Machine learning a data mining 1

Jan Outrata

KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI

přednášky

Data

Data pro DM a ML

Významné „znát data“:

- **typ dat** – např. typy atributů popisujících objekty/záznamy (kvalitativní, kvantitativní), obecné a speciální vlastnosti (časoprostornost, vztahy mezi objekty); určuje použité metody
- **kvalita dat** – např. šum, anomálie (outliers), chybějící hodnoty, nekonzistence, duplicity, vychýlení (bias), nereprezentativnost; zvýšení zvyšuje kvalitu výsledků analýzy
- **předzpracování** – např. převedení spojitých atributů na diskrétní, snížení počtu atributů; pro zvýšení kvality nebo přizpůsobení zvolené metodě analýzy
- **vztahy** (předem) – např. podobnosti nebo vzdálenosti mezi objekty; analýza nad vztahy mezi objekty místo přímo nad objekty (např. shlukování)

Data pro DM a ML

Významné „znát data“:

- **typ dat** – např. typy atributů popisujících objekty/záznamy (kvalitativní, kvantitativní), obecné a speciální vlastnosti (časoprostornost, vztahy mezi objekty); určuje použité metody
- **kvalita dat** – např. šum, anomálie (outliers), chybějící hodnoty, nekonzistence, duplicity, vychýlení (bias), nereprezentativnost; zvýšení zvyšuje kvalitu výsledků analýzy
- **předzpracování** – např. převedení spojitých atributů na diskrétní, snížení počtu atributů; pro zvýšení kvality nebo přizpůsobení zvolené metodě analýzy
- **vztahy** (předem) – např. podobnosti nebo vzdálenosti mezi objekty; analýza nad vztahy mezi objekty místo přímo nad objekty (např. shlukování)

Př. vědět, že nějaký atribut je jen identifikátor objektů

Data pro DM a ML

dataset (data set) = kolekce (množina) **objektů**/záznamů (bodů, vektorů, vzorů, událostí, případů, pozorování apod.)

objekty popsány **atributy** (proměnnými, charakteristikami, vlastnostmi apod.) = vlastnost charakterizující objekt, s různou hodnotou pro různé objekty, např. tvar, nebo v čase, např. velikost

$\Rightarrow$ **objekt-atributová data** – strukturovaná, charakteristická pro DM a ML

- reprezentace datasetu typicky tabulkou – objekty = řádky, atributy = sloupce, ale i jinak (např. grafem)

Data pro DM a ML

dataset (data set) = kolekce (množina) **objektů**/záznamů (bodů, vektorů, vzorů, událostí, případů, pozorování apod.)

objekty popsány **atributy** (proměnnými, charakteristikami, vlastnostmi apod.) = vlastnost charakterizující objekt, s různou hodnotou pro různé objekty, např. tvar, nebo v čase, např. velikost

⇒ **objekt-atributová data** – strukturovaná, charakteristická pro DM a ML

- reprezentace datasetu typicky tabulkou – objekty = řádky, atributy = sloupce, ale i jinak (např. grafem)

atribut → Teorie měření (measurement theory):

míra (měření) = předpis (funkce) přiřazující atributu symbolickou nebo číselnou hodnotu, např. malý/střední/velký vs. číslo

(proces) **měření** = aplikace míry (funkční hodnota), přiřazení konkrétní hodnoty atributu pro daný objekt, např. změření velikosti

Typ atributu

- vlastnosti atributu nemusí být stejné jako vlastnosti hodnot míry

→ jaké vlastnosti atributu jsou reflektovány hodnotami míry (a obráceně, jaké vlastnosti hodnot, např. různost nebo uspořádání, jsou konzistentní s vlastnostmi atributu), např. ID vs. velikost a číslo

∼ **typ míry**

⇒ identifikace vlastností (operací) hodnot korespondujících s vlastnostmi atributu: **různost**, **uspořádání**, **sčítání** (a odčítání), **násobení** (a dělení)

- významný, pro „zacházení s atributem“, např. nepočítat průměr ID, nebo výběr metody analýzy

Asymetrické atributy

= důležité jen nenulové hodnoty = prezence

- často u datasetů většina hodnot pro objekty nulových
- významné pro např. asociační analýzu

Typy atributů

(S. S. Stevens, psycholog)

- **kategorické (kvalitativní)** $\sim$ symboly
  - **nominální**: hodnoty jen různá jména, jen pro rozlišení objektů (různost), např. ID, tvar, validní operace např. výběr, kontingence, korelace, transformace neměnící význam bijektivní
  - **ordinální**: pro uspořádání objektů, např. pořadí, známky, navíc operace např. medián, korelace s uspořádáním (rank), transformace zachovávající pořadí
- **numerické (kvantitativní)** $\sim$ čísla
  - **intervalové**: smysluplné rozdíly mezi hodnotami (existuje jednotka), např. datum, navíc operace např. (aritmetický) průměr, směrodatná odchylka, transformace lineární
  - **poměrové (ratio)**: smysluplné i podíly hodnot, např. počet, délka, navíc operace např. (geometrický) průměr, procenta, transformace i nelineární

- **diskrétní** = konečně nebo spočetně mnoho hodnot, typicky symbolických nebo celočíselných, **binární** = dvě hodnoty, typicky kategorické (ale i např. počet)
- **spojité (continuous)** = nespočetně mnoho hodnot, reálně-hodnotové (reprezentované s plovoucí řádovou čárkou s omezenou přesností), typicky numerické

Typy datasetů

Charakteristiky mající vliv na DM a ML:

- **dimenze** = počet atributů – mnoho → curse of dimensionality → předzpracování redukce
- **řídkost (sparsity)** = relativní množství nulových hodnot (asymetrických) atributů – často vysoká
- **rozlišení** – vlastnosti dat různé pro různé „úrovně pohledu“, např. rozdíly hodnot dle jejich přesnosti, „viditelnost“ vzorů nebo šumu, typicky u časoprostorových dat

Záznamová data

= všechny objekty popsány stejnou pevnou množinou atributů

- žádné (explicitní) vztahy mezi objekty nebo atributy
- reprezentace tabulkou, uložení **flat** file nebo relační databáze
- typická pro DM a ML

Typy datasetů

Záznamová data

- **transakční** = kolekce transakcí (objekty) jako množin položek (items, atributy) ∼ „nákupní košíky“ produktů... market basket data, položky → asymetrické atributy, i nebinární (počet, cena)
- **maticová** – číselné atributy, objekty ∼ body (vektory) vícerozměrného prostoru, rozměr ∼ atribut, → $n \times m$ matice pro n objektů (řádky) a m atributů (sloupce) – maticové operace
- **řídká (sparse)** – asymetrické atributy stejného typu, např. také transakční, document-term (dokumenty = objekty, termy/slova v nich = atributy, počet výskytů = hodnota, na pořadí slov nezáleží)

Grafová data

- **graf pro vztahy mezi objekty** – objekty = uzly, vztahy = hrany a vlastnosti hran (orientace, váha apod.), např. hypertextové dokumenty s odkazy
- **objekt = graf** = strukturované objekty – mají „podobjekty“ se vztahy → **subgraph mining**, např. chemické sloučeniny

Typy datasetů

Uspořádaná data

= uspořádané atributy, např. v čase nebo prostoru

- **sekvenční** = sekvence (hodnot) atributů, např. genomická (geny jako sekvence nukleotidů) – typicky predikce podobností genů z podobností sekvencí
- **temporální** (časová, sequential) – čas pro objekt nebo hodnotu atributu, rozšíření záznamových dat, např. transakční s časem celých transakcí nebo zařazení položek do transakce, **časová autokorelace** = podobné hodnoty atributů blízké v čase
  - **časových řad** – objekt = časová řada, např. měření v čase
- **prostorová** – prostorové informace k hodnotě atributu, např. měření/modelování pomocí mřížky, meteorologická, **prostorová autokorelace** = podobné hodnoty atributů blízké v prostoru

Ne objekt-atributová data:

→ extrakce atributů a vytvoření objekt-atributových

- např. společné části prvků dat = objektů jako (asymetrické) binární atributy
- problém postihnout všechny informace v datech, např. vztahy mezi objekty nebo atributy navzájem: např. časové řady pro body (= objekty) v prostoru

Kvalita dat

- analyzovaná data často sbíraná za jiným účelem (nebo „do budoucna“) $\Rightarrow$ nemožnost „řešení kvality u zdroje“, prevence problémů

$\rightarrow$ dosažení požadované úrovně u dat i výsledků analýzy:

- detekce a oprava problémů = **data cleaning**
- metody s tolerancí problémů

Problémy při sběru/měření dat

- způsobené lidskými chybami, omezeními sběrného/měřícího zařízení, vlivy okolí, systematickými chybami apod.

$\rightarrow$ chyby způsobu měření: šum, artefakty, bias, nedostatečná přesnost hodnot aj.

$\rightarrow$ chyby sběru/procesu měření: anomálie (outliers), chybějící a nekonzistentní hodnoty, duplicitní objekty aj.

- systematické i náhodné
- dále obecné, pro (běžné) specifické, např. časté překlepy, specifické metody detekce a opravy

Kvalita dat

Šum a artefakty

= náhodné (šum) nebo systematické (artefakty) chybné hodnoty nebo přidané objekty, např. kvůli vadě měřidla

- často pro maticová, časoprostorová data → metody redukce šumu ve zpracování signálu, obrazu apod.
- obecně obtížné odstranit (co je „šum“ a co není?) → metody s **odolností vůči šumu**

Přesnost (precision) a bias

- pro informování o nebo stanovení kvality měření (dat) a výsledků
- **přesnost (precision)** = blízkost opakovaných měření (stejné veličiny), měřena obvykle směrodatnou odchylkou, použití pouze **platných číslic (significant digits)** – pro vyjádření hodnot jen tolik číslic, kolik odpovídá přesnosti, např. měření pravítkem s milimetry jen na milimetry, s přesností $\pm 0,5$ mm
- **bias** = systematická odchylka měření od skutečné hodnoty, obvykle rozdíl mezi průměrem a hodnotou
- **přesnost (accuracy)** = blízkost opakovaných měření skutečné hodnotě, obecnější ∼ stupeň chyby měření, závisí na precision a bias

Kvalita dat

Anomálie (Outliers)

= vlastnostmi odlišné objekty od většiny ostatních nebo netypické hodnoty atributu, mnoho definic, typicky „extrémní“

- mohou být legitimní (a hledané) – rozdíl oproti šumu nebo artefaktům

Chybějící hodnoty (nebo i celé objekty)

- nezískané, atribut pro nějaký objekt (podmíněně) neplatný apod.

→ vyřazení objektů nebo atributů – ne mnoho, i neúplně popsané objekty mohou být užitečné nebo atributy významné

→ odhad – např. nejčastější (kategorická) hodnota, průměr, interpolace z ostatních („sousedních/nejbližších“ numerických) hodnot/objektů, hodnota z „nejbližšího/nejpodobnějšího“ objektu aj.

→ ignorování – adaptace metod, vynechání při používání atributu (např. výpočet podobnosti objektů), může vést k jen přibližným nebo i jiným výsledkům

Kvalita dat

Nekonzistentní hodnoty

- vzhledem k významu atributu nebo explicitním vztahům mezi atributy
- často překlepy a zjevné chyby, snadno zjistitelné a opravitelné
- pro opravu potřeba dodatečná nebo externí informace

Duplicitní objekty

= reprezentující stejný „skutečný“ – potřeba znát, i s příp. různými hodnotami některých atributů (= nekonzistence) → **deduplikace**

! ne „podobné“ objekty reprezentující různé „skutečné“

Kvalita dat

Problémy při použití dat

= vhodná pro zamýšlené použití?

- aktuálnost – užitečnost dat i výsledků analýzy jen po omezenou dobu
- relevance – potřebné informace?, jinak nízká vypovídající hodnota výsledků, problém **biasu vzorkování (sampling bias)** = data (vzorek) neobsahují dostatek různých objektů podle jejich zastoupení ve skutečnosti $\Rightarrow$ chybné (zavádějící) výsledky analýzy
- znalost (dokumentace) – např. o typech atributů, přesnosti hodnot, chybějících hodnotách (jejich specifikace), provázaných atributech $\rightarrow$ redundantní, výběr jen některých, původu dat atd.

Předzpracování dat (data preprocessing)

→ výběr objektů a/nebo atributů nebo vytvoření/změna atributů pro zlepšení analýzy (čas, kvalita)

Agregace

= kombinace více objektů do jednoho

- jak zkombinovat hodnoty atributů? – např. součet, průměr (kvantitativní), výběr nejčastější, sjednocení (kvalitativní)
- efektivně zrušení atributů nebo redukce hodnot atributu, např. dnů na měsíce
- menší data ⇒ možné náročnější metody, „pohled z vyšší úrovně“, agregované „stabilnější“ – menší variabilita (rozptyl), ale možná ztráta detailů

Vzorkování (Sampling)

= výběr objektů; pro prvotní průzkum (statistika) i finální analýzu (schůdnější, náročnějším algoritmem)

∼ celá data, jestliže je **reprezentativní** = maximálně stejné vlastnosti (zájmu) jako celá data, např. průměr hodnot atributů

→ **náhodné** – stejná pravděpodobnost výběru každého objektu

- bez ponechání – v populaci k výběru, pravděpodobnost výběru roste
- s ponecháním – objekt může být vybrán vícekrát
- nezohlednění (různých) četností výskytu objektů různých typů – méně četných méně vybraných

→ **stratifikované** – vybraný stejný počet objektů z každé dané skupiny (typu) objektů nebo relativní k velikosti skupiny

! velikost vzorku – větší (pravděpodobně) reprezentativnější, menší možná ztráta informace → dostatečná pravděpodobnost výběru

→ **adaptivní/progresivní** – zvyšující se velikost vzorku dokud ne dostatečná, např. dokud se zvyšuje kvalita výsledků, např. přesnost modelu

Redukce dimenzionality

- pro metody lepší menší počet atributů (dimenze dat): eliminace redundantních nebo irelevantních, redukce šumu, srozumitelnější model, snadnější vizualizace (dat i výsledků, často po dvojicích nebo trojicích atributů), výpočetní nároky, ...
- **curse of dimensionality** = (značně) náročnější analýza při zvyšující se dimenzi dat $\Rightarrow$ data řidší, relativně málo objektů pro tvorbu modelu (např. klasifikace), méně významné hustota a vzdálenost mezi objekty (např. shlukování)

$\rightarrow$ výběr atributů z původních = **feature selection**

$\rightarrow$ vytvoření nových atributů z původních = **feature creation**

$\rightarrow$ lineární algebra: projekce dat z vysokodimenzionálního prostoru do ménědimenzionálního, typicky pro spojité atributy, **analýza hlavních komponent (principal component analysis, PCA)** (nové atributy = hlavní komponenty = navzájem ortogonální lineární kombinace původních zachycující maximum variace dat) a **singular value decomposition (SVD)**, **faktorová analýza** (původní atributy = lineární kombinace nových „skrytých“), locally linear embedding (LLE), multidimensional scaling (MDS) aj.

Převod typů atributů

- některé metody vyžadují kategorické atributy, např. klasifikace, nebo (asymetrické) binární atributy, např. asociační analýza

→ **diskretizace** = převod spojitého atributu na diskrétní (kategorický)

→ **binarizace** = převod spojitého nebo diskrétního atributu na binární atributy

- také snížení počtu hodnot kategorického atributu – diskretizace (ordinální), sloučení více hodnot do jedné (nominální) – na základě např. vztahů mezi hodnotami
- vliv na výsledky metody, ideálně dle metody

Binarizace

1. jednoznačné přiřazení čísla z intervalu $[0, m-1]$ každé z m hodnot kategorického atributu, se zachováním pořadí u ordinálního atributu
2. reprezentace čísel ve dvojkové soustavě s $\lceil \log_2(m) \rceil$ ciframi a nový binární atribut pro každou cifru

! nechtěné vztahy mezi binárními atributy – původní hodnoty reprezentované více (korelovanými) atributy, ne asymetrické

→ nový (asymetrický) binární atribut pro každou hodnotu kategorického atributu (včetně dvouhodnotového)

Diskretizace atributů

1. rozdělení (rozklad) hodnot spojitého atributu, po jejich setřízení, do n (disjunktních) intervalů specifikováním $n-1$ dělicích bodů
2. zobrazení (hodnot) intervalu na (stejnou) hodnotu nového kategorického atributu

? kolik intervalů/dělicích bodů a jakých hodnot $x_1, \ldots, x_{n-1}$

- $\{(x_0, x_1], (x_1, x_2], \ldots, (x_{n-1}, x_n)\}$, x_0, x_n příp. $\pm\infty$, nebo $x_0 < x \leq x_1 < x \leq x_2 \ldots x_{n-1} < x < x_n$

$\rightarrow$ zadaný počet stejně velkých intervalů – ovlivnění anomáliemi (outliers)

$\rightarrow$ zadaný počet intervalů s maximálně stejným počtem hodnot pro (dané) objekty

$\rightarrow$ intervaly blízkých hodnot $\rightsquigarrow$ shlukovací metoda, např. K-means; vizuálně, doménově specificky aj.

$\sim$ unsupervised diskretizace

Supervised diskretizace atributů

= využití hodnot cílového atributu = **tříd klasifikace (class labels)**

- cíl: maximální jedinečnost hodnoty cílového atributu (class label) pro objekty s hodnotou (diskretizovaného atributu) v intervalu (= „class label v intervalu“) = „čistota“ (purity) intervalu $\rightsquigarrow$ jak měřit?, minimálně velké intervaly

$\rightarrow$ interval = každá hodnota a opakované slučování „podobných“ sousedních – jak podobných (na základě class labels)?

- impurity i-tého intervalu = **entropie**: $e_i = \sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$, k počet různých class labels, $p_{ij} = m_{ij}/m_i$ podíl class label j v intervalu i, m_{ij} počet výskytů class label j v intervalu i, m_i počet všech class labels v intervalu i
- entropie rozdělení (rozkladu): $e = \sum_{i=1}^{n} w_i e_i$, n počet intervalů, $w_i = m_i/m$ podíl class labels v intervalu i, m počet všech class labels

$\rightarrow$ interval = všechny setřízené hodnoty a opakovaně bisekce intervalu s nejvyšší entropií tak, aby rozdělení mělo nejmenší entropii, dokud ne zadaný počet intervalů nebo dostatečně nízká nebo již neklesající entropie – potřeba testovat jako dělicí bod (libovolný) bod mezi každými dvěma sousedními hodnotami intervalu pro objekty s různými class labels

Úpravy hodnot atributů

$\sim$ transformace proměnných (variable transformation)

$=$ stejná úprava všech hodnot atributu x pro všechny objekty

Jednoduché funkce

- pro číselné atributy např. $\log x$, $\sqrt{x}$, $1/x$ aj.

! změna rozsahu hodnot, jejich pravděpodobnostního rozložení, např. na normální, uspořádání (!), nedefinování funkce pro některé hodnoty aj.

Normalizace/standardizace

$=$ zajištění nějaké vlastnosti (číselných) hodnot

- např. běžná „normalizace“ $(x - \min(x))/(\max(x) - \min(x))$ pro $\min(x) = 0$, $\max(x) = 1$, (statistická) standardizace $(x - \overline{x})/s_x$, $\overline{x}$ průměr hodnot x, s_x směrodatná odchylka hodnot x pro $\overline{x} = 0$ a $s_x = 1$
- potřebné pro „férové“ porovnávání nebo kombinaci atributů, např. s různými rozsahy hodnot („škálami“), nebo „korektní“ porovnání objektů na základě atributů, např. podobnost/vzdálenost objektů

Podobnost a nepodobnost objektů

- používané v mnoha metodách (shlukování, klasifikace metodou nejbližšího souseda, detekce anomálií aj.) – převod dat do prostoru (ne)podobností a analýza v něm
- **podobnost s** objektů $\sim$ (numerická) míra stupně stejnosti objektů, typicky mezi 0 (žádná) a 1 (plná)
- **nepodobnost d** objektů $\sim$ (numerická) míra stupně různosti objektů, často $\sim$ **vzdálenost** – speciální případ, běžně od 0 (žádná) do ∞

$\rightsquigarrow$ **blízkost (proximity) p**

Transformace

- blízkost (podobnost) $\rightarrow [0, 1]$: pro podíl, typicky $p' = (p - \min(p))/(\max(p) - \min(p))$ pro konečné $\max(p) - \min(p)$, pro $p \in [0, \infty]$ např. $p' = p/(p + 1)$ – jiné (nelineární) vztahy mezi hodnotami, možná ztráta informace nebo jiný význam
- podobnost $\longleftrightarrow$ nepodobnost: typicky $d = 1 - s$ pro $s, d \in [0, 1]$, jinak např. $s = 1/(d + 1), s = e^{-d}, s = 1 - (d - \min(d))/(\max(d) - \min(d))$, obecně jakákoliv monotónní klesající funkce

Blízkost objektů

... kombinace blízkostí hodnot atributů objektů

- měla by respektovat typy atributů, může být potřeba normalizace/standardizace

Atributy:

- nominální: jen různost hodnot $\Rightarrow s = 1$ pro stejné hodnoty, jinak $s = 0$, d opačně
- ordinální: pořadí hodnot, bližší $\sim$ podobnější $\Rightarrow$ jednoznačné zobrazení hodnot na čísla se zachováním pořadí a $d =$ rozdíl odpovídajících čísel / rozsah čísel – jaká čísla (a rozdíly mezi nimi odpovídající rozdílům mezi hodnotami)?
- intervalové a poměrové: $d =$ (absolutní) rozdíl hodnot

- atributy stejného typu $\rightarrow$ nepodobnosti (vzdálenosti) a podobnosti objektů dále
- atributy různého typu $\rightarrow$ blízkosti hodnot atributů jednotlivě (nebo po skupinách stejného typu) a kombinace, typicky průměr – s výjimkou asymerických atributů s nulovými hodnotami pro objekty a atributů s chybějící hodnotou pro objekt
- různé váhy atributů

Nepodobnosti (vzdálenosti) objektů

- pro objekty ... vektory $\mathbf{x} = (x_1, \ldots, x_m)$ hodnot m (číselných) atributů
- **euklidovská vzdálenost**:

$$\sqrt{\sum_{k=1}^{m} (x_k - y_k)^2}$$

- **Minkowského vzdálenost**:

$$\left(\sum_{k=1}^{m} |x_k - y_k|^r \right)^{1/r}$$

  - $r = 1$: city block (Manhattan, taxicab, L_1 norm), maximální, pro binární atributy **Hammingova vzdálenost** = počet atributů (bitů) s rozdílnými hodnotami pro objekty
  - $r = 2$: euklidovská (L_2 norm)
  - $r = \infty$: supremum (L_{max}, L_∞ norm), $= \lim_{r \to \infty} \ldots$, minimální

- atributy s různými rozsahy hodnot („škálami“) $\rightarrow$ (statistická) standardizace

Nepodobnosti (vzdálenosti) objektů

- **Mahalanobisova vzdálenost**: pro atributy s různými rozsahy (rozptyly) hodnot a korelacemi mezi sebou

$$\sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$$

  $\mathbf{S}^{-1}$ inverzní kovariantní matice dat (matice kovariancí mezi každými dvěma atributy)
- pro spojité atributy, např. (hustá) číselná data

Míry (ne)podobnosti

Metrika

1. **pozitivita**: $d(\mathbf{x}, \mathbf{y}) \geq 0 \ \forall \mathbf{x}, \mathbf{y}$, $d(\mathbf{x}, \mathbf{y}) = 0 \leftrightarrow \mathbf{x} = \mathbf{y}$
2. **symetrie**: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \ \forall \mathbf{x}, \mathbf{y}$
3. **trojúhelníková nerovnost**: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \ \forall \mathbf{x}, \mathbf{y}, \mathbf{z}$

$\sim$ vzdálenost, např. Minkowského, Mahalanobisova

- ne metrika např. velikost (počet prvků) rozdílu množin: $d(A, B) = |A \setminus B| \rightarrow$ metrika?

Podobnost (typicky)

1. $s(\mathbf{x}, \mathbf{y}) \in [0, 1] \ \forall \mathbf{x}, \mathbf{y}$, $s(\mathbf{x}, \mathbf{y}) = 1 \leftrightarrow \mathbf{x} = \mathbf{y}$
2. **symetrie**: $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x}) \ \forall \mathbf{x}, \mathbf{y}$
3. některé lze převést na metriky, např. Jaccard koeficient, kosinovou
4. ne symetrická např. podíl počtu binárních atributů s hodnotou 1 pro jednotlivé objekty $\rightarrow$ symetrická?

Podobnosti objektů

Podobnostní koeficienty

= podobnosti pro binární atributy, $\in [0, 1]$

- f_{xy} = počet atributů s hodnotou x pro $\mathbf{x}$ a y pro $\mathbf{y}$, $xy \in \{00, 01, 10, 11\}$
- **simple matching koeficient**:

$$\frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

- **Jaccard koeficient**: pro asymetrické binární atributy, např. (řídká) transakční data

$$\frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Podobnosti objektů

Pro (asymetrické) číselné atributy, např. (řídká) document-term data, $\in [0, 1]$:

- **kosinová podobnost**: kosinus úhlu mezi (vektory) $\mathbf{x}$ a $\mathbf{y}$

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \, ||\mathbf{y}||} = \frac{\sum_{k=1}^{m} x_k y_k}{\sqrt{\sum_{k=1}^{m} x_k^2} \sqrt{\sum_{k=1}^{m} y_k^2}} = \frac{\mathbf{x}}{||\mathbf{x}||} \cdot \frac{\mathbf{y}}{||\mathbf{y}||}$$

- **rozšířený Jaccard (Tanimoto) koeficient**:

$$\frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||^2 + ||\mathbf{y}||^2 - \mathbf{x} \cdot \mathbf{y}}$$

Korelace objektů

= míra lineární závislosti mezi hodnotami atributů objektů ($x_k = ay_k + b$), $\in [-1, 1]$

- podobně i korelace mezi hodnotami atributů (napříč objekty)
- **Pearsonův korelační koeficient**:

$$corr(\mathbf{x}, \mathbf{y}) = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}} s_{\mathbf{y}}}$$

$s_{\mathbf{xy}}$ kovariance mezi $\mathbf{x}$ a $\mathbf{y}$, $s_{\mathbf{x}}$ směrodatná odchylka hodnot $\mathbf{x}$

= kosinová podobnost při $\overline{\mathbf{x}} = \overline{\mathbf{y}} = 0$

Zkoumání dat (data exploration) [SAMOSTUDIUM]

= předběžný průzkum dat pro prvotní charakteristiky $\Rightarrow$ pro výběr metod předzpracování a analýzy

- může „již něco odhalit“, např. vzory po vizualizaci
- použití pro porozumění a interpretaci výsledků analýzy, zejména vizualizace

$\rightarrow$ **souhrnné statistiky**: četnosti, percentily, průměr a medián, rozptyl a směrodatná odchylka, kovariance a korelace aj.

$\rightarrow$ **vizualizace**: histogram, grafy, diagramy, matice atd.

$\rightarrow$ **On-Line Analytical Processing (OLAP)** – data jako vícedimenzionální pole hodnot, souhrnné tabulky, agregace dat (přes dimenze nebo hodnoty)

$\sim$ část **explorativní analýzy dat** (Tukey, statistik, 1970s)

Souhrnné statistiky