 1 - ε
- Build the full tree, then replace nodes with leaves labelled with the majority class, if classification accuracy does not change
- Apply Minimum Description Length (MDL) principle

A common problem in classification is that a classifier may overspecialize and capture noise and outliers in the data, rather than general properties. One possibility to limit overspecialization would be to stop the partitioning of tree nodes when some specific criteria is met (e.g., number of samples assigned to the leaf node). A possible criterion is to stop partitioning when the majority of remaining samples falls into one class. However, in general it is difficult to specify a suitable criterion a priori (e.g. choosing the right value of epsilon).

Another alternative is to first build the complete classification tree, and then, in a second phase, prune subtrees that do not contribute to an efficient classification. Different approaches can be applied to that end: heuristic approaches can identify subtrees that do not contribute to the classification accuracy, and eliminate those. A more principled approach is the use of the minimum description length principle. (MDL).

Minimum Description Length Pruning

Let M_1, M_2, …, M_n be a list of candidate models (i.e., trees). The best model is the one that minimizes

$$L(M) + L(D|M)$$

where

- L(M) is the length in bits of the description of the model (#nodes, #leaves, #arcs …)
- L(D|M) is the is the length in bits of the description of the data when encoded with the model (#misclassifications)

The MDL is based on the following consideration: if the effort in order to specify a class (the implicit description of the class extension through a decision tree) exceeds the effort to enumerate all class members (the explicit description of the class by enumerating its extension), then the subtree is over classifying and non-optimal. To measure the description cost a suitable metrics for the encoding cost, both for trees and data sets is required. For trees this can be done by suitably counting the various structural elements needed to encode the tree (#nodes, #test predicates, # arcs), whereas for explicit classification, it is sufficient to count the number of misclassifications that occur in a tree node.

Extracting Classification Rules from Trees

Represent the knowledge in the form of IF-THEN rules

- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction
- The leaf node holds the class prediction

Rules are easier for humans to understand

Example

IF *age* = "<=30" AND *student* = "*no*"	THEN *buys_computer* = "*no*"
IF *age* = "<=30" AND *student* = "*yes*"	THEN *buys_computer* = "*yes*"
IF *age* = "31…40"	THEN *buys_computer* = "*yes*"
IF *age* = ">40" AND *credit_rating* = "*excellent*"	THEN *buys_computer* = "*yes*"
IF *age* = ">40" AND *credit_rating* = "*fair*"	THEN *buys_computer* = "*no*"

A decision tree can also be seen as an implicit description of a set of classification rules. Classification rules represent the classification knowledge as IF-THEN rules and are easier to understand for human users. They can be easily extracted from the classification tree as described.

Decision Trees: Continuous Attributes

With continuous attributes we can not have a separate branch for each value

- use **binary decision trees**

Binary decision trees

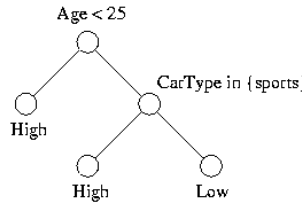
- For continuous attributes A a split is defined by val(A) < X
- For categorical attributes A a split is defined by a subset X $\subseteq$ domain(A)

With continuous attributes it does not make sense to create a separate path in the decision tree for every possible attribute value. Instead, in such a case, a binary decision tree is constructed. Binary decisions can be specified both fro continuous and categorical attributes. For continuous attributes, the binary split is performed by selecting a threshold that separates the instances in those that have a larger and a smaller value than the threshold. For categorical attributes, a subset of attribute values can be chosen that distinguishes the instances in two subsets.

Example: Binary Decision Tree

rid	Age	Car Type	Risk
0	23	family	High
1	17	sports	High
2	43	sports	High
3	68	family	Low
4	32	truck	Low
5	20	family	High

(a) Training Set



(b) Decision Tree

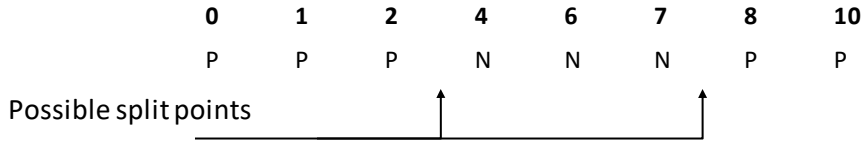
This example shows a dataset with both categorical and continuous attributes and a possible binary decision tree for such a dataset. The class label in the example is Risk.

Splitting Continuous Attributes

Approach

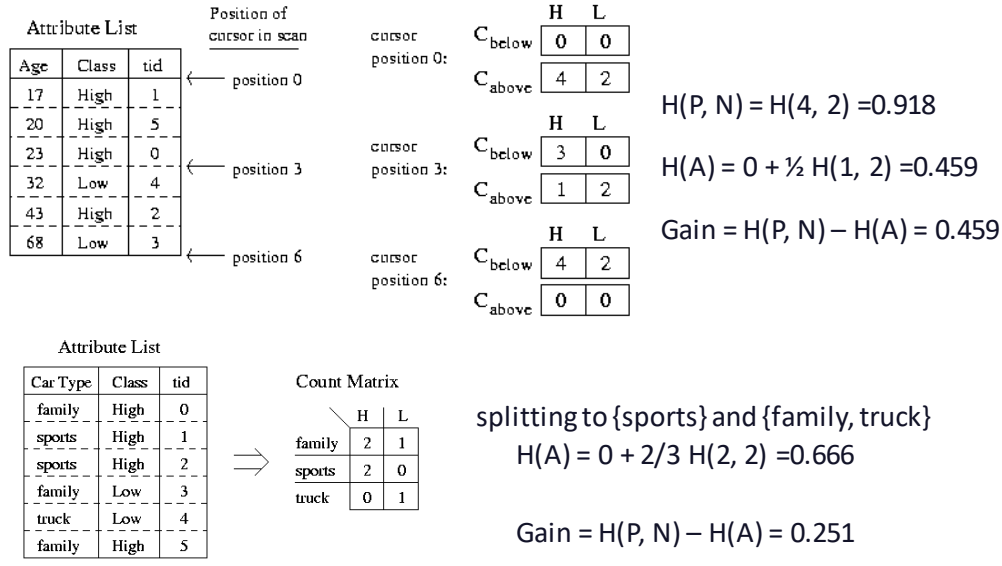
- Sort the data according to attribute value
- Determine the value of X which maximizes information gain by scanning through the data items

Only if the class label changes, a relevant decision point exists



When splitting the dataset using a continuous attribute, we need to determine which is the optimal value to split the dataset based on this attribute. To that end, first the set of attribute values is sorted. Then the class labels are traversed, and whenever it changes a possible split point is found (it can be shown that splitting where class labels do not change is provably sub-optimal). At these points the information gain needs to be computed.

Example



This example illustrates the principle of how splitting is performed both for continuous and categorical attributes.

For reasons we will discuss later, we construct separate attribute lists for each attribute, that is used for classification. The attribute list contains the attribute for which it is constructed (i.e. Age and Car Type), the class label attribute and the transaction identifier tid. The attribute list is sorted for continuous attributes.

Now let us see of how a split point is found for the continuous attribute Age. The distribution of the class attribute for the whole data set (4 High and 2 Low) is stored in variables C_above and a pointer is positioned on top of the attribute list. Then the pointer is moved downwards. Whenever the class label changes the values C_below and C_above are updated (such that they always keep the distribution of H and L values above and below the pointer). At this step the information gain is computed, if the split were performed at that point (in the same way as done for categorical attributes before). After passing through the attribute list the optimal split value for the Age attribute is known.

For the categorical attribute we have to establish a statistics of the distribution of the classes for each of the possible attribute values and store it in a matrix. Then we check the information gain that can be obtained for each of the possible subsets of attribute values and thus determine the optimal "split" for the categorical attribute. Note that this method will be very inefficient with attributes that have large numbers of different values.

Finally, the attribute is chosen that results in the best (binary) split.

Scalability of Continuous Attribute Splits

Naive implementation

- At each step the data set is split in subsets that are associated with a tree node

Problem

- For evaluating which attribute to split, data needs to be sorted according to these attributes
- Becomes dominating cost

In a naïve implementation of the splitting process, we would keep all data in a single table. This would imply that we would have for traversing the attributes in order to resort that table every time an attribute is investigated. Therefore, a more efficient approach is needed.

Scalability of Continuous Attribute Splits

Idea: Presorting of data and maintaining order throughout tree construction

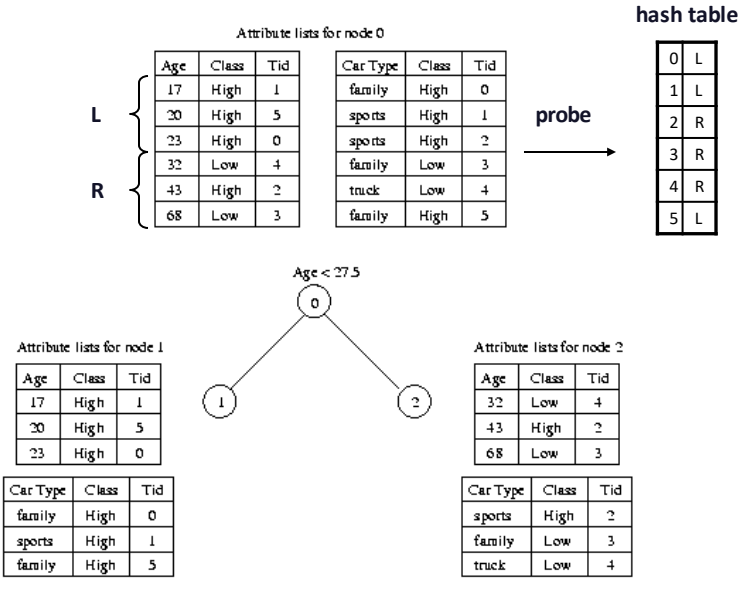
- Requires separate sorted attribute tables for each attribute

Updating attribute tables

- Attribute used for split: splitting attribute table straightforward
- Other attributes
  - Build Hash Table associating tuple identifiers (TIDs) of data items with partitions
  - Select data from other attribute tables by scanning and probing the hash table

To avoid repeated resorting of data, for every attribute a separate and presorted table is kept. Once a split is chosen, we find two different situation. For the table that keeps the attribute that was used in the split, the table needs just to be partitioned into two subtables, maintaining the order. For the other attributes we have to select the subtables corresponding to the instances of the two partitions that have been formed. To that end a temporary hash table is constructed that allows to associate to each dataitem its partition. Then the attribute table is scanned and partitioned using the hashtable to decide for each entry to which partition it belongs. Note that in this approach, for continuous attributes, the resulting tables are again sorted as the order is preserved from the original table.

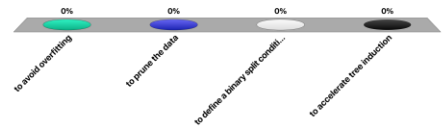
Example



In this example we demonstrate of how attribute tables are split, when a decision node is introduced in the decision tree. Since the split is based on attribute Age, the table for Age can simply be split into two subtables at the threshold value. For the Car Type table we use the temporary hash table indicating partition membership to separate it into two subtables.

When splitting a continuous attribute, its values need to be sorted …

A. to avoid overfitting
B. to prune the data
C. to define a binary split condition
D. to accelerate tree induction



Clustering 64

Characteristics of Decision Tree Induction

Strengths

- Automatic feature selection
- Minimal data preparation
- Non-linear model
- Easy to interpret and explain

Weaknesses

- Sensitive to small perturbation in the data
- Tend to overfit
- No incremental updates

We summarize here some of the major strengths and weaknesses of standard decision tree induction.

Decision trees advantages

- The information theoretic criteria used to select the most discriminative attribute is an embedded feature selection
- No data preparation is needed, such as normalisation of data
- The best aspect of using trees for analytics is that they are easy to interpret and explain, while more sophisticated ML algorithms (ANN, SVM) are seen as black-boxes that do not “explain” the classification decisions they make

Decision trees drawbacks

- They can be extremely sensitive to small perturbations in the data: a slight change can result in a drastically different tree.
- They can easily overfit. This can be compensated by validation methods and pruning, but remains a problem.
- They are not incremental. If new data is available, the existing tree cannot be incrementally modified, but the whole tree must be reconstructed from scratch

Decision Tree Induction: Properties

Model: flow-chart like tree structure

Score function: classification accuracy

Optimisation: top-down tree construction + pruning

Data Management: avoiding sorting during splits

Classification Algorithms

Decision tree induction is a (well-known) example of a classification algorithm

Alternatives

- Basic methods: Naïve Bayes, kNN, logistic regression, ..
- Ensemble methods: random forest, gradient boosting, …
- Support vector machines
- Neural networks: CNN, rNN, LSTM, …

Decision trees is one of the best known and historically first examples of a classification approach. Many other methods have been devised in studied over tie. These include basic methods (we will see some examples later), ensemble methods (discussed in the following), support vector machines (a paradigm based on splitting the space through hyper-planes), and neural networks (which are attracting recently significant attention and are nowadays among the best performing classifiers if very large training sets are available).

Ensemble Methods

Idea

- Take a collection of simple or **weak** learners
- Combine their results to make a single, **strong** learner

Types

- **Bagging:** train learners in parallel on different samples of the data, then combine outputs through voting or averaging
- **Stacking:** combine model outputs using a second-stage learner like linear regression
- **Boosting:** train learners on the filtered output of other learners

One important development in decision trees was the introduction of the idea of ensemble methods. The basic principle is simple: instead of constructing a single model, many different models are constructed independently. Even if each model is not very expressive (weak learners) their combination can be powerful (strong learner). Different ensemble methods are distinguished by the type of approach they are based on. Ensemble methods, which we will discuss in the following, learn several models in parallel, and combine then their predictions by voting or averaging. Stacking methods use more sophisticated techniques to combine model outputs, based themselves on learning methods. Finally, boosting learn models in sequence. In each step the samples of the training data are reweighted depending on whether they have been correctly classified.

Random Forests

Learn K different decision trees from independent samples of the data (bagging)

- vote between different learners, so models should not be too similar

Aggregate output: majority vote

Random forests are an ensemble method based on bagging. The principle is very simple: K different decision trees are learnt in parallel from different (independent) samples of the data, and the classification is derived from a majority vote of the predictions.

Why do Ensemble Methods Work?

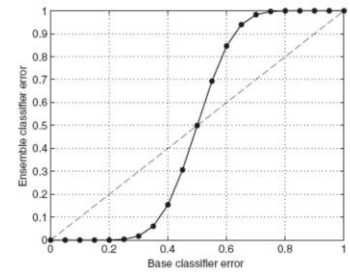
Assume there are 25 base classifiers

– Each classifier has error rate = 0.35

– Assume classifiers are independent

Probability that the ensemble classifier makes a wrong prediction

$$P(\text{wrong prediction}) = \sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$



Tan, Steinbach, Kumar

Here we give the argument why ensemble methods work: even if the individual classifiers are not very good (e.g. make 35% errors in prediction) their aggregate will be very strong. For example, if we have 25 classifiers, the probability that a majority of them, namely at least 13, make a wrong prediction is very small, namely 6%. In general, ensemble methods work well, if the individual models are better than random guessing. The figure illustrates the relation between the classification errors of individual classifiers and the aggregate classification accuracy.

Sampling Strategies

Two sampling strategies

Sampling data

- select a subset of the data → Each tree is trained on different data

Sampling attributes

- select a subset of attributes → corresponding nodes in different trees (usually) don't use the same feature to split

For random forests the main issue is the choice of the sampling strategy, i.e., the generation of samples that are used for learning the individual models. Specifically, it consists of two different sampling strategy. The first, sampling data selects from the original dataset a sample. Thus each decision tree is trained on a different sample of data. The second, sampling attributes selects from the attributes available to take a decision a random subset. Thus even if a tree would have been constructed in the same way up to a level, the continuation might become different due to attribute sampling (e.g. the optimal attribute in one tree is not available for splitting in the other tree).

Random Forests: Algorithm

1. Draw K bootstrap **samples of size N** from original dataset, with replacement (bootstrapping)
2. While constructing the decision tree, select a random set of **m attributes** out of the p attributes available to infer split (feature bagging)

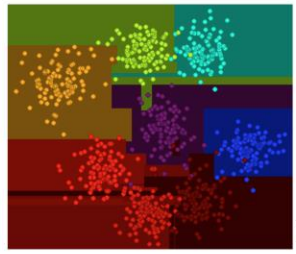
Typical parameters

- m ≈ sqrt(p), or smaller
- K ≈ 500

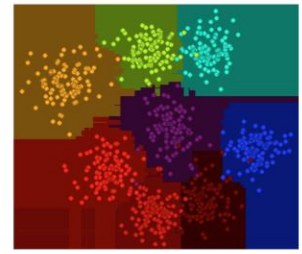
The random forest algorithm is based on specific choices of sampling strategies, when constructing multiple decision trees in parallel. For sampling the data, for each of the K trees to be constructed, a sample of size N (the original size of the dataset) is selected with replacement. This step is called bootstrapping. Note, since sampling is done with replacement the same object might occur repeatedly in a sample. Thus even if the sample has the same size as the original datasets, not all original data instances will occur in the sample. For selecting the attributes a proper subset of attributes is chosen to infer the next split. The number of attributes considered, is significantly smaller than the whole set, e.g. in the order of the square root of the number of all attributes.

Bootstrapping has been original conceived, for training datasets that are not exceedingly large, and thus available data would have to be used very carefully. In cases were training data is available abundantly, an alternative approach, called su-bagging, can be used. There sampling is done without replacement.

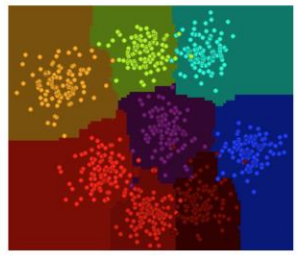
Illustration of Random Forests



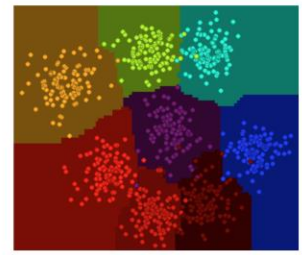
1 rCART



10 rCARTs



100 rCARTs

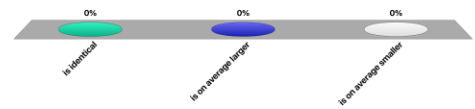


500 rCARTs

Random forests allow to learn much more complex functions than basic decision trees. This fact is illustrated in this visualization. For the same training data set different numbers of decision trees are constructed (rCART is a variant of decision trees). We observe that with increasing numbers of trees the decision boundaries become increasingly more complex and smoother, and thus a better separation among different classes can be achieved.

The computational cost for constructing a RF with K as compared to constructing K decision trees on the same data

A. is identical
B. is on average larger
C. is on average smaller



Characteristics of Random Forests

Strengths

- Ensembles can model extremely complex decision boundaries without overfitting
- Probably the most popular classifier for **dense data** (<= a few thousand features)
- Easy to implement (train a lot of trees)
- Parallelizes easily, good match for MapReduce

Random forests are a very popular method for classification due to the many advantages they offer. They are considered as the method of choice in cases where the data is dense, which means that the number of features is relatively low (in the thousands). Sparse data would be, for example, vector space representation of documents with very large vocabularies. In such cases, before applying a method such as random forests, a dimensionality reduction would have to be applied. This could be accomplished in the case of documents by creating a word embedding.

Characteristics of Random Forests

Weaknesses

- Deep Neural Networks generally do better
- Needs many passes over the data – at least the max depth of the trees
- Relatively easy to overfit – hard to balance accuracy/fit tradeoff

More recently, in cases where large training sets are available or number of features is very large, deep neural networks exhibit better performance than random forests.

References

Textbook

- Jiawei Han, Data Mining: concepts and techniques, Morgan Kaufman, 2000, ISBN 1-55860-489-8

References

- Leo Breiman (2001) “Random Forests” Machine Learning, 45, 5-32.
- Shafer, John, Rakesh Agrawal, and Manish Mehta. "SPRINT: A scalable parallel classifier for data mining." *Proc. 1996 Int. Conf. Very Large Data Bases*. 1996.