KMI/ZZD – Získávání znalostí z dat

Úvod, motivace, modely KPD, úlohy DM

Jan Konečný

17. února 2015

Rozvrh a sylabus

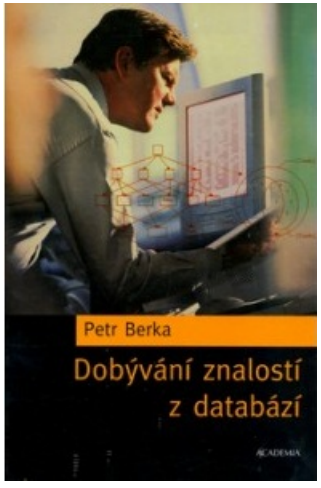
`http://phoenix.inf.upol.cz/~konecnyj/vyuka/ZZD.html`

Rozvrh:
Úterý: 8:00–10:15 (v tom je 1×90 min. přednášky, 1×45 min. cv.)

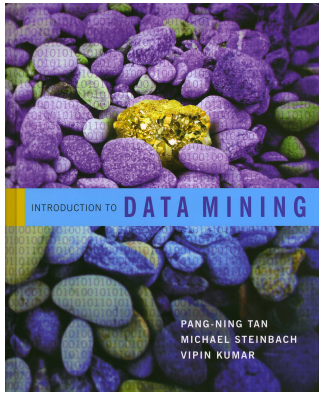
Sylabus Úvod, data mining, data, základní pojmy, reprezentace znalostí. Redukce dimenzionality: principal component analysis, independent component analysis; metody diskretizace dat. Učení bez učitele: shlukování a asociační pravidla. Učení s učitelem: Rozhodovací stromy, pravidlové algoritmy, hybridní algoritmy. Neuronové sítě. Metoda GUHA, další okrajová témata.

Zápočet & zkouška obojí za úkoly zadávané v průběhu semestru.

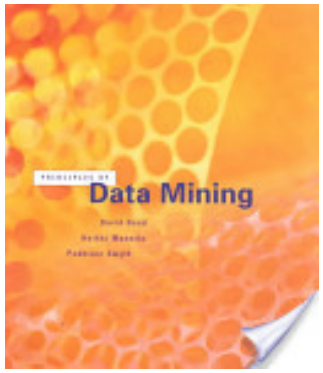
Doporučená literatura



Berka P.
Dobývání znalostí z databází.
Academia, Praha, 2003.



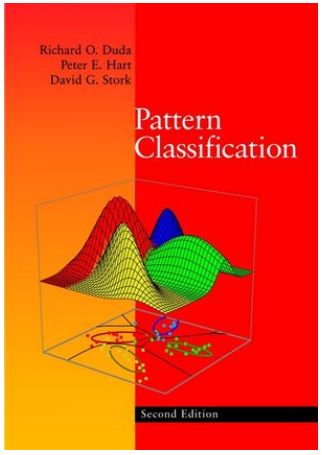
Tan P., Steinbach M., Kumar V.
Introduction to Data Mining.
Addison Wesley, 2005.



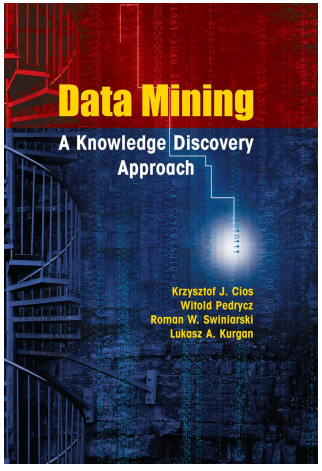
Hand D. J., Mannila H., Smyth P.
Principles of Data Mining.
MIT Press, Cambridge, MA, 2001.

- – absenční výpujčka v knihovně PřF,
- – prezenční výpujčka v knihovně PřF,
- z , z – výpujčka v knihovně Zbrojnice, ftk , ftk – výpujčka v knihovně FTK, . . .
- w – k dispozici na webu,
- – není k dispozici v knihovnách UP.

Další doporučená literatura



Duda R. O., Hart P. E., Stork D. G.,
Pattern Classification,
Wiley, 2000



Cios K. J., Pedrycz W., Swiniarski R. W., Kurgan L. A.,
Data Mining: A Knowledge Discovery Approach,
Springer, 2007

Co je Data Mining (DM)?

„Cíl DM je dát smysl velkému množství z velké části neoznačených dat v nějaké doméně .“

dát smysl – najít novou znalost, která má být:

- *pochopitelná* – pro uživatele/vlastníka dat, který z nich chce mít užitek. Nejvhodnější bude znalost nebo model dat, který může být popsán snadno-pochopitelnými pojmy.
  Např. pravidly jako:

  IF *abnormalita (obstrukce) věnčitých tepen* **THEN** *ischemická srdeční choroba*

- *platná* – jasné;
- *nová* – nalezení obecně známých nebo triviálních věcí by bylo uživatelem/vlastníkem dat považováno za selhání.
- *užitečná*.

Co je Data Mining (DM)?

velkému množství dat – DM není o zkoumání malých datasetů; ty mohou být zkoumány běžnými technikami, nebo dokonce manuálně.

- AT&T – obslouží 300 milionů hovorů denně, obslouží 100 milionů zákazníků děnně, uloží asi terabyte dat denně.
- Wal-Mart – 21 milionů transakcí denně, zhruba 12 terabytů
- NASA – generuje několik gigabajtů za hodinu (Earth Observing System)
- Moderní biologie – data pro lidský genom jsou v řádu terabytů/pentabytů.

Taková data nemůžou být přímo zkoumána algoritmicky, natož tak manuálně. Je potřeba provést redukci dat – kvantity i dimenze.

z velké části neoznačených dat – Je levnější a jednodušší sbírat neoznačená data. Označená data musí mít známe vstupy asociované se známými výstupy.

Např. **vstup:** obraz srdce a okolních žil, **výstup:** diagnóza.
(musí určovat kardiolog – náročný a na chyby náchylný proces)

Když jsou sbírána jenom neoznačená data: Potřebujeme algoritmy, které jsou schopny najít „přirozené“ seskupení/shluky, vztahy a asociace v datech.

Pokud bychom našli shluky, může je pojmenovat doménový expert. Tím se z neoznačených dat stanou označená (jednoduchým procesem).

Hledání shluků, vztahů a asociace je otevřený vědecký problém. Současné algoritmy mají pořád ještě vady:

- Shlukování (clustering) – nutno předem specifikovat počet shluků.
- Dolování asociačních pravidel – nutno předem zadat číselné parametry k vygenerování vhodně velké množiny asociací.

pokud jsou sbíraná data z části označená: tj. pár vstupů s výstupy a k tomu velké množství neoznačených dat. Existují techniky (*semi-supervised learning*), které může pro označené vstupy využít.

pokud jsou sbíraná data zcela označená: většina DM technik s nimi pracuje velice dobře (možná až na škálovatelnost).

. . . dat v nějaké doméně – Úspěch DM projektu velmi závisí na přístupu ke znalosti domény. Ti, kdo dělají DM musí úzce spolupracovat s doménovými experty/vlastníky dat.

Objevení nových znalostí z dat je interaktivní a iterativní proces.

Nemůžeme jednoduše vzít DM systém vybudovaný pro nějakou doménu, aplikovat ji na jinou a očekávat dobré výsledky.

Co je DM?

DM vzniklo jako odpověď na technologický pokrok v mnoha různých disciplínách.

počítačové inženýrství $\rightarrow$ silnější počítače (rychlost, pamět)
informatika + matematika $\rightarrow$ efektivnější architektura databází a vyhledávací algoritmy
kombinace obojího $\rightarrow$ vylepšení technik pro sběr, ukládání a přenos velkých objemů dat (pro image processing, digital signal processing, text processing,...)

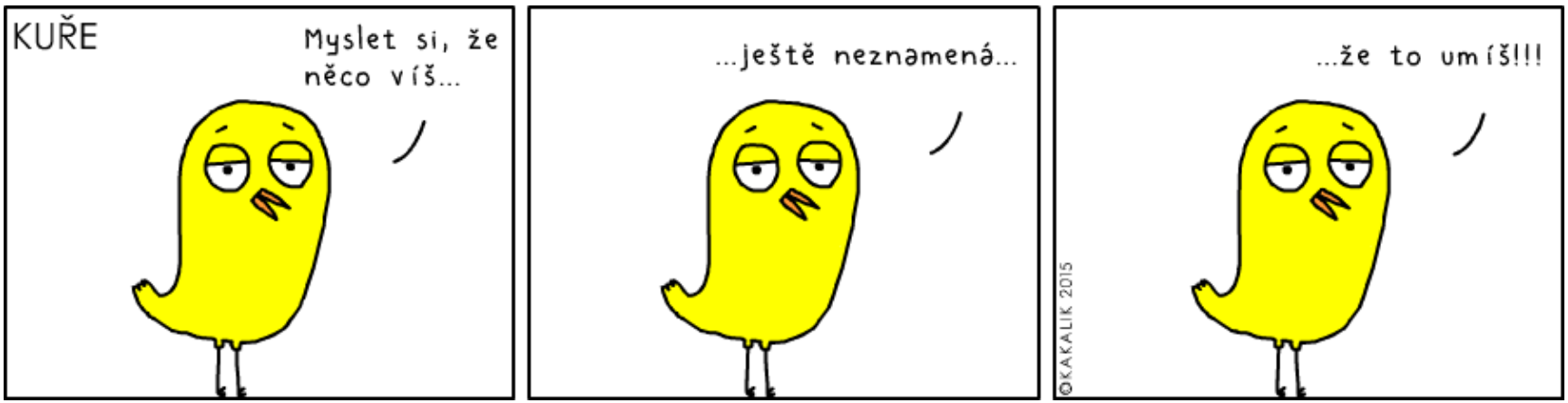
Potřeba lepších, rychlejších a levnějších způsobů, jak s těmi daty zacházet:

> *Data jsou k ničemu bez mechanismů, které z nich efektivně extrahují znalosti.*

První DM průkopníci: U. Fayyad, H. Mannila, G. Piatetsky-Shapiro, G. Djorgovski, W. Frawley, P. Smith a další.

Proces získávaní znalosti z dat (Knowledge discovery process)

- před pokusem extrahovat užitečné znalosti z dat, je důležité pochopit celkový přístup.
- to, že zám mnoho algoritmů pro analýzu dat, není dostatečné pro úspěšný projekt DM.



teď se budeme zabývat **procesem**, který vede k objevení nových znalostí (KDP).

KDP definuje sekvenci kroků (s možnými cykly), které by měly být následovány k objevení znalosti v datech.

Každý krok je obvykle realizován pomocí softwarových nástrojů.

Pojem **model procesu**

- pomáhá organizacím lépe porozumět KDP,
- nabízí plán k plánování a vykonání procesu,
- vede k úsporám (času i financí), lepšímu porozumění a přijatelnosti výsledků.

Knowledge discovery process

Je několik důvodů proč strukturovat KDP jako **standardizovaný model procesu**:

- *Koncový produkt musí být užitečný pro uživatele/vlastníka dat.*
  Nestrukturované používání technik DM „na slepo“ (tzv. bagrování dat; *data dredging*) často vyprodukuje výsledky, které sice můžou být zajímavé, ale nepřispívají k řešení uživatelova problému.
  Skrze dobře definované KDP modely se dosáhne výsledků, které jsou nové, platné, pochopitelné a užitečné.
- *Dobře definovaný KDP model by měl být logický, soudržný, mít dobře promyšlenou strukturu a přístup, který může být prezentován zákazníkovi .*
  Smrtelníci mnohdy nechápou potenciál znalostí skrytých ve velkých datech a nechtějí věnovat čas a prostředky k jejich extrakci.
  Místo toho se spoléhají schopnosti a znalosti jiných (např. doménových expertů).
  Je potřeba jim představit dobrý KPD model.

- *Projekty KD vyžadují značné úsilí v project managementu, které potřebuje být zakotvený v pevném frameworku.*
  Většina KD projektů zahrnuje týmovou práci, plánovaní a rozvrhování.
  Specialisté na project management obvykle nejsou obeznámeni s pojmy DM a KPD – model KPD jim má pomoct vytvořit vhodný plán.
- *KD by mělo následovat příkladu jiných inženýrských disciplín, které už mají stanovené modely.*
  Dobrý příklad je softwarové inženýrství – relativně nová disciplína, která má hodně společného s KD.
  Softwarové inženýrství adoptovalo několik vývojových modelů (vodopád, spirála, . . .), které se staly známými standardy v této oblasti.

Co je KDP?

Data mining – už jsme popsali. (alternativní názvy jsou: knowledge extraction, information discovery, information harvesting, data archelogy, data pattern processing).

Knowledge discovery process – Netriviální proces identifikace platných, nových, potenciálně užitečných a pochopitelných vzorů v datech. Skládá se z více kroků (jeden z nich je DM).
KD se zabývá celým procesem extrakce znalostí (knowledge extraction), včetně toho, jak jsou data uložena, jak je k nim přistupováno, jak použít efektivní a škálovatelné algoritmy k analýze masivních datasetů, jak interpretovat vizualizovat a výsledky...

Modely KDP

KPD model sestává z kroků, které by měly být následovány, když je spouštěn KDP.
Od 90. let několik bylo vyvinuto různých KDP.
První v akademickém výzkumu, průmysl rychle následoval.
Společné pro všechny modely je:

- skládají z nějakého počtu kroků (liší se počtem a rozsahem). Každý krok je inicializován úspěšným dokončením předchozího kroku. Výstup předchozího kroku slouží jako vstup do nového kroku.
- definice vstupů a výstupů:
  typický vstup zahrnuje data v nějake podobě. typický výstup je vygenerovaná nová znalost – obvykle vyjádřena jako pravidla, vzory, klasifikační modely, asociace, tredny, statistické analýzy, ...

Akademické modely

Snaha zavést KDP model začala v akademickém prostředí.

V pol. 90. let, výzkumníci (akademici) začali definovat několika-krokové procedury, které měly provádět uživatele DM nástrojů světem KD.

Hlavní důraz: nabídnout posloupnost aktivit, které pomohou vykonat KDP v libovolné doméně.

Dva KDP:

- 1996 – 9ti krokový model – Fayyad (a spol.)
- 1998 – 8mi krokový model – Anand a Buchner.

Fayyad (a spol.) KDP model – skládá se z devíti kroků:

1. *Vývoj a pochopení aplikační domény* – zahrnuje získání apriorních znalostí a pochopení cílů koncového uživatele.
2. *Vytvoření cílového datasetu* – výběr podmnožiny proměnných a datapointů, které budou použity pro KD.
3. *Čištění dat a preprocessing* – odstranění odlehlých bodů, vypořádání se s šumem a chybějícími hodnotami.
4. *Redukce a projekce dat* – nalezení použitelných atributů použitím redukce dimenze a transformačních metod.
5. *Výběr DM úlohy* – na základě cílů def. v bodě 1 vybere úlohu DM (jako klasifikace, regrese, shlukování, . . .)
6. *Výběr DM algoritmu* – výběr metod pro hledání vzorů datech a rozhodnutí jaké modely a parametry pro použitou metodu budou vhodné.

7. *Data Mining* – vyprodukovaní vzorů.
8. *Interpretace vytěžených vzorů* – vizualizace vyextrahovaných vzorů a modelů, vizualizace dat na základě vyextrahovaných modelů.
9. *Konsolidace výsledků* – zahrnutí objevené znalosti do systému, dokumentace a reportování interesovaným stranám. Může zahrnovat porovnání a hledání konfliktů s dřívější „znalostí“.

hlavní aplikace: Tento 9ti krokový model byl zahrnut do komerčního KD systému MineSet
(viz Purple Insight Ltd. `http://www.purpleinsight.com`).

Model byl také použit v mnoha různých doménách včetně inženýrství, medicíny, výroby, e-businessu a vývoje softwaru.

Průmyslové modely

Průmyslové modely rychle následovaly akademické snahy.
Dva reprezentanti:

- 5ti krokový model Cabena a spol. (s podporou IBM)
- 6ti krokový model CRISP-DM – (stal se hlavním KDP modelem)

CRISP-DM (**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)

- konec 90.let
- Integral Solutions Ltd., NCR, DaimlerChrysler, OHRA.

Vývoj tohoto procesu měl velkou podporu, též podporováno programem ESPIRIT financovaným EU.

CRISP-DM

Šest kroků CRISP-DM:

Porozumění problematice

Zaměřuje se na pochopení cílů a pořadavků, konvertuje tyto do definice problému DM, navrhuje předběžný plán projektu k dosažení těch cílů.

Tento krok je dále rozložen na několik podkroků:

- stanovení cílů v problémové doméně
- zhodnocení situace
- určení cílů DM
- generování projektového plánu

Data Understanding

Porozumění datům

Začíná inicialním sběrem dat a seznámení se s nimi.
Specifické cíle zahrnují identifikaci problémů v kvalitě dat, počáteční vhled do dat, detekce zajímavých podmnožin dat.

Podkroky

- prvotní sběr dat,
- popis dat,
- průzkum dat,
- ověření kvality dat.

Data preparation

Příprava dat

Zahrnuje činnosti, které vedou k vytvoření datového souboru, který bude zpracováván DM metodami.
Tato data by měla

- obsahovat údaje význačné pro danou úlohu,
- mít podobu, která je vyžadována vlastními DM algoritmy.

Zahrnuje selekci dat, čištění dat, transformaci dat, vytváření dat, integrování dat a formátování dat.
Toto je nejpracnější část celého procesu.

Modeling

Modelování

Nasazení metod DM, výběr vhodné metody, určení parametrů, kombinace výsledků různých metod.
Součástí tohoto kroku je i ověřování nalezených znalostí z pohledu metod dobývání znalostí.
(např. testování klasifikačních znalostí na nezávislých datech).

Evaluation

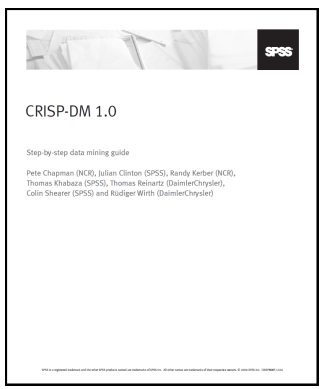
Evaluace – Vyhodnocení výsledných znalostí z pohledu zákazníka – zda byly splněny cíle formulované při zadání.

Deployment

Využití výsledků

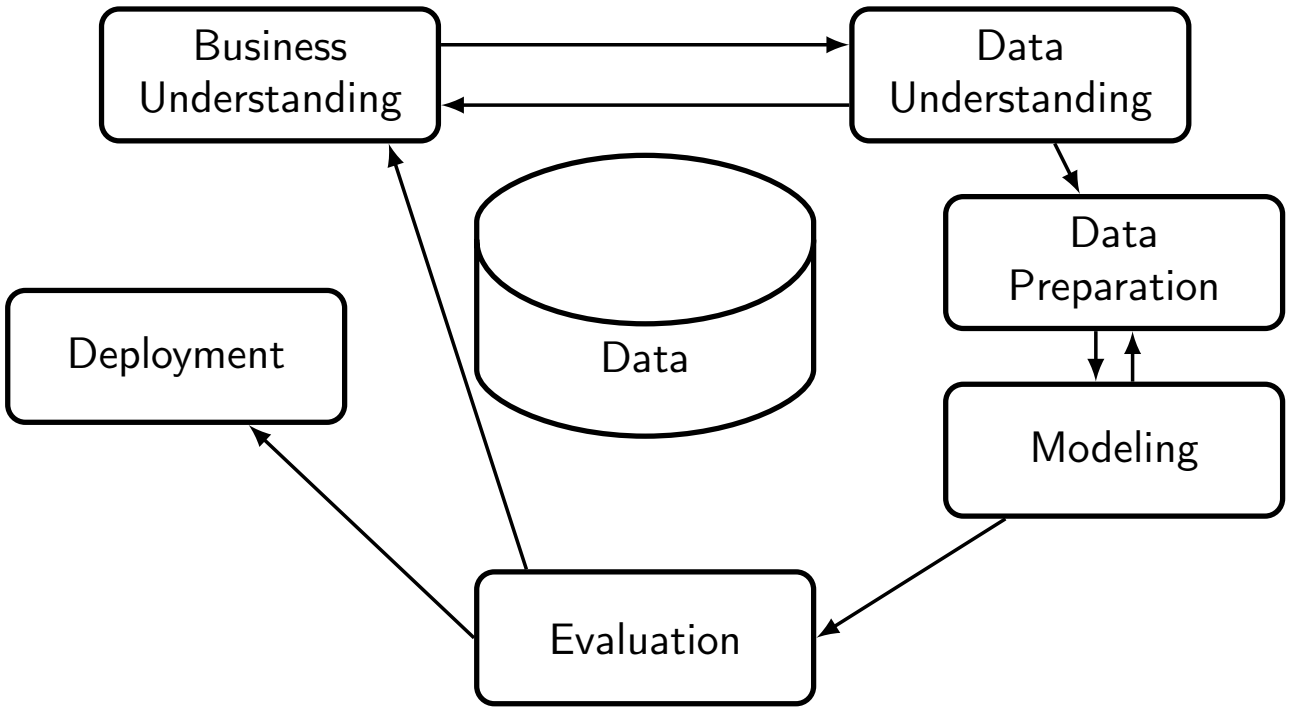
Úprava do podoby, která je přijatelná pro zákazníka.
Aplikace výsledků.

Chcete vědět víc?



Chapman P., Clinton J., Kerber R., Khabaza T., Reinhartz T., Shearer C., Rüdiger W.
CRISP-DM 1.0 – Step-by-step data mining guide
SPSS Inc. 2000
W

Schéma CRISP-DM



Kategorie reprezentace dat

Pravidla
V jejich nejobecnějším formátu, pravidla jsou podmiňovací tvrzení ve tvaru:

IF *podmínka* **THEN** *následek (akce)*,

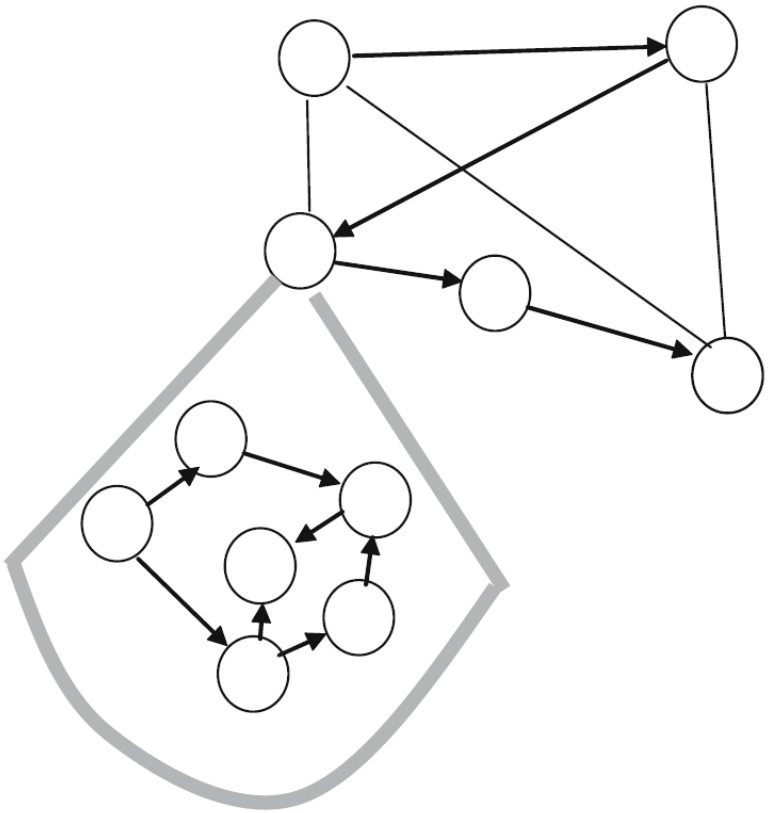
podmínka a *následek* jsou deskriptory kousků znalosti o doméně.
Pravidlo samotné představuje vztah mezi těmito deskriptory.

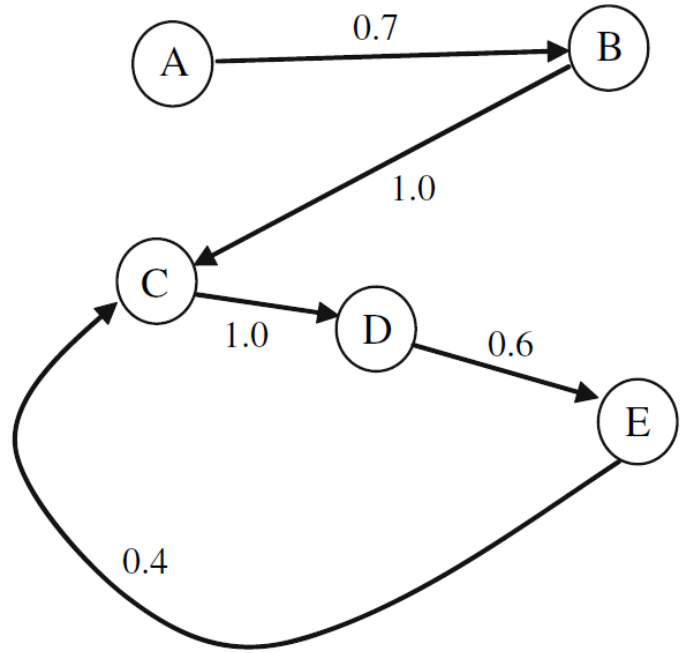
podmínka a *následek* jsou tvořeny inf. granulemi. Operační kontext, ve kterém jsou inf. granule formovány a používány, může být stanoven uvážením dostupného formálního systému: množiny, fuzzy množiny, rough sets, ...
V praxi, doménová znalost je typicky strukurovaná do kolekce pravidel, které mají stejný (nebo podobný) formát

IF *podmínka* je A_i **THEN** *následek* je B_i,

Grafy a orientované grafy

Uzlu představují koncepty nebo atributy, hrany jsou vztahy mezi nimi.





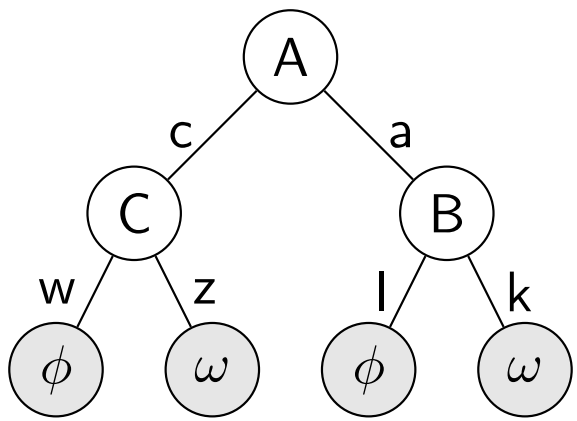
Speciální případ tohoto je konceptuální svaz ve FCA

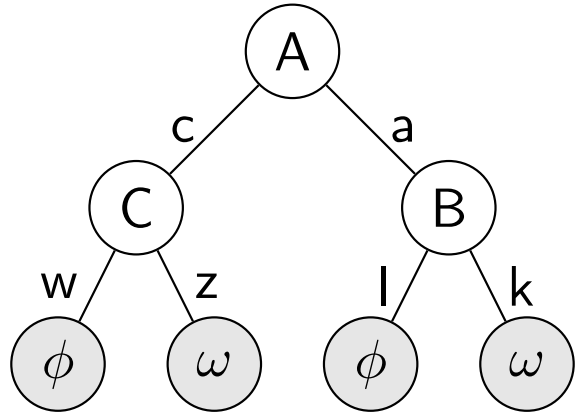
Stromy, rozhodovací stromy

(Zakořeněné stromy) jsou speciální kategorie grafů: mají určený kořen, neobsahují cykly.

Jednou z nejběžněji používaných struktur jsou **rozhodovací stromy**.

- Každý uzel reprezentuje atribut, který nabývá konečného počtu diskrétních hodnot.
- Hrany vycházející z každého uzlu jsou označeny odpovídajícími hodnotami.





Rozhodovací stromy mohou být jednoduše přeloženy do kolekce pravidel:

IF A je c & C je w **THEN** ϕ (1)

IF A je c & C je z **THEN** ω (2)

IF A je a & B je k **THEN** ω (3)

IF A je a & B je l **THEN** ϕ (4)

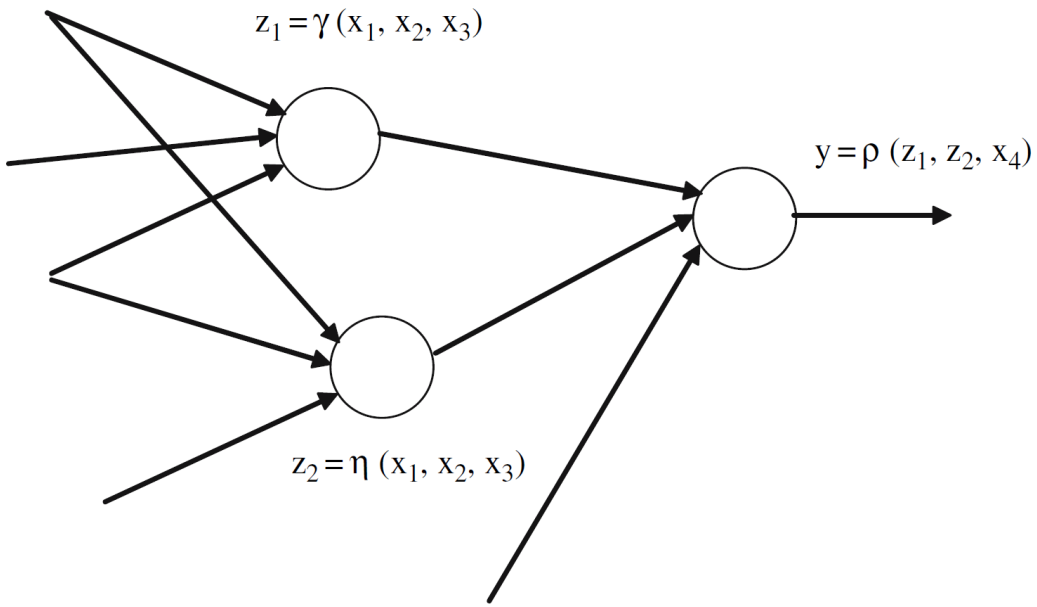
Stromy ale navíc obsahují hierarchii (říkají podle jakých atributů se rozhodovat dříve).

Shluky

Sítě

Mohou být považovány za rozšířené grafy – v tom smyslu, že každý uzel má nějakou lokální výpočetní schopnost.
Představují nejen znalost samotnou ale obsahují taky výpočet.

KMI/UNS – Umělé neuronové sítě



Úlohy DM

Učení bez učitele (unsupervised learning)

Paradigma učení bez učitele zahrnuje proces, který automaticky objeví strukturu v datech a nezahrnuje žádný dohled (supervision).

shlukování (clustering)

Je dán N-dimenzionální dataset,

$$\mathbf{X} = \{x_1, x_2, \ldots, x_N\},$$

kde každé x_k je charakterizováno množinou atributů.
Chceme určit strukturu $\mathbf{X}$, t.j. identifikovat a popsat skupiny (shluky, clusters) v $\mathbf{X}$.

Algoritmus K-means shlukování

Když máme N datapointů v $\mathbb{R}^n$ a předpokládáme, že chceme zformovat c shluků.
Vypočítáme součet rozptylů mezi datapointy a množinou prototypů $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_c$:

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik} \|\vec{x}_k - \vec{v}_i\|,$$

kde $\|\,.\,\|$ je Euklidovská vzdálenost mezi $\vec{x}_k$ a $\vec{v}_i$.
Matice rozkladu $U = [u_{ik}]$, $i = 1, 2, \ldots, c; k = 1, 2, \ldots, N$

- její role je přidělit datapointy shlukum.
- prvky matice U jsou binární:

$$u_{ik} = \begin{cases} 1 & \text{pokud datapoint } k \text{ patří do shluku } i, \\ 0 & \textit{jinak}. \end{cases}$$

Matice rozkladu splňuje následující podmínky

- každý shluk je netriviální, tj. neobsahuje všechny datapointy a jsou neprázdné

$$0 < \sum_{k=1}^{N} u_{ik} < N, \quad i = 1, 2, \ldots, c$$

- každý datapoint patří do jednoho shluku

$$\sum_{i=1}^{c} u_{ik} = 1, \quad k = 1, 2, \ldots, N$$

Kolekce všech matic rozkladu bude značena $\mathbb{U}$
Jako výsledek minimalizace Q konstruujeme matici rozkladu a množinu prototypů.
Formálně vyjadřujeme tento konstrukt jako optimalizační problém s omezeními:

$$\text{Minimální } Q \text{ vzhledem k } \vec{v}_1, \vec{v}_2, \ldots, \vec{v}_c \text{ a } \mathbf{U} \in \mathbb{U}.$$

Existuje mnoho přístupů k této optimalizaci.
Nejběžnější je K-means.

Algoritmus K-means shlukování

inicializuj prototypy $\vec{v}_i$, $i = 1, 2, \ldots, c$ (např. náhodně)

iteruj, dokud se Q neustálí:

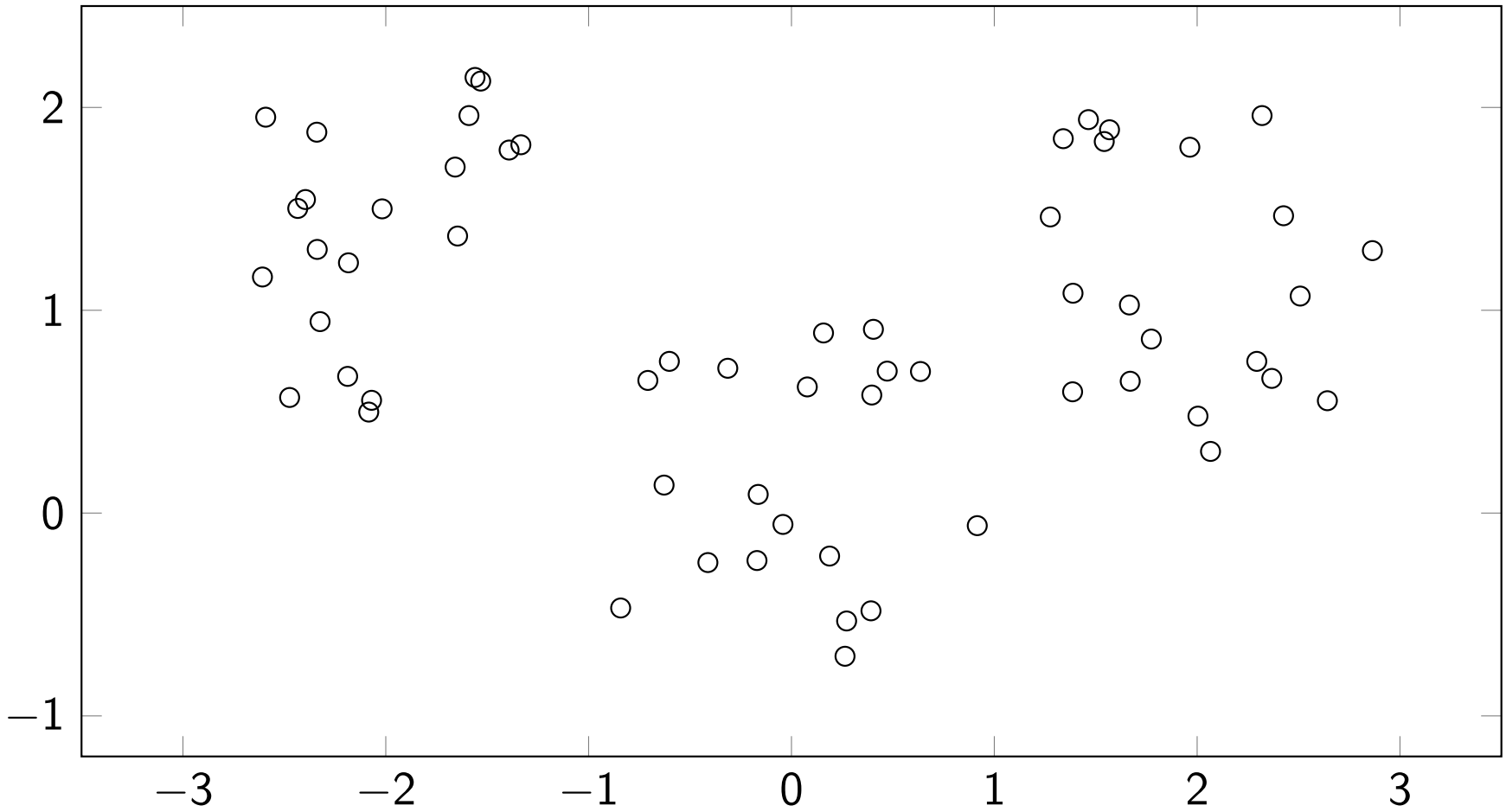
- zkonstruuj rozkladovou matici U takto:

$$u_{ik} = \begin{cases} 1, & \text{pokud } d(\vec{x}_k, \vec{v}_i) = \min d(\vec{x}_k, \vec{v}_i), \\ 0, & \text{jinak.} \end{cases}$$

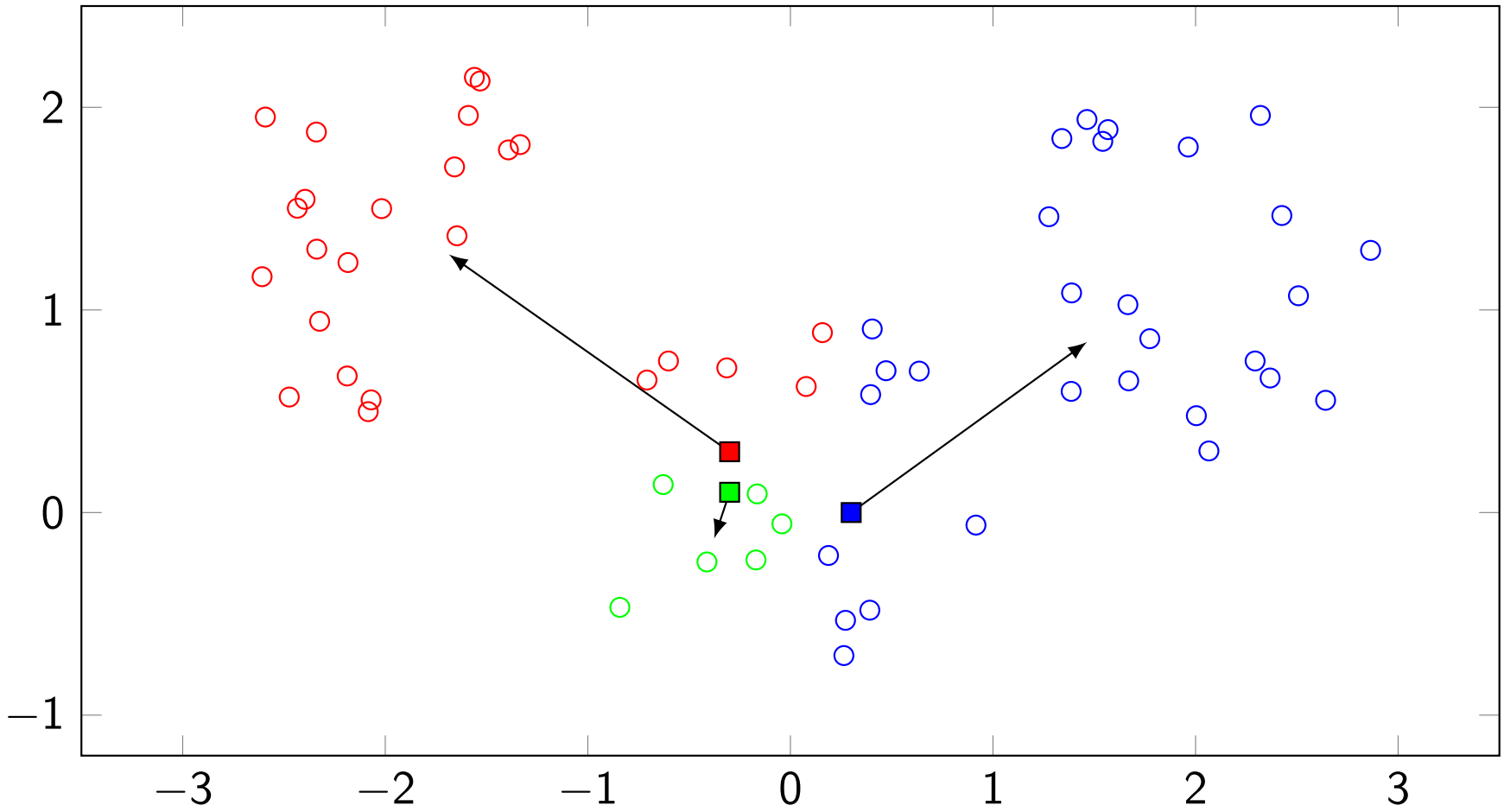
- změň prototypy výpočtem průměru

$$\vec{v}_i = \frac{\sum_{k=1}^{N} u_{ik} \vec{x}_k}{\sum_{k=1}^{N} u_{ik}}$$

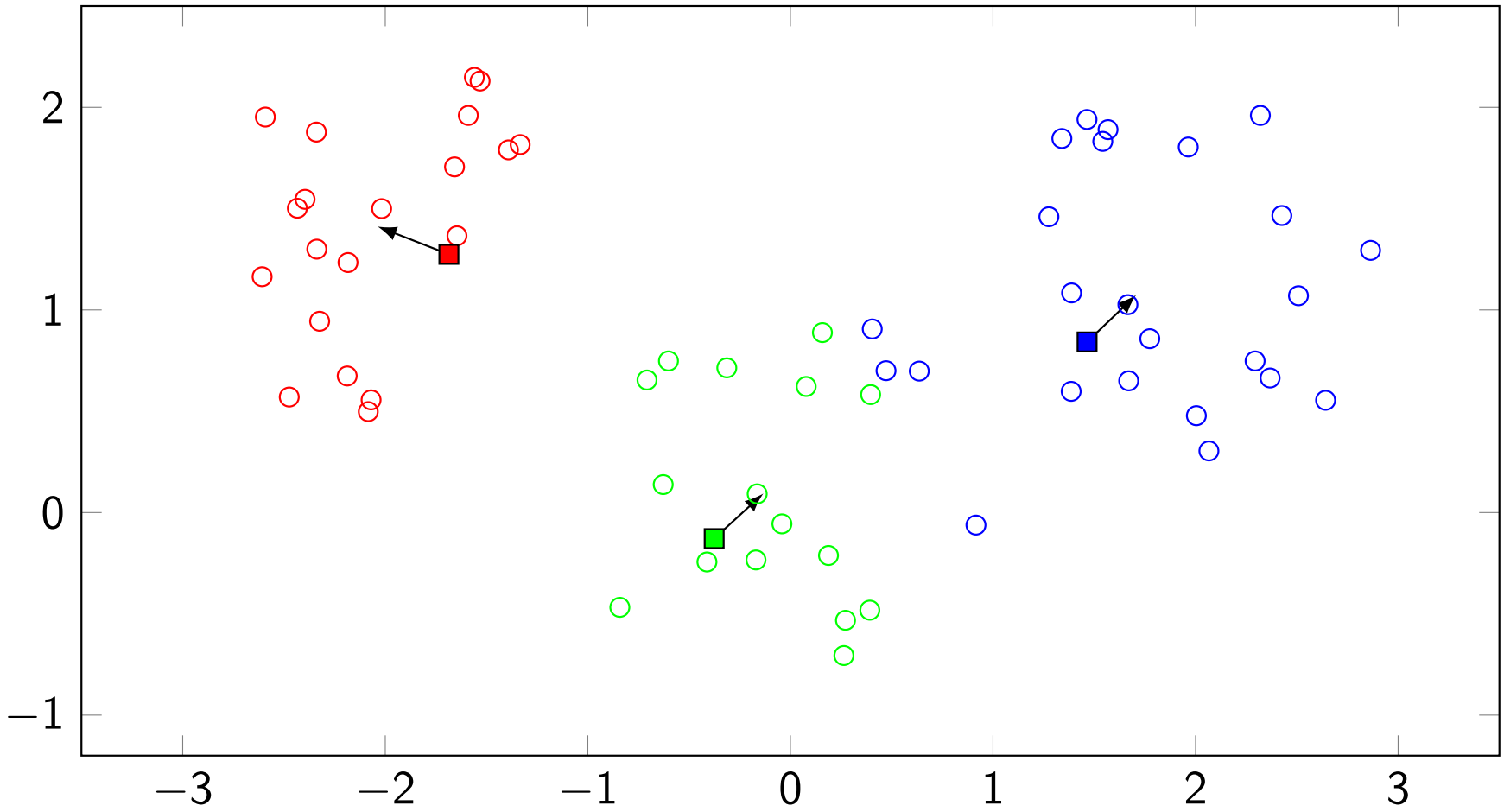
Demonstrace algoritmu K-means shlukování



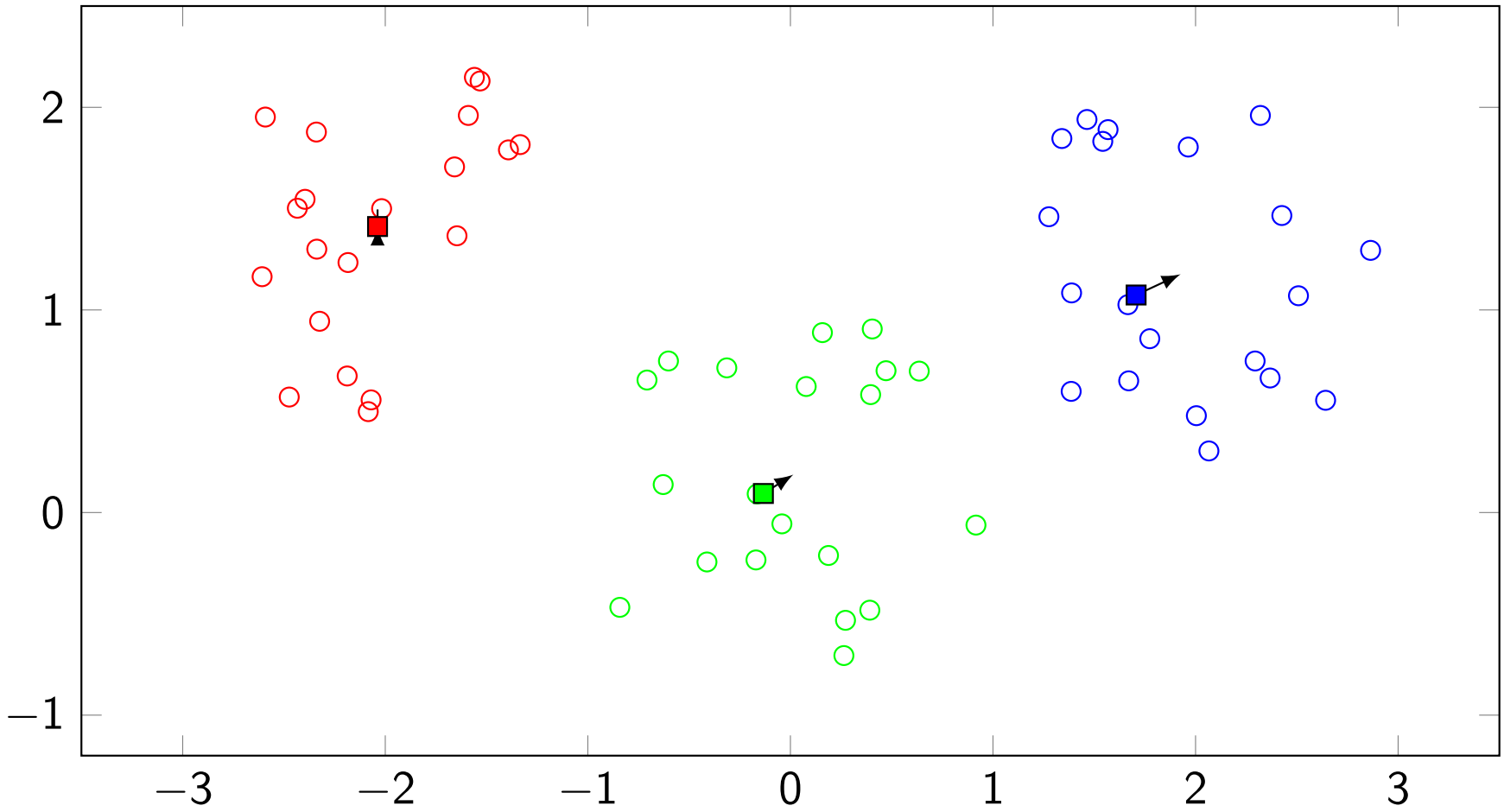
Demonstrace algoritmu K-means shlukování



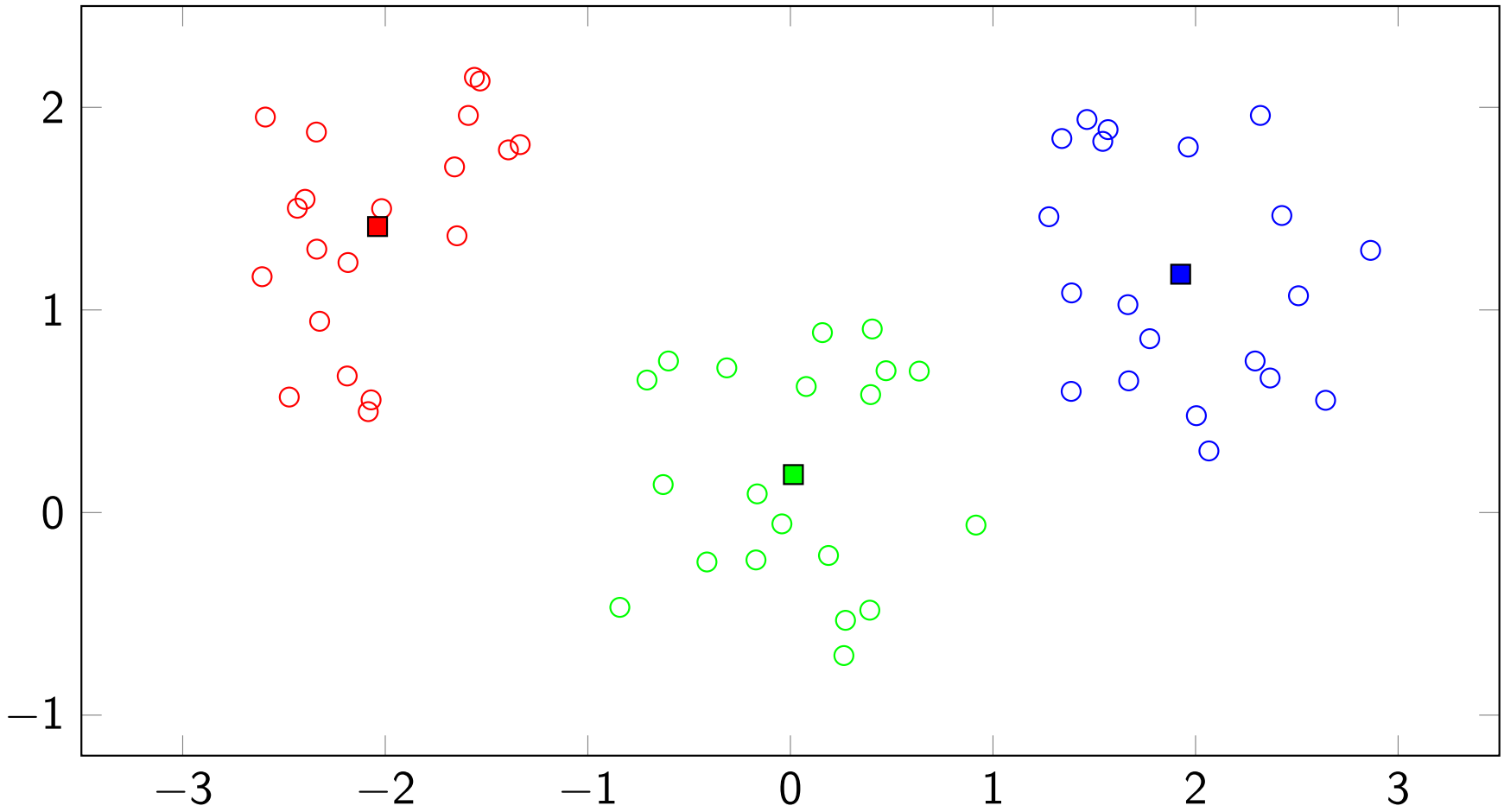
Demonstrace algoritmu K-means shlukování



Demonstrace algoritmu K-means shlukování



Demonstrace algoritmu K-means shlukování



Dolování asociačních pravidel (association rules mining)
hledájí se zajímavé asociace (vztahy, závislosti) ve velkých datasetech.
Aplikace:
Market-basket analysis (analýza nákupního košíku, MBA).

snaží se najít vzory v chování zákazníků v podobě asociací mezi zbožím, které si zákazníci kupují společně (v jednom nákupu).
Například je možno objevit, že zákazníci si kupují mléko a chleba společně, a dokonce že určitý chléb je kupován společně s určitým mlékem. (např. vícezrný chléb a sojové mléko).

To může být využito k uspořádání zboží v obchodě; v rozvržení slev (aby neslevili vícezrný chléb i sojové mléko současně)...

$$\{\text{mléko, máslo}\} \Rightarrow \{\text{chleba} \quad [25\%, 60\%]\}$$

Toto pravidlo říká,

- že když si někdo koupil mléko, tak si koupil i chleba.
- support 25% znamená, že mléko a chleba dohromady bylo koupeno v 25% případů.
- confidence 60% znamená, že 60% košíků, které obsahovaly mléko obsahovaly taky chleba.

Úlohy DM

Učení s učitelem (supervised learning)

Máme k dispozici

- kolekci dat (vzorů)

$$\mathbf{X} = \{x_1, x_2, \ldots, x_N\},$$

- jejich charakterizaci:
  - hodnoty kvalitativní proměnné $\rightarrow$ klasifikace
  - hodnoty spojité proměnné $\rightarrow$ regrese, aproximace

V klasifikaci: každý datapoint $\vec{x}_k$ má určitý label ω_k. kde hodnoty ω_k přicházejí z nějaké malé množiny čísel

$$\omega \in \{1, 2, \ldots, c\},$$

kde c je počet tříd (*class*)
Cílem je vytvořit klasifikátor (*classifier*), který je konstrukcí funkce Φ, která datapointům přiřazuje labely.

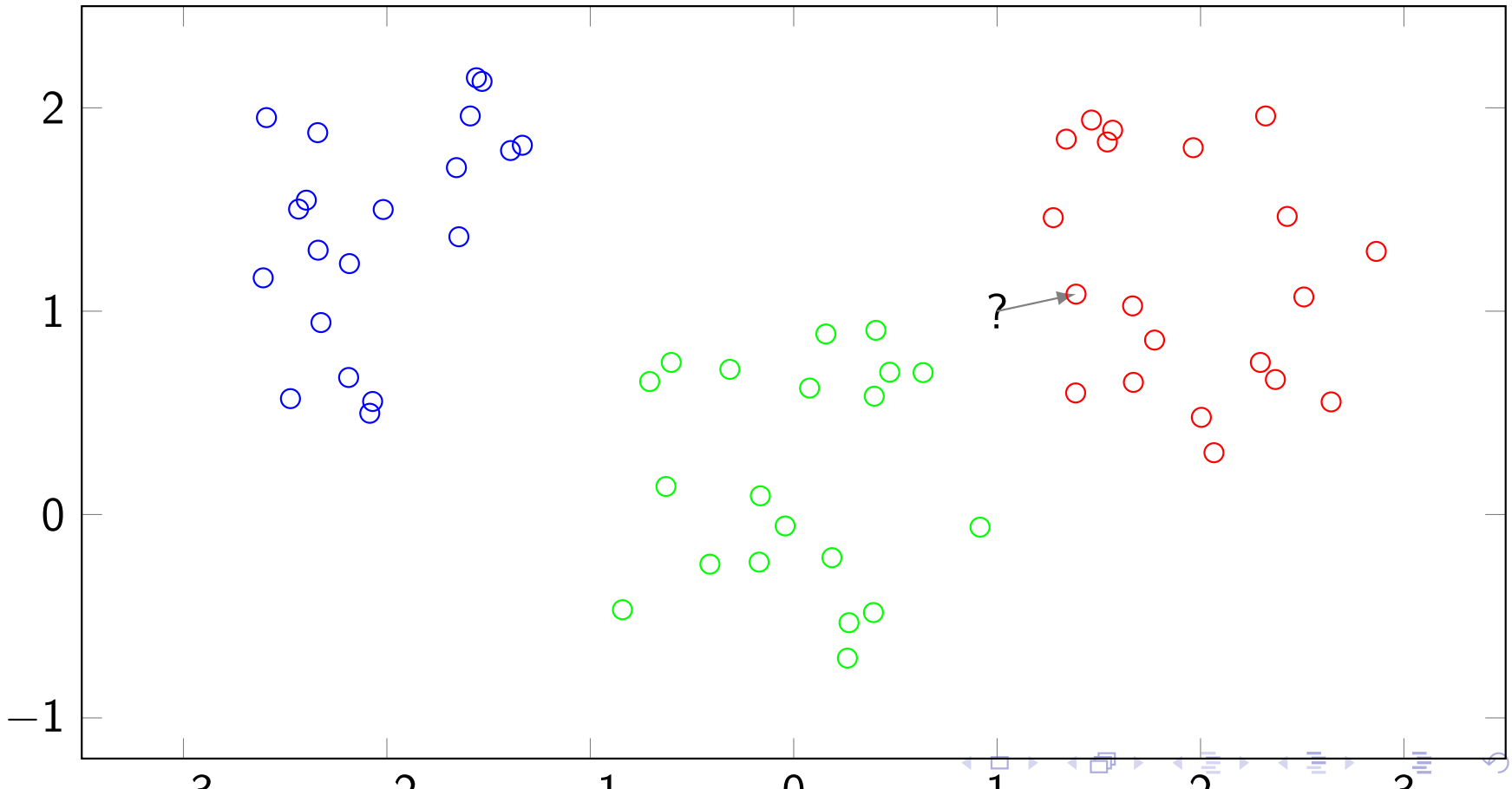
Klasifikace

Příklad jednoduchého klasifikátoru:

nearest-neighbor klasifikátor

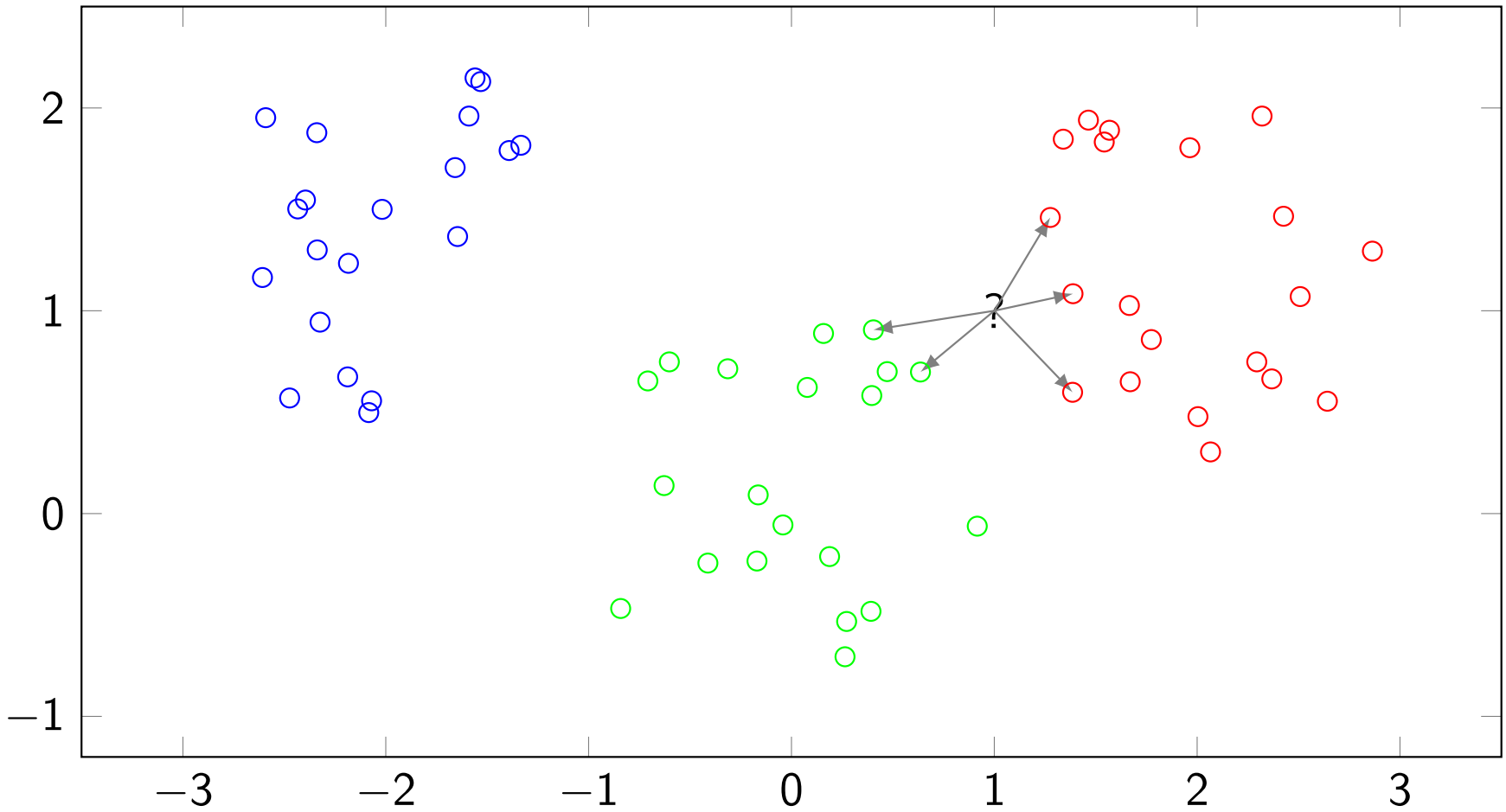
Je dána kolekce označených dat, nový datapoint $\vec{x}$ je klasifikován na základě vzdálenosti mezi ním a ostatních dat.

Klasifikujeme $\vec{x}$ stejnou třídou, jako má její nejbližší soused.



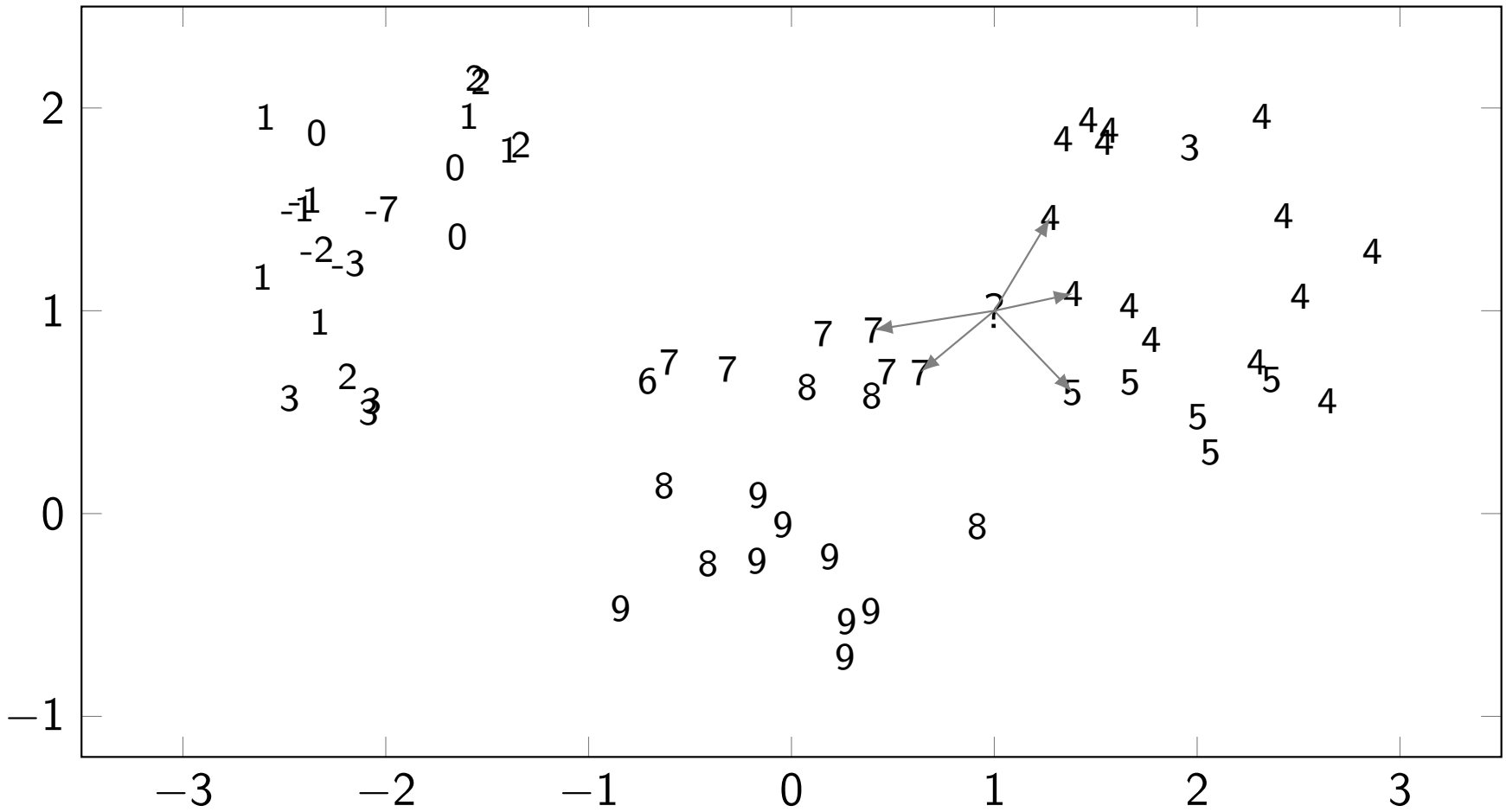
k-nearest neighbor klasifikátor

Klasifikujeme nový datapoint $\vec{x}$, podle hlasování k-nejbližších sousedů (k je obvykle malé liché číslo)



k-nearest neighbor regresor

Výsledek regrese pro nový datapoint $\vec{x}$, podle hlasování k-nejbližších sousedů je dán jako vážený průměr



Další učení: reinforced learning, learning with knowledge hints and semi-supervised learning

Reinforced learning – něco mezi supervised a unsupervised.

Místo přiřazení datapointů k třídám máme méně detailní informaci.
Jen potvrzení; potvrzující signál (*reinforcement; reinforcement signal*).

Např. pro c tříd, potvrzující signál $r(\vec{x})$ může být

$$r(\vec{x}) = \begin{cases} 1 & \text{pokud je označení sudé číslo } (\omega_2, \omega_4, \dots), \\ -1 & \text{jinak.} \end{cases}$$

Dá se říct, že reinforced learning je učení vedené signálem, který je agregátem detailnějších signálů (používaných v supervised learning)

Learning with knowledge hints and semi-supervised learning
Málokdy máme data, která jsou supervised, a málokdy máme data bez jakékoli doménové znalosti.
Ve velkém datasetu $\mathbf{X}$ máme malou část označených datapointů, které vedou k pojmu shlukování s částečným dohledem (partial supervision).
Tyto datapointy jsou pevné body, které nám pomáhají navigovat proces určování (objevování) shluků.