Ordered Sets for Data Analysis: An Algorithmic Approach

Lecture 1
Overview. Relations, Their Matrices and Graphs

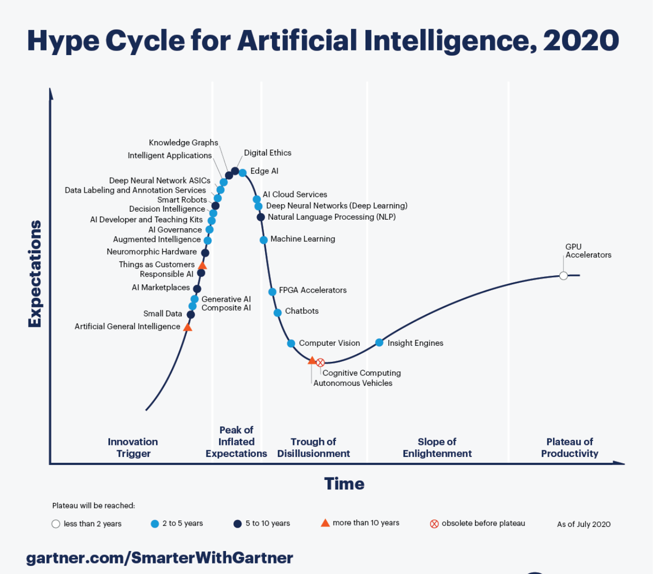
Sergei O. Kuznetsov

Master Program in Data Science
National Research University Higher School of Economics, Moscow, Russia

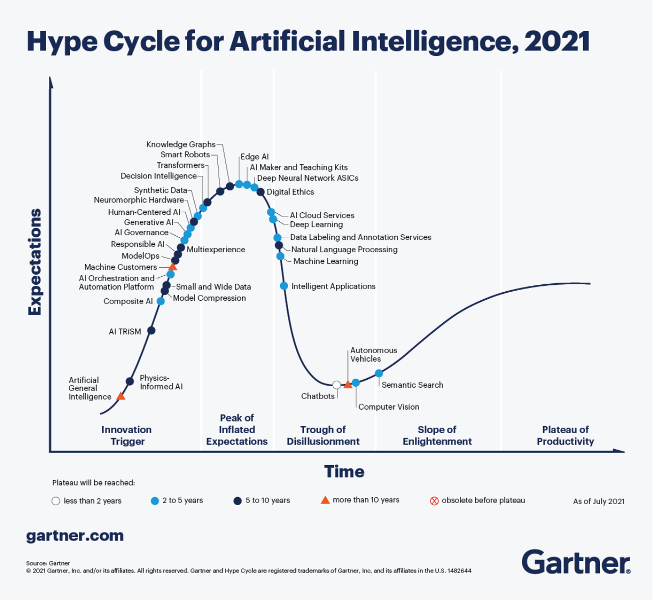
September 13, 2022

AI hype curve and interpretable AI

Gartner hype curve describes standard dynamics of expectations related to a new technology



AI hype curve 2021



Interpretable and Explainable AI. A Guide to AI research

C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019. **More than 1500 citations in google.scholar**

A new manual on AI: A Guided Tour of AI Research (3 volumes), Springer, 2020.

Outline

The main goal of the course: Getting acquainted with interpretable approaches to knowledge discovery based on order and lattice theory.

1. Relations, binary relations, their matrices and graphs. Operations over relations, their properties, and types of relations.
   Main topics: relations, graphs, their properties and operations over them
2. Introduction to the theory of algorithmic complexity
   Main topics: Worst-case complexity, classes NP, co-NP, #P. Delay, total complexity, amortized complexity.
3. Partial order and its diagram
   Main topics: graphs and diagrams of partial order, topological sorting, Galois connection between two partial orders
4. Semilattices and lattices
   Main topics: infimums, supremums, classes of lattices

Outline

5. Introduction to Formal Concept Analysis (FCA)
   Main topics: (formal) concept, concept extent, concept intent, order on concepts, concept lattice
6. Implications and functional dependencies
   Main topics: implication bases, implicational closure, Attribute Exploration
7. Association rules
   Main topics: confidence, support of association rules, concept lattice as a concise representation of association rules
8. Algorithmic problems of computing concept lattices and implication bases
   Main topics: intractability of generation of concept lattices and implication bases, efficient algorithms

The course: important information

1. 6 lectures $\times$ 4 hours (almost) every next Tuesday;
   September 13, 27
   October 11
   November 1, 15, 29
2. practical exercise with
   Egor Dudyrev, Eric George Parakal, Abdulrahim Ghazal
   September 20
   October 4, 18
   November 8, 22
   December 6
3. regular homework (twice a month);
4. big home work: designing an interpretable classification method (the work starting from October 1) or
   making a survey on a new research area in interpretable AI (presentation starting from November 1)
5. manual S.O. Kuznetsov, "Ordered Sets in Data Analysis. Part 1"

PRELIMINARIES

Binary relations

Cartesian (direct) product of sets A and B: the set of ordered pairs, the first element of which belongs to A and the second element belongs to B:

$$A \times B := \{(a, b) \mid a \in A, b \in B\}.$$

Binary relation R from set A in set B is a subset of Cartesian product of sets A and B: $R \subseteq A \times B$.
Infix form for denoting relation R: $aRb \Leftrightarrow (a, b) \in R \subseteq A \times B$.
If $A = B$, then R is **relation on set** A.
Identity relation $I := \{(a, a) \mid a \in A)\}$.
Universal relation $U := \{(a, b) \mid a \in A, b \in A)\}$.

Matrix representation of a binary relation

Matrix of a binary relation

Let $R \subseteq A \times A$. Relation R is represented in the form of the matrix

$$\begin{array}{c|ccc} R & \cdots & a_j & \cdots \\ \hline \vdots & & \vdots & \\ a_i & \cdots & \varepsilon_{ij} & \\ \vdots & & & \end{array} \qquad \varepsilon_{ij} = \begin{cases} 1, & \text{if } (a_i, a_j) \in R \\ 0, & \text{if } (a_i, a_j) \notin R \end{cases}$$

Derivatives of a relation $R \subseteq A \times A$

dual relation
$R^d := \{(a, b) \mid (b, a) \in R\}$

complement
$R^c = \overline{R} := \{(a, b) \mid (a, b) \notin R\}$

incomparability relation
$I_R = A \times A \setminus (R \cup R^d) = (R \cup R^d)^c = R^c \cap R^{cd}$

product of relations
$P \cdot R = \{(x, y) \mid \exists z \quad ((x, z) \in P \text{ and } (z, y) \in R)\}$

degree of relation $R^n = \underbrace{R \cdot R \cdot \ldots \cdot R}_{n \text{ times}}$

transitive closure of relation
$R^T = R \cup R^2 \cup R^3 \cup \ldots = \bigcup_{i=1}^{\infty} R^i$

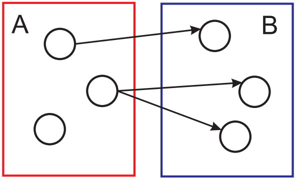
Functions

Relation $f \subseteq A \times B$ is a **function** from A to B (denoted by $f : A \to B$) if for every $a \in A$ there is $b \in B$ such that $(a, b) \in f$ and

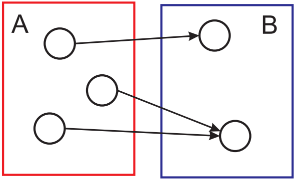
$$(a, b) \in f, (a, c) \in f \Rightarrow b = c$$

A function $f : A \to B$ is
injective if $b = f(a_1)$ and $b = f(a_2) \Rightarrow a_1 = a_2$;
surjective if for every b from B there is a from A such that $b = f(a)$ (or $\forall b \in B\ \exists a \in A\ b = f(a)$);
bijection if it is surjective and injective.

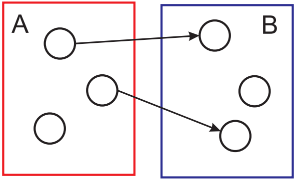
Function types



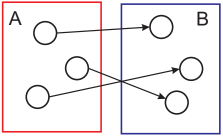
relation, not a function



surjective, not injective



injective, not surjective



bijective

Properties of binary relations

Let $R \subset A \times A$, then R is called

reflexive if $\forall a \in A \; aRa$
antireflexive if $\forall a \in A \quad \neg(aRa) \; (\Leftrightarrow aR^c a)$
symmetric if $\forall a, b \in A \quad aRb \Rightarrow bRa$
asymmetric if $\forall a, b \in A \quad aRb \Rightarrow \neg(bRa) \; (\Leftrightarrow bR^c a)$
antisymmetric if $\forall a, b \in A \quad aRb \; \& \; bRa \Rightarrow a = b$
transitive if $\forall a, b, c \in A \quad aRb \; \& \; bRc \Rightarrow aRc$
complete or linear if $\forall a, b \in A \quad a \neq b \Rightarrow aRb \vee bRa$.

Types of binary relations

- **Tolerance** is a reflexive and symmetric binary relation;
- **Equivalence** is a reflexive, symmetric, and transitive binary relation;
- **Quasi-order** or **preorder** is a reflexive and transitive binary relation;
- **Partial order** is a reflexive, transitive, and antisymmetric binary relation;
- **Strict order** is antireflexive and transitive binary relation.

Representing relations with graphs

Directed graph G is a pair of the form (V, A), where V is called the set of graph **vertices**and $A \subseteq V \times V$ is called the set of graph **arcs**.
Directed graphs with the set of vertices V represent relations on set V.

Representing relations with graphs

Undirected graph is a pair of the form $G = (V, E)$. The set V is called set of **vertices** of the graph. The elements of the set $E = \{\{v, u\} \mid v, u \in V\} \cup E_0$, where $E = \{\{v, u\} \mid v, u \in V\}$ is the set of unordered pairs of elements of V, is called an **edge**, and elements of $E_0 \subseteq V$ are called **loops**. If $E_0 = \emptyset$, then G is called **loopless**.
Undirected graphs with vertex set V represent symmetric relations on set V, i.e., $R \subseteq V \times V$: $(a, b) \in R \Leftrightarrow (b, a) \in R$.

Matrix graph representation

(Vertex-vertex) adjacency matrix
(Directed) graph $G = (V, E)$ can be represented

$$\begin{array}{c|ccc} & \cdots & v_j & \cdots \\ \hline \vdots & & \vdots & \\ v_i & \cdots & \varepsilon_{ij} & \\ \vdots & & & \end{array}$$

$$\varepsilon_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{if } (v_i, v_j) \notin E \end{cases}$$

In undirected graph $\varepsilon_{ij} = \varepsilon_{ji}$

Matrix graph representation

(Vertex-edge) incidence matrix

(Directed) graph $G = (V, E)$ can be represented by a matrix

$$\begin{array}{c|ccc} & \cdots & e_j & \cdots \\ \hline \vdots & & \vdots & \\ v_i & \cdots & \varepsilon_{ij} & \\ \vdots & & & \end{array} \qquad \varepsilon_{ij} = \begin{cases} -1, \text{ if } \exists v_k \in V : e_j = (v_i, v_k) \\ 1, \text{ if } \exists v_k \in V : e_j = (v_k, v_i) \\ 0, \text{ if } v_i \notin e_j \end{cases}$$

For undirected graph

$$\varepsilon_{ij} = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{if } v_i \notin e_j \end{cases}$$

Vertex v_i is incident to arc (edge) e_j if $\varepsilon_{ij} \neq 0$.

Representing relation with graphs

Directed bipartite graph is a pair of the form (V, A), where $V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$, and $A \subseteq V_1 \times V_2$, i.e., any arc from A connects vertex from V_1 c with vertex from V_2. Seta V_1 and V_2 are called **parts** of the graph.

Directed bipartite graphs on parts V_1 and V_2 represent binary relations from V_1 to V_2.

Representing relations with graphs

(Undirected) bipartite graph is a pair of the form (V, E), where V is the set of vertices, $V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$, and $E = \{\{v_1, v_2\} \mid v_1 \in V_1, v_2 \in V_2\}$ is the set of undirected edges connecting vertices from V_1 with vertices from V_2.

Undirected bipartite graphs on parts V_1, V_2 represent symmetric binary relations from V_1 to V_2.

Undirected **complete graph** is a graph $G = (V, E)$, where each pair of vertices is connected by an edge.
Complete graph with the set of vertices V is a universal relation on set V.

Tolerance classes

Let set M and relation of tolerance on $M \times M$ be given.

Tolerance class is a maximal subset of elements M, all pairs of which belong to relation.

Graph interpretation of a tolerance class is a maximal complete subgraph (clique) of a undirected graph with loops.

Equivalence classes and partitions

Equivalence class is a maximal subset of M, each element of which is equivalent to some $x \in M$.

Graph interpretation of equivalence class is a connected component of a graph.

Partition of set M is a family of sets $\{M_1, \ldots, M_n\}$, such that

$$\bigcup_{i \in [1,n]} M_i = M, \quad \forall i, j \in [1, n] \quad M_i \cap M_j = \emptyset.$$

There is a bijection between partitions and equivalences on set M.

Parts of (undirected) graphs

Graph $H = (V_H, E_H)$ is a **part** (subgraph) of graph $G = (V_G, E_G)$ if all vertices and edges of H are vertices and edges of G, i.e. $V_H \subseteq V_G$ and $E_H \subseteq E_G$.

Graph $H = (V_H, E_H)$ is an **(induced) subgraph** of graph $G = (V_G, E_G)$ if H is a part (subgraph) of G, and edges of H are all edges of G, both vertices of which lie in H.

Paths and connectivity in undirected graphs

route (path) is an alternating sequence of graph vertices and edges $v_0, e_1, v_1, e_2, v_2, \ldots, e_k, v_k$, where each of two neighboring edges have a common vertex.
chain is a route where all edges are different.
simple chain is a route where all vertices (an hence, edges) are different.
Two vertices are called **connected** if there is a simple chain joining them.
A graph is **connected** if all pairs of vertices are connected.
connected component of a graph is maximal (by containment order) subset of graph vertices, each pair of which is connected.
A connected component is an equivalence class on the set of vertices wrt. relation "be connected".

Paths and connectivity in directed graphs

A **(directed) route** or **(directed) path** is an alternating sequence of graph vertices and arcs of the form $v_0, a_1, v_1, a_2, v_2, \ldots, a_k, v_k$, where the end of any arc (may be, except for the last one) coincides with the beginning of the next arc, i.e. $a_i = (v_{i-1}, v_i)$.
chain is a route where all arcs are different.
simple chain is a route where all vertices (and hence, arcs) are different.
Vertex v_j is **reachable** from vertex v_i if there is a route that starts in v_i and ends in v_j (it is considered that a vertex is reachable from itself).
A graph is **strongly connected** if any two vertices are reachable from each other.
A graph is **one-way** or **weakly connected** if for any pair of vertices at least one of them is reachable from the other.

Cycles

(oriented) cycle is an (oriented) route where the first and the last vertices are the same.
simple cycle is a cycle where every vertex occurs at most once.
undirected cycle (contour) in directed graph is a sequence of vertices and arcs of a directed graph which can be transformed in a directed cycle by changing orientations of some arcs.
directed acyclic graph is a directed graph with no directed cycles.

Quasiorders, partial orders, and strong orders are naturally represented by acyclic graphs (loops are usually not drawn).

Trees

(Undirected) tree is an undirected acyclic graph.
Directed tree is a directed graph with no cycles, either directed or undirected.
Rooted tree is a tree with a distinguished vertex called **root**.
The root of a rooted tree defines the direction from the root to leaves (vertices with one neighbor), so rooted trees may be considered as directed.
Star is a tree of the form $G = (V, E)$, where $V = \{v_0\} \cup V_1$, $E = \{\{v_0, v_i\} \mid v_i \in V_1\}$.

Partial order

(Partial) order is a reflexive, transitive, and antisymmetric binary relation.

(Partially) ordered set $(P, \leq)$ is a set P with a (partial) order $\leq$ on it.
Ordered sets are naturally represented by acyclic directed graphs.

Strict order $<$ related to partial order $\leq$:

$$x < y := x \leq y \text{ and } x \neq y$$

is an antireflexive asymmetric transitive relation

Example. Order on partitions

Partition of a set S is a family of it subsets (called **blocks** of partition) $\{S_1, \dots, S_n\}$, such that

$$\bigcup_{i \in [1,n]} S_i = S, \quad \forall i, j \in [1, n] \quad S_i \cap S_j = \emptyset.$$

Partitions are denoted by $S_1 \mid S_2 \mid \dots \mid S_n$.

Partition P_1 is **finer** than partition P_2 (equivalently, partition P_2 is **rougher** than partition P_1) or $P_1 \leq P_2$ if for any block of partition P_1 there is a block of partition P_2 containing it.

Proposition. Relation $\leq$ on partitions is an order.

Example. For $S = \{a, b, c, d\}$ one has
$\{a, b\} \mid \{c\} \mid \{d\} \leq \{a, b, c\} \mid \{d\}$.

Example. Partial order on multisets

Multiset on a set S is set S equipped with the function $r\colon S \to N \cup \{0\}$ giving multiplicity of elements of S.
Multiset M on S is usually denoted by $\{a_{m_a} \mid a \in M\}$, where m_a is multiplicity of element a.
Multiset $L = \{a_{l_a} \mid a \in L\}$ is a subset of multiset $M = \{a_{m_a} \mid a \in M\}$ ($L \subseteq M$), if for all a one has $l_a \leq m_a$.

Proposition. Relation $\subseteq$ on multisets on a set is an order.

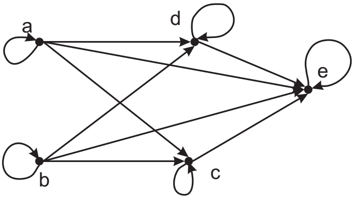
Example. For $S = \{a, b, c, d\}$ one has $\{a_1, b_5, c_2\} \subseteq \{a_3, b_6, c_2, d_2\}$.

Example. Order relation

	a	b	c	d	e
a	1	0	1	1	1
b	0	1	1	1	1
c	0	0	1	0	1
d	0	0	0	1	1
e	0	0	0	0	1

Example. Graph of an ordered set

	a	b	c	d	e
a	1	0	1	1	1
b	0	1	1	1	1
c	0	0	1	0	1
d	0	0	0	1	1
e	0	0	0	0	1



acyclic graph

Linear order

Linear or **complete** order is an order with linearity (completeness) property:

$$\text{for any } x, y \quad \text{either } x \leq y \text{ or } y \leq x.$$

A linearly ordered set is also called a **chain**. Indeed, a linearly ordered set correspond to a chain in the acyclic order graph.

Strict partial order that corresponds to the linear order is called a strict order.

Example. Lexicographic order

Let A be a finite set of characters (alphabet) linearly ordered by relation $\prec$. A word in alphabet A is a finite (may be empty) sequence of charactersfrom A. The set of all words is denoted by A^*.

Lexicographic order $<$ on words from A^* is defined in the following way:

$w_1 < w_2$ for $w_1, w_2 \in A^*$ if

either w_1 is a prefix of w_2 (i.e., $w_2 = w_1 v$, where $v \in A^*$), or its first character which is different for w_1 and w_2 is less in w_1 than w_2 wrt. linear order $\prec$ (i.e., $w_2 = wav_1$, $w_1 = wbv_2$, where $w, v_1, v_2 \in A^*$, $a, b \in A$, $a \prec b$).

Proposition. Lexicographic order on the set of words A^* is a strict linear order.

Exercises

1. Determine the properties of the relation $Q := \{(m, n) \mid m, n \in \mathbb{N} \;\&\; m = n^2\}$
2. What is the number of injections from a finite set A to a finite set B, where $|A| < |B|$.
3. Find the number of parts of a finite graph with the set of edges E.
4. Degree of a vertex of an (undirected) graph is the number of edges incident to the vertex. Prove that in an arbitrary graph the number of vertices with odd degree is even.
5. Prove that the incomparability relation for an order is a tolerance relation
6. Every subset of an ordered set is an ordered set (w.r.t. the restriction of the order relation on the subset)
7. Is a strict order antisymmetric?
8. Prove that for an order relation P one has $(P^c)^d = (P^d)^c$.
9. Prove that the dual relation to an order is an order relation.
10. Prove that intersection of orders on a set is an order relation
11. Prove that a lexicographic order is linear.

Literature on relations, orders and graphs

G. Birkhoff and T.C. Bartee *Modern Applied Algebra*, McGraw-Hill, N.Y., 1970. (Chapter 2)

J.A. Bondy, U.S.R. Murty, Graph Theory, Springer, 2008.

F. Harary, Graph Theory, Reading, MA: Addison-Wesley, 1969.

B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, Cambridge University Press, 1990.