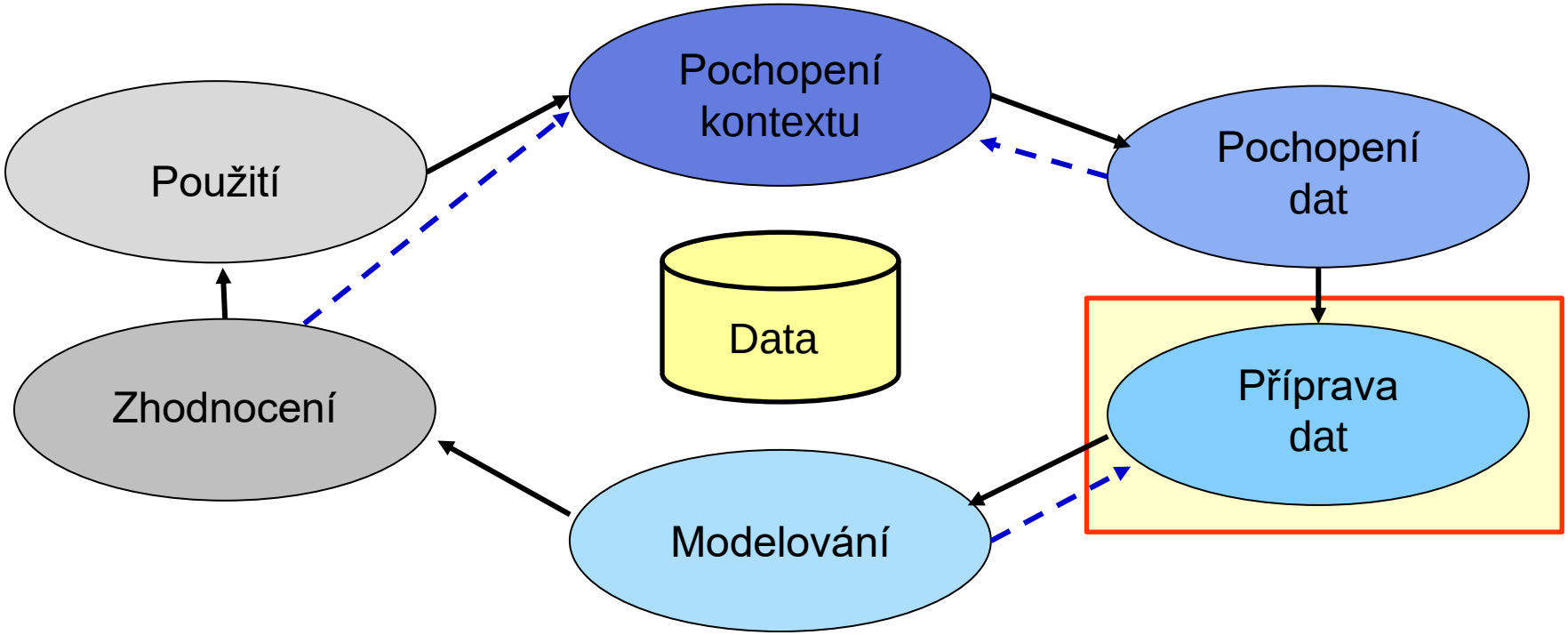
Příprava dat

Ivana Burgetová

Obsah

Krok přípravy dat

- Model životního cyklu analytického projektu CRISP



Krok přípravy dat

- Třetí krok|/fáze modelu procesu CRISP-DM
- Cíl: Připravit data pro aplikaci dolovacího algoritmu.
- Důvody:
  - Zajistit co nejvyšší kvalitu dat
  - Upravit data do podoby vyžadované dolovacím algoritmem, který bude použit pro modelování (pokud už byl zvolen)

Kvalita dat

- Některá kritéria kvality dat
  - Přesnost – vztah k realitě
  - Úplnost – do šířky (atributy), hloubky (záznamy)
  - Konzistence
  - Aktuálnost – podpora rozhodnutí do budoucnosti
  - Věrohodnost
  - Přidaná hodnota – prospěšnost pro danou úlohu
  - Interpretovatelnost hodnot – kódování intervalů, klasifikační třídy atd.
  - Dostupnost, relevance, srozumitelnost, …
- Řada z nich posuzována už v kroku *Pochopení kontextu*, jiné v kroku *Pochopení dat*

Kvalita dat

- Reálná data jsou nekvalitní (*dirty*)
  - neúplná: chybějící hodnoty, chybějící atributy, pouze agregovaná data
    - *např.:* zaměstnání=“ ”
  - zašuměná: obsahují chyby a odlehlé hodnoty
    - *např.:* plat = -10
  - nekonzistentní: rozpory v kódech/identifikátorech, jménech atd.
    - *např.:* věk= 42, dat_narození=03/07/1997
    - *např.:* známky 1,2,3, nyní A, B, C
    - *např.:* rozpory mezi duplicitami
  - navíc rozsáhlá nebo jich je naopak málo: dimenze, počet záznamů

Kvalita dat – příčiny

- Příčiny nízké kvality dat
  - Příčinou neúplnosti může být
    - *Nepoužitelné/nerelevantní (not applicable)* při sběru dat → NULL
    - Rozdílné požadavky při pořizování dat (návrh DB) a jejich analýze
    - Problémy HW, SW nebo člověka
  - Příčinou zašuměných dat může být
    - Porucha zařízení pro sběr
    - Chyba člověka při pořizováni/sběru
    - Chyba při přenosu
  - Zdrojem nekonzistence může být
    - Různé zdroje dat
    - Porušení funkčních závislostí (modifikace jen něčeho při redundanci)

Kvalita dat - důsledek

- Důsledek nízké kvality dat:
  - Nízká kvalita dat → nízká kvalita výsledků
    - Duplicitní, chybějící, nekonzistentní, neaktuální nebo nevěrohodná data mohou vést k nepřesným nebo i nesprávným výsledkům.
  - Kvalitní rozhodování a věrohodné ověřování hypotéz musí vycházet z kvalitních dat.
- Příprava (reálných) dat pro získávání znalostí, stejně tak pro uložení do datového skladu (tzv. ETL proces) má výrazný podíl na celkové pracnosti projektu.

Příprava dat – typické aktivity

- Typické aktivity kroku přípravy dat:
  1. Výběr dat – jaká data použijeme a proč, ve značné míře už na začátku kroku pochopení dat, kritéria:
     - relevance k cílům dolování
     - kvalita dat
     - technická omezení (datové typy, množství, …)
  2. Čištění dat – zvýšení kvality dat
     - chybějící hodnoty
     - šum
     - nekonzistence a redundance
  3. Integrace dat – kombinace dat z několika zdrojů
     - spojování
     - agregace

Příprava dat – typické aktivity

- Typické aktivity kroku přípravy dat:
  4. Úprava datové sady
     - redukce – dimenzionality, počtu záznamů, počtu hodnot
     - řešení nevyváženosti
  5. Transformace dat – úprava hodnot do podoby vhodné/nutné pro dolování
     - konstrukce atributů
     - agregace
     - normalizace
     - diskretizace numerických dat
     - transformace kategorických dat

Obsah

- Krok přípravy dat
- Čištění dat
- Integrace dat
- Úprava datové sady
- Transformace dat
- Další úlohy předzpracování

Čištění dat

- Podstata
  - Proces přípravy dat k analýze odstraněním nebo úpravou dat, která jsou neúplná, nesprávná nebo duplicitní.
- Čištění dat zahrnuje
  - Ošetření chybějících hodnot (*imputation*)
  - Identifikaci odlehlých hodnot a vyhlazení zašuměných dat
  - Úpravu nekonzistentních dat
  - Řešení redundance způsobené integrací dat

Čištění dat – chybějící hodnoty

- Chybějící hodnoty
  - Podoba
    - U datových objektů chybí hodnoty některých atributů

      Příklad: chybí informace o střední škole studenta univerzity
  - Důvody chybějících hodnot
    - hodnota nedodána z důvodu volitelnosti nebo nepochopení
    - určitá data nemusí být považována za důležitá v okamžiku zadávání
    - výpadek zařízení pro sběr dat
    - hodnota nekonzistentní s ostatními, proto vypuštěna
  - Může být potřeba chybějící data doplnit.

Čištění dat – chybějící hodnoty

- Ošetření chybějících hodnot – možnosti:
  1. Ignorování záznamu:
     - Obvykle tehdy, když chybí návěští třídy při klasifikaci, nebo když chybí hodnoty více atributů daného objektu.
     - Není efektivní, když procento chybějících hodnot u jednotlivých atributů značně kolísá, nebo by zůstalo málo dat.
  2. Doplnění chybějících hodnot: ručně nebo automaticky.

- Ručně: pracné, příp. neproveditelné
- Automatické doplnění:
  - globální konstantou

    Příklad: “neznámá”, “ třída?“
  - hodnota charakterizující střed

    Příklad: průměr, medián, modus
  - hodnota charakterizující střed hodnot záznamů patřících do téže třídy
  - nejpravděpodobnější hodnota

    Příklad: odvození založené na bayesovské klasifikaci nebo rozhodovacím stromu

Čištění dat – zašuměná data

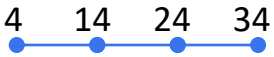
- Zašuměná data:
  - Šum – náhodná chyba nebo odchylka zaznamenaných hodnot.
  - Může se týkat:
    - hodnot atributů – odlehlé a *rozptýlené* hodnoty
    - atributu jako celku – atribut není pro úlohu přínosný
    - celého datového objektu – odlehlý od ostatních
  - Důvody nesprávných hodnot atributů:
    - porucha zařízení pro sběr dat
    - problémy vstupu dat
    - problémy při přenosu dat
    - nekonzistence v konvencích, formátech atd.

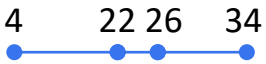
Čištění dat – zašuměná data

- Metody odstranění šumu u numerických atributů:
  1. Plnění (binning)
     - vyhlazení na základě okolních hodnot v seřazených datech.
  2. Regrese
     - vyhlazení vyrovnáním dat podle regresní funkce.
  3. Analýza odlehlých hodnot/objektů
     - detekce a odstranění odlehlých hodnot/objektů, např. využitím shlukování.
- Metody odstranění šumu u kategorických atributů:
  - Oprava textových hodnot (ideálně oproti známému oboru hodnot).
  - Odstranění odlehlých (málo četných) hodnot → prázdná hodnota.

Čištění dat – zašuměná data

- **Plnění:**
  1. Uspořádání dat a rozdělení do tzv. **košů**, varianty:
     - Rozdělení podle stejné šířky (vzdálenosti) (equi/equal-width)
       - Rozdělení rozsahu hodnot na *N* intervalů stejné velikosti, rovnoměrně
       - Jsou-li *A* a *B* nejmenší a největší hodnota atributu, bude šířka intervalu: $w = (B - A)/N$.
       - Nejjednodušší, ale odlehlé hodnoty mohou mít dominantní vliv.
       - Nevypořádá se dobře ani s vychýlenými daty.
     - Rozdělení podle stejné hloubky (frekvence) (equi/equal-depth)
       - Rozdělení rozsahu na *N* intervalů, z nichž každý obsahuje přibližně stejný počet hodnot
       - Dobře škálovatelná metoda (vzhledem k rozložení hodnot)
  2. Vyhlazení dat v koši - např. průměrem koše, mediánem koše, bližší hraniční hodnotou apod.
- Používá se také pro diskretizaci kvantitativních atributů





Čištění dat – zašuměná data

- Plnění:

Příklad: Uspořádaná data podle jednotkové ceny:

4, 9, 15, 21, 24, 24, 24, 26, 27, 28, 29, 34

1.) Rozčlenění do košů stejné hloubky:

- Koš 1: 4, 9, 15, 21
- Koš 2: 24, 24, 24, 26
- Koš 3: 27, 28, 29, 34

2a.) Vyhlazení průměrem koše:

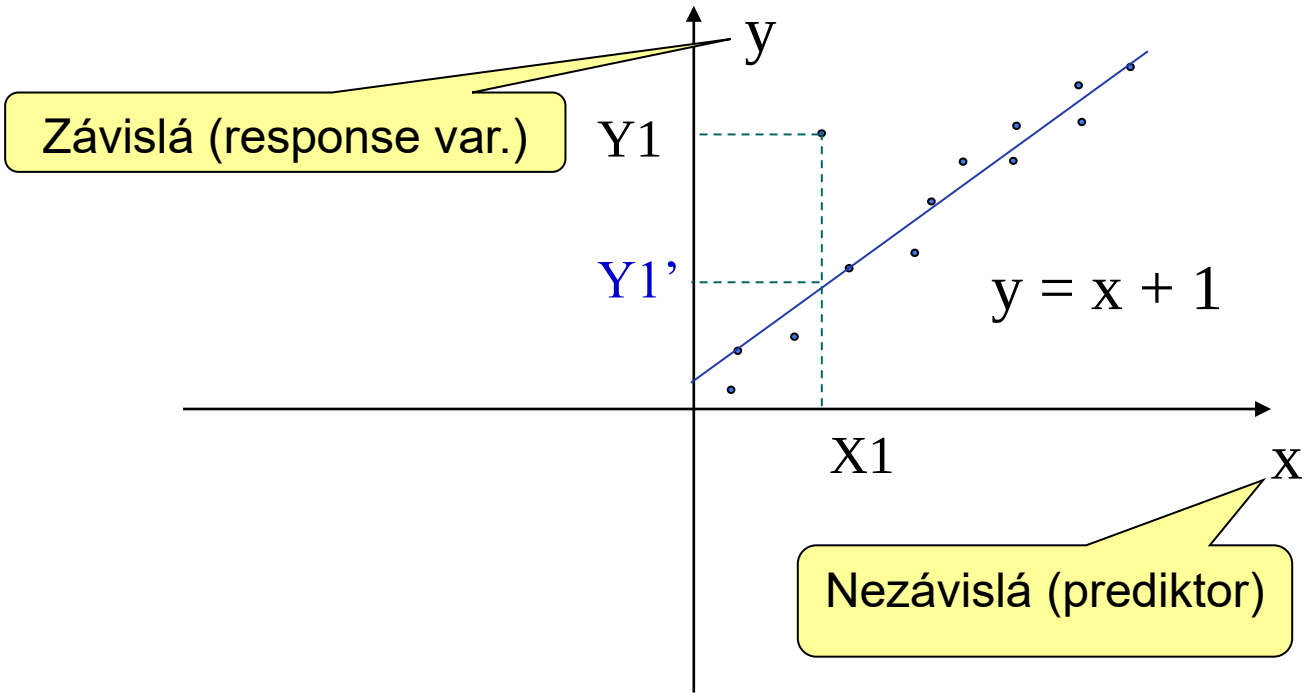
- Koš 1: 12, 12, 12, 12
- Koš 2: 25, 25, 25, 25
- Koš 3: 30, 30, 30, 30

2b) Vyhlazení hraničními hodnotami:

- Koš 1: 4, 4, 21, 21
- Koš 2: 24, 24, 24, 26
- Koš 3: 27, 27, 27, 34

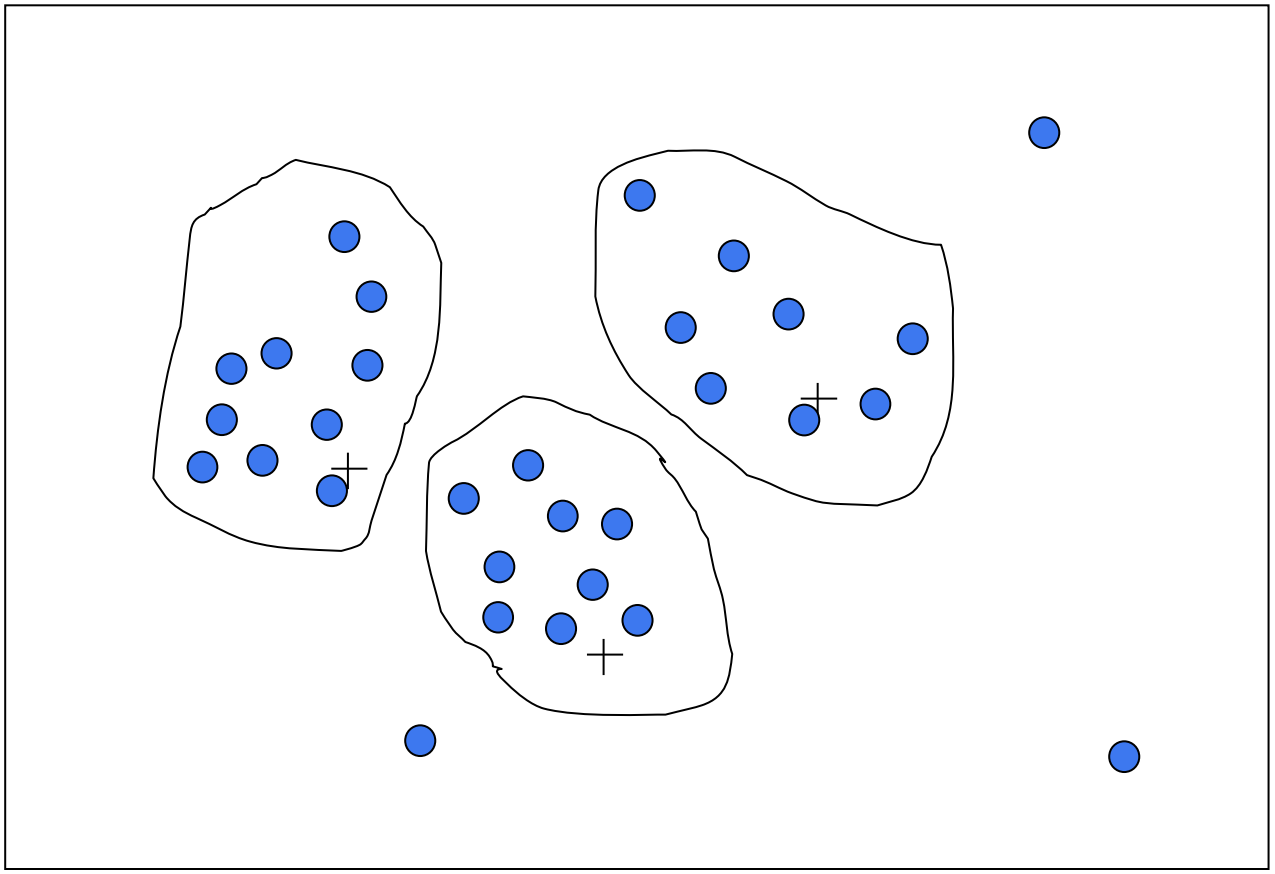
Čištění dat – zašuměná data

- **Regrese:**
  - Hodnota jednoho nebo několika atributů jsou použity k predikci (zde vyhlazení) hodnoty jiného atributu → vyhlazení pomocí regresní funkce
  - Např. lineární, vícenásobná lineární regrese



Čištění dat – zašuměná data

- Analýza odlehlých objektů shlukováním



Čištění dat jako proces

- Čištění dat jako proces – kroky:
  1. Detekce nesrovnalostí
     - Použití výsledků kroku pochopení dat a použití metadat (např.: doména, rozsah, závislosti, rozdělení)
     - Kontrola přetěžování polí
     - Kontrola omezení typu PRIMARY KEY, UNIQUE, NULL/NOT NULL/vyhrazená hodnota pro NULL , souvislé řady (faktury), existující PSČ, …
     - Kontrola specifických integritních omezení (čísla účtů, rodná čísla, existující PSČ, jazyk)
  2. Odstranění nesrovnalostí
  - Typicky potřeba více iterací s interakcí uživatele.

Obsah

- Krok přípravy dat
- Čištění dat
- Integrace dat
- Úprava datové sady
- Transformace dat
- Další úlohy předzpracování

Integrace dat

- Podstata:
  - Kombinace dat z několika zdrojů do jednoho úložiště/souboru.
- Úrovně integrace a řešení konfliktů
  - metadata (schémata):

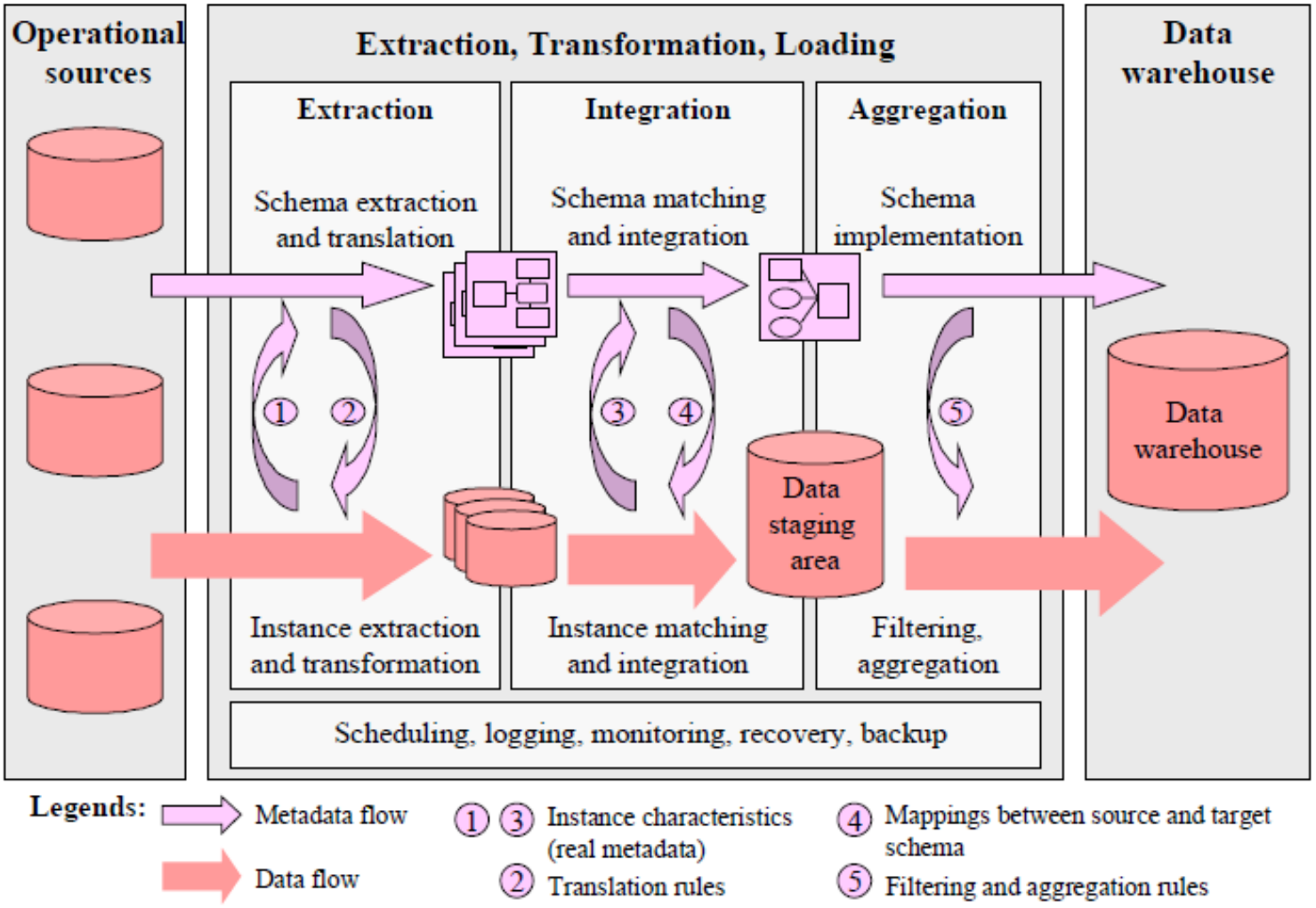
    Příklad: atributy *c_zakaznika* a *zakaznikID* jsou významově totožné
  - data:
    - Problém identifikace entit – tatáž entita reálného světa identifikována v různých zdrojích různě.

      Příklad: rodné číslo vs. osobní číslo
    - Detekce a řešení konfliktů hodnot – pro tutéž entitu jsou hodnoty atributů z různých zdrojů různé.

      Příklad: Jan Novák vs. Novák Jan

      Možné důvody: různé reprezentace, různé formáty, míry
- Integraci řeší i proces ETL u datového skladu

ETL proces


Operational sources
Extraction, Transformation, Loading
Extraction
Integration
Aggregation
Schema extraction and translation
Schema matching and integration
Schema implementation
1
2
3
4
5
Data staging area
Data warehouse
Data warehouse
Instance extraction and transformation
Instance matching and integration
Filtering, aggregation
Scheduling, logging, monitoring, recovery, backup
Legends:
Metadata flow
Data flow
1 3 Instance characteristics (real metadata)
2 Translation rules
4 Mappings between source and target schema
5 Filtering and aggregation rules

Integrace dat

- Řešení redundance (a případné nekonzistence) při integraci
  - Častý problém při integraci z několika zdrojů
    - Identifikace objektů, názvy atributů: Tentýž atribut nebo objekt může mít různá jména ve schématu v různých zdrojích.
    - Odvoditelná data: Jeden atribut může být odvozeným atributem v jiné tabulce, např. celková cena u faktury.
  - Redundantní atributy lze detekovat *korelační analýzou.*
  - Redukce/eliminace redundance a nekonzistence může zvýšit rychlost dolování a zlepšit kvalitu výsledku.
- Integrace zpravidla zahrnuje nejvíc *ruční* práce

Obsah

- Krok přípravy dat
- Čištění dat
- Integrace dat
- Úprava datové sady
- Transformace dat
- Další úlohy předzpracování

Úprava datové sady

- Důvody:
  - Analýza a získávání znalostí rozsáhlých datových sad může být časově velmi náročná → redukce dat.
  - Výrazně se liší zastoupení datových objektů jednotlivých tříd v datech pro klasifikaci → řešení nevyváženosti dat.
- Cíl:
  - Získání redukované reprezentace datového souboru, který je mnohem menší, ale při analýze/získávání znalostí dává stejné (nebo téměř stejné) výsledky jako originální soubor.
  - Získání vyvážené datové sady.

Úprava datové sady

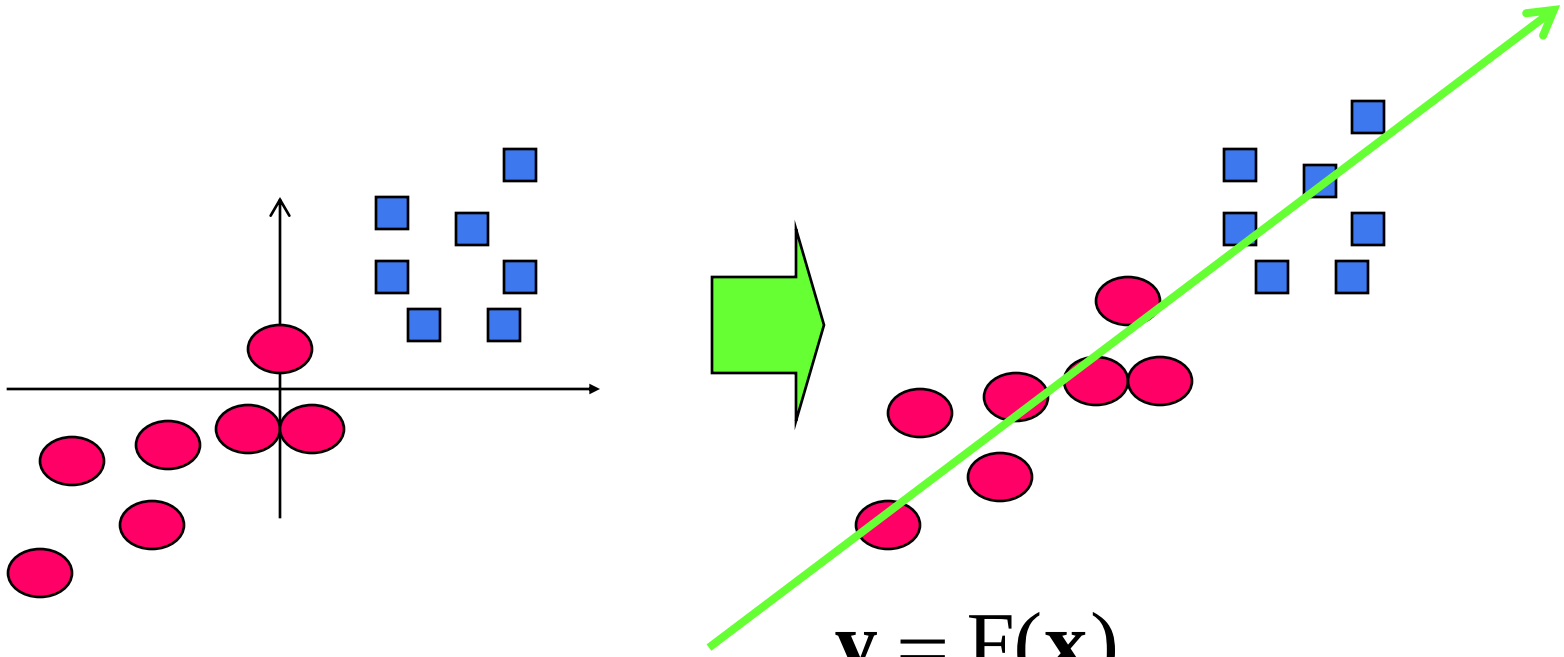
- Strategie redukce dat
  - Redukce dimenzionality
  - Komprese dat
  - Redukce objemu dat (*numerosity*) – parametrické metody (např. regresní – jen parametry), neparametrické (např. shlukování, vzorkování, agregace do datové kostky)
  - Diskretizace a redukce počtu hodnot kategorických atributů (viz transformace dále)

Redukce dimenzionality

- Základní metody
  - Konstrukce rysů/atributů (Attribute Construction)
    - Vytvoření nového atributu z existujících

      Příklad: zboží_kuchyň (místo nábytek, nádobí, spotřebiče)
  - Extrakce rysů/atributů (Feature Extraction)
    - Proces hledání deskriptorů ve zdrojových datech, které nejlépe reprezentují zdrojová data s ohledem na cíl zpracování. Obvykle jde o transformaci do nového prostoru rysů/atributů.
  - Výběr rysů/atributů (Feature Selection)
    - Proces hledání nejlepší podmnožiny rysů/atributů zdrojových dat s ohledem na cíl zpracování nebo kritérium výběru.

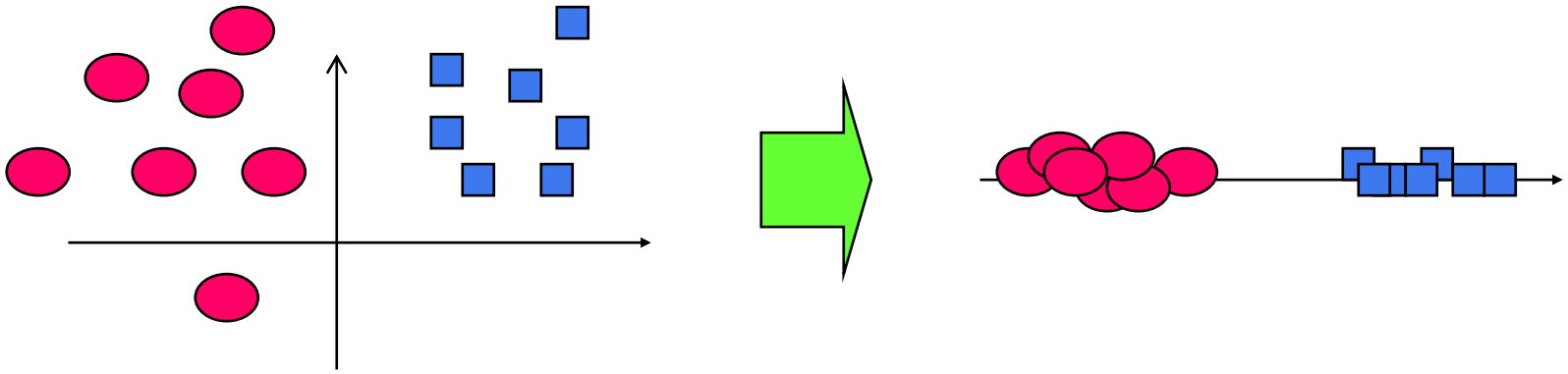
Extrakce rysů/atributů



$$\mathbf{y} = \mathrm{F}(\mathbf{x})$$

$$\dim(\mathbf{y}) < \dim(\mathbf{x})$$

Výběr rysů/atributů



$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

Metoda PCA

- *Principal Component Analysis*
- **Cíl:** snížit dimenzionalitu ortogonální transformací původních n-dimenzionálních datových objektů do nového k-dimenzionálního prostoru ($k <= n$) definovaného tzv. hlavními komponentami splňujícími podmínky:
  - jsou lineárně nezávislé,
  - chyba projekce by měla být co nejmenší.

Metoda PCA

- **Vstup:** datový soubor $[X]_{d*n}$ s *n* objekty a *d* dimenzemi
- **Výstup:** transformovaný soubor $[Y]_{k*n}$

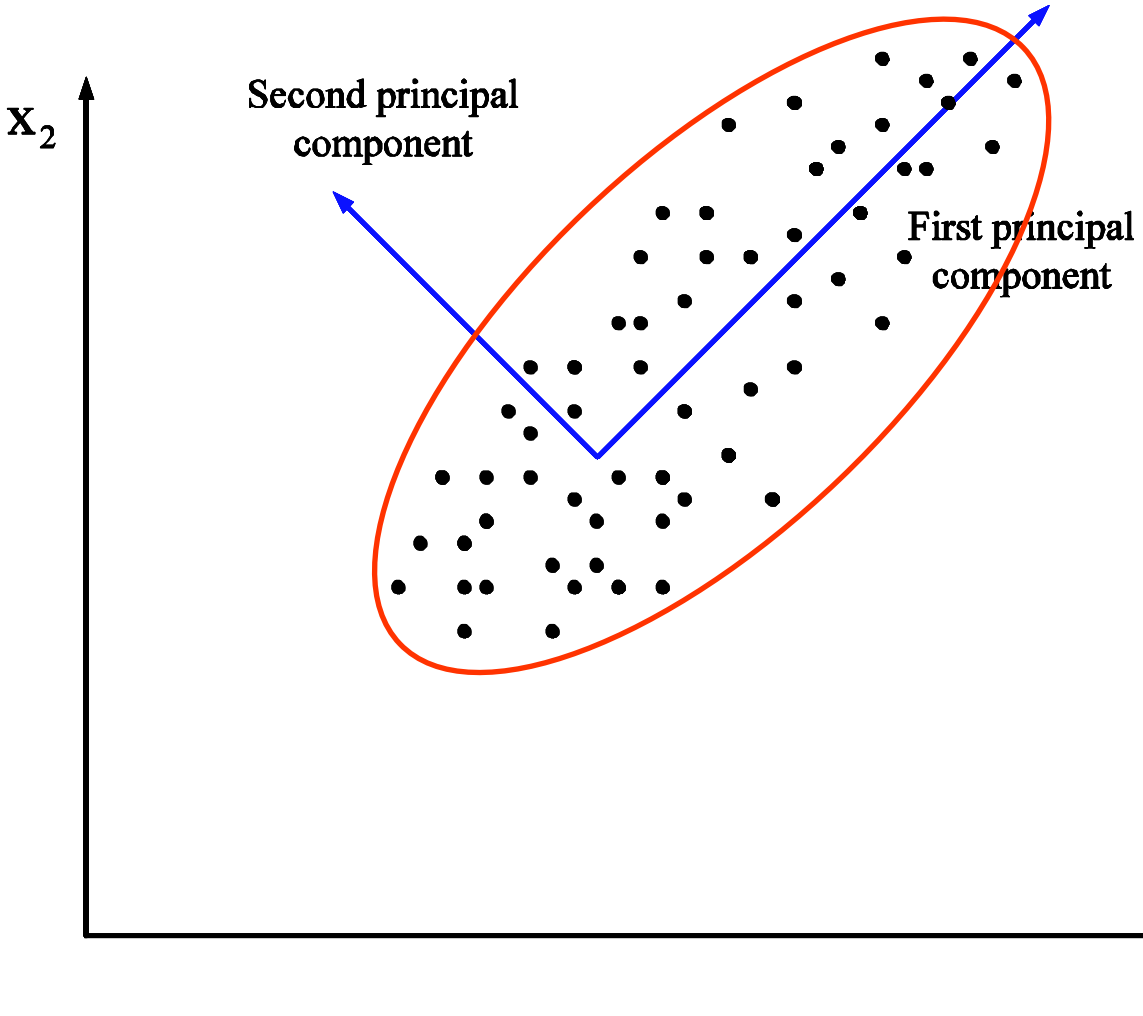
1. Spočítej matici kovariance pro X

$$[C_X] = \frac{1}{n-1}([X] - [\bar{X}])([X] - [\bar{X}])^T$$

$$\begin{bmatrix} V_a & C_{a,b} & C_{a,c} & C_{a,d} & C_{a,e} \\ C_{a,b} & V_b & C_{b,c} & C_{b,d} & C_{b,e} \\ C_{a,c} & C_{b,c} & V_c & C_{c,d} & C_{c,e} \\ C_{a,d} & C_{b,d} & C_{c,d} & V_d & C_{d,e} \\ C_{a,e} & C_{b,e} & C_{c,e} & C_{d,e} & V_e \end{bmatrix}$$

2. Spočítej vlastní čísla a příslušné vlastní vektory.
3. Uspořádej sestupně vlastní vektory dle vlastních čísel.
4. Vyber prvních *k* vlastních vektorů $[EV]_{k*d}$.
5. Transformuj datový soubor $[Y]_{k*n} = [EV]_{k*d}[X]_{d*n}$

Metoda PCA


Second principal component
First principal component
x_2
x_1

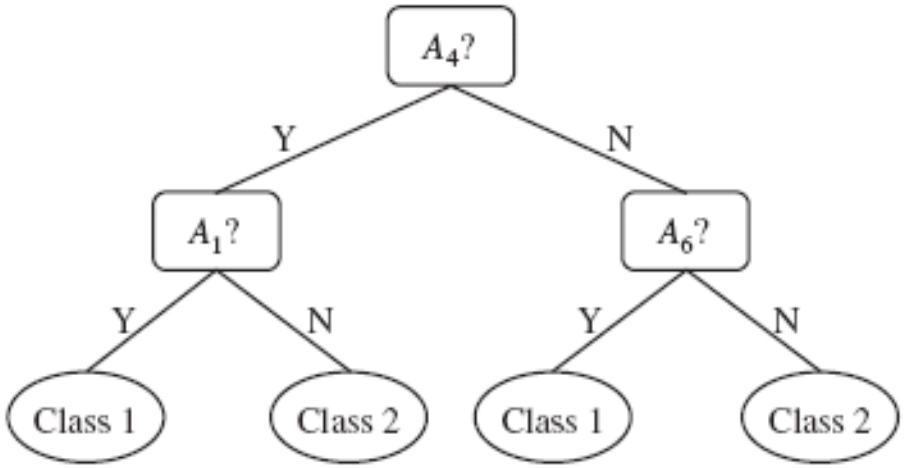
Metoda PCA

- Podrobnosti a vysvětlení lze nalézt např. v [3].
- Další metody redukce dimenzionality
  - SVD – *Singular Value Decomposition* (používaná pro implementaci PCA)
  - LDA – *Linear Discriminant Analysis* – pracuje s informací o třídě objektů → řízená (supervised)
  - ICA – *Independent Component Analysis*

Výběr rysů/atributů

- Cíl:
  - Výběr minimální podmnožiny atributů takové, že rozdělení pravděpodobnosti různých tříd pro zadané hodnoty těchto atributů je co nejbližší původnímu rozdělení při zadaných hodnotách všech atributů.
- Důsledek:
  - Snížení časové náročnosti.
  - Redukce počtu atributů vede na snadnější pochopení.
- Heuristické metody (suboptimální):
  - Postupný výběr/postupná eliminace– typicky greedy (*hladový*) algoritmus → lokální optimalizace, výběr na základě statistické významnosti
  - Kombinace výběru a eliminace
  - Indukce rozhodovacím stromem

Výběr rysů/atributů

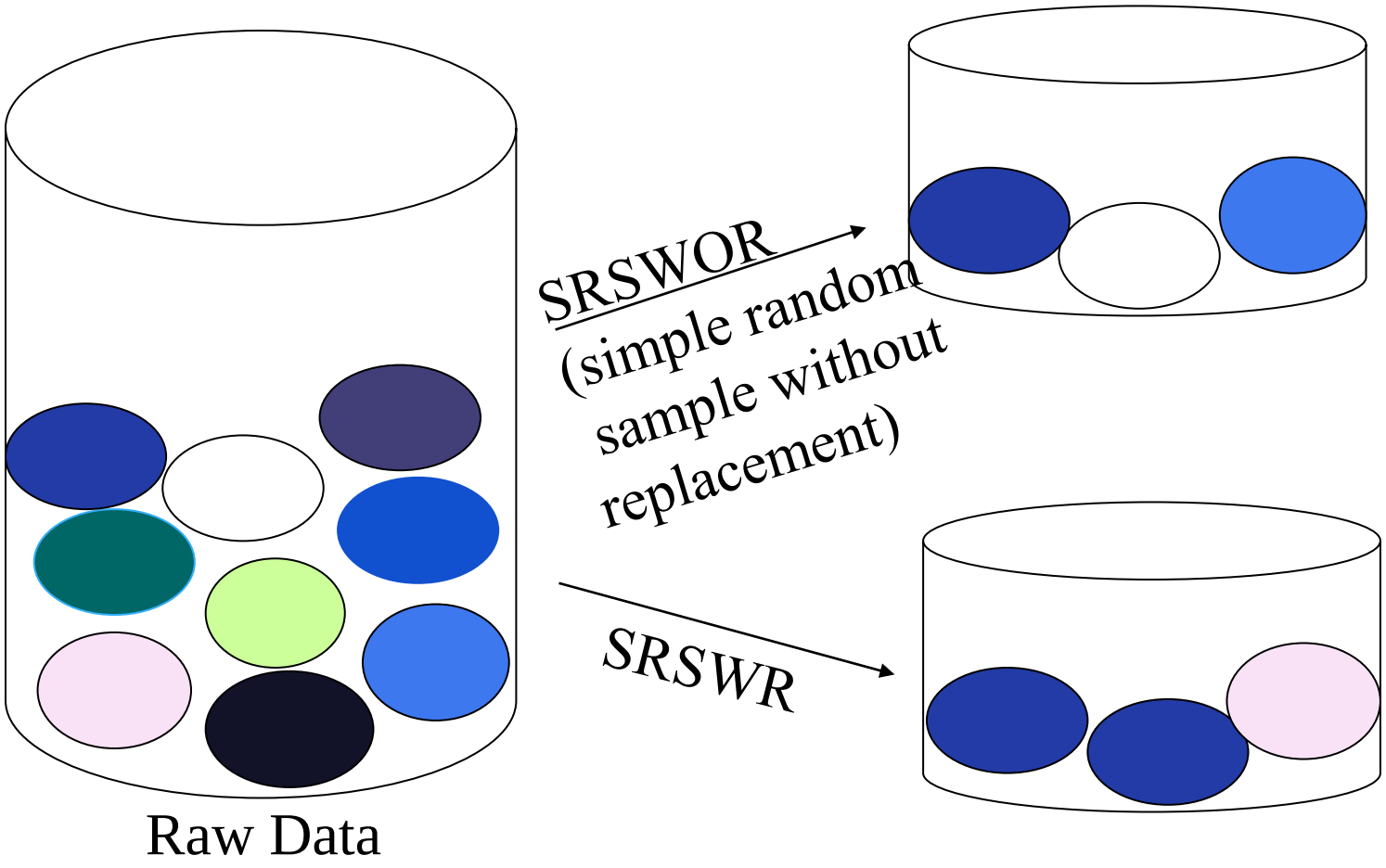
Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: { } => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Redukce dat - vzorkování

- Cíl
  - Získání *malého* vzorku *s* k reprezentaci celého datového souboru s *N* záznamy.
- Výběr reprezentativního vzorku
  - Jednoduché náhodné vzorkování – může být nedostatečné v případě vychýlení dat → adaptivní vzorkování.
  - Adaptivní metody vzorkování
    - Stratifikované vzorkování – aproximuje procentuální zastoupení každé třídy.

Vzorkování

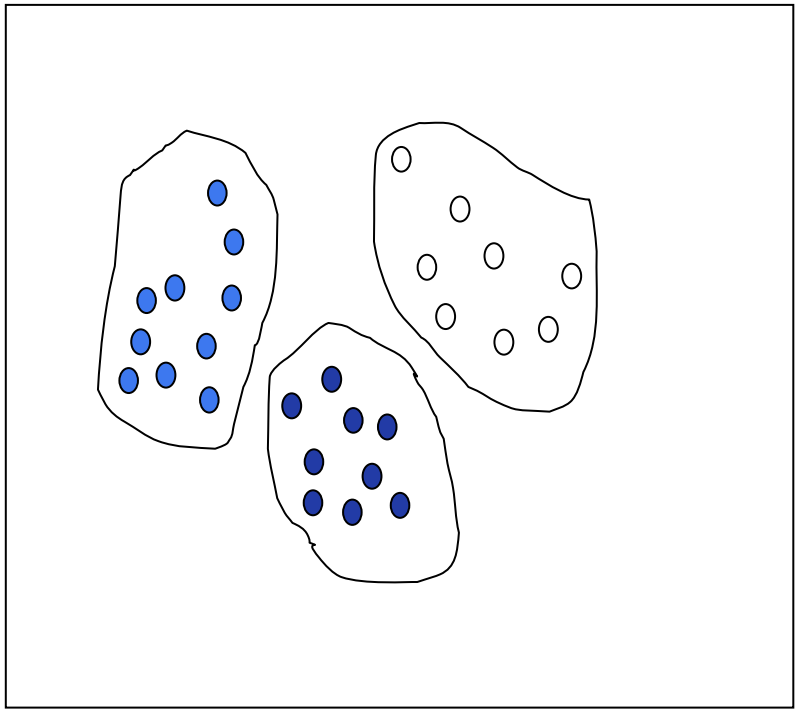
- Jednoduché náhodné vzorkování



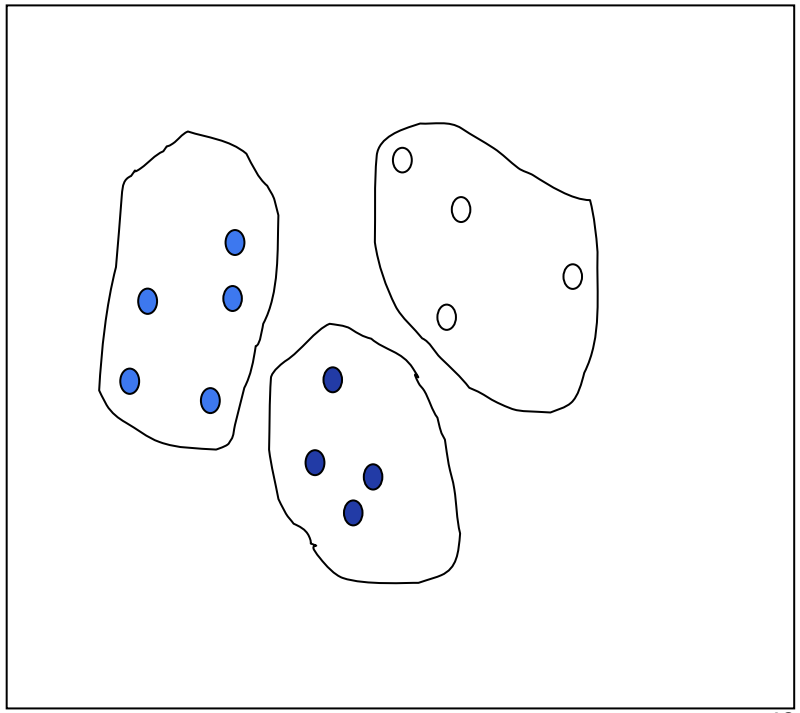
Vzorkování

- Stratifikované vzorkování

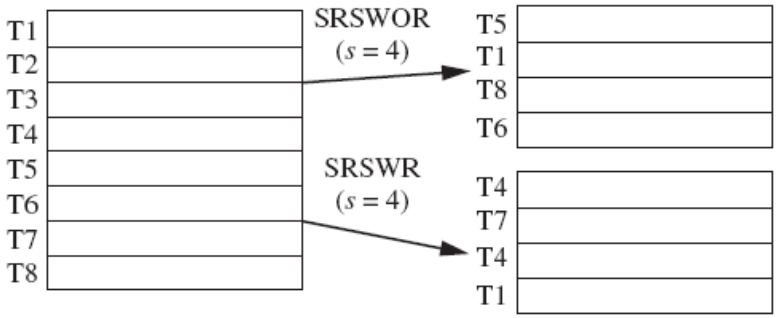
Raw Data

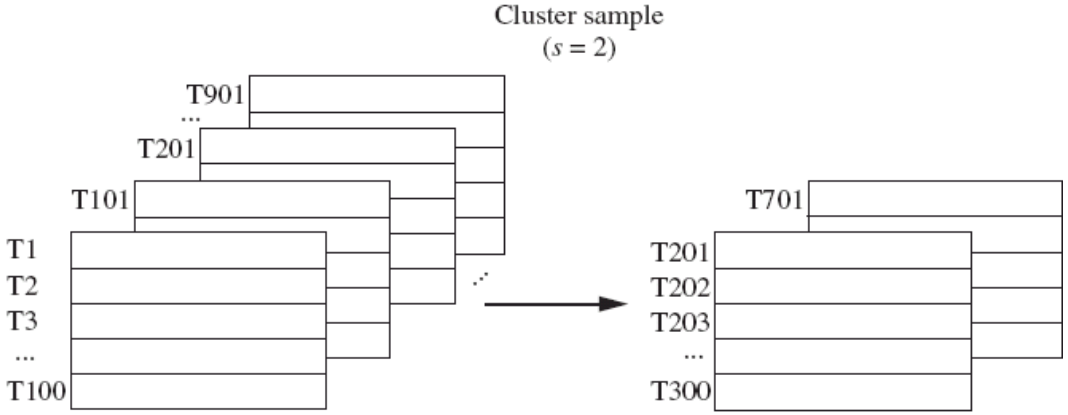


Stratified Sample



Vzorkování





Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

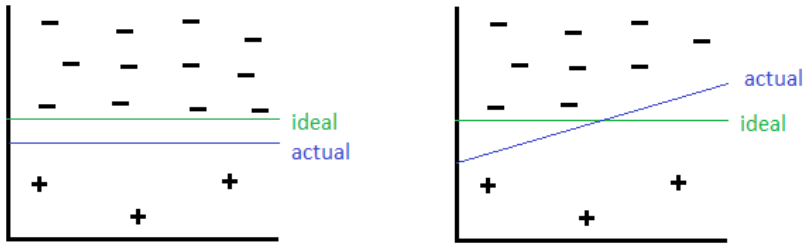
Problém nevyváženosti dat

- Podstata:
  - Řada klasifikačních algoritmů předpokládá vyváženost cílových tříd
  - Řada aplikačních oblastí s nevyváženými daty

    Příklad: Běžné/podvodné transakce, napadení sítě, málo běžné nemoci
  - Důsledek: při trénování je klasifikátor zahlcen daty majoritní třídy a má tendenci minoritní třídu ignorovat → horší výsledky klasifikace
- Řešení:
  - Úprava distribuce dat
  - Zohlednění při výběru atributů/rysů
  - Úprava klasifikátoru
  - Použití sestavy klasifikátorů

Problém nevyváženosti dat

- Úprava distribuce dat:
  - Podvzorkováni (under-sampling) majoritní třídy
    - Náhodné vzorkování

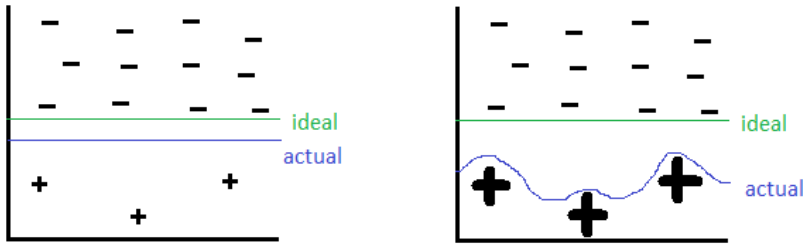


Zdroj: http://www.chioka.in/class-imbalance-problem/

    - Heuristické metody vzorkování – snaží se vyloučit *méně důležitá* data/šum

Problém nevyváženosti dat

- Úprava distribuce dat:
  - Nadvzorkování (over-sampling) minoritní třídy
    - Náhodné s návratem



Zdroj:
http://www.chioka.in/class-imbalance-problem/

    - Heuristické metody, např. SMOTE

Problém nevyváženosti dat

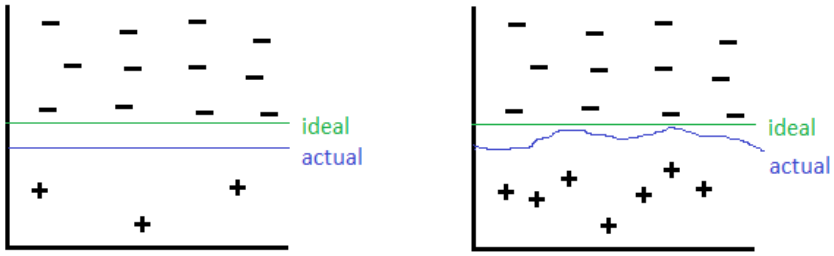
- Úprava distribuce dat:
  - Metoda SMOTE (*Synthetic Minority Over-sampling Technique*)
    - **Idea:** Data blízká datům minoritní třídy budou pravděpodobně patřit do téže třídy
    - **Podstata:** Generování syntetických dat minoritní třídy
    - **Postup:**

FOR EACH instanci minoritní třídy $\bar{c}$ DO
Najdi jeho K nejbližších sousedů
Náhodně vyber jednoho z nich $\bar{k}$
Vytvoř novou instanci minoritní třídy $\bar{n}$ jako

$$\bar{n} = \bar{c} + (\bar{k} - \bar{c}) * rand(0, 1)$$

Problém nevyváženosti dat

- Metoda SMOTE:



Zdroj:
http://www.chioka.in/class-imbalance-problem/

Obsah

- Krok přípravy dat
- Čištění dat
- Integrace
- Úprava datové sady
- Transformace dat
- Další úlohy předzpracování

Transformace dat

- Cíl:
  - Obecně každá úprava hodnot, zde transformace dat do podoby vhodné/nutné pro dolování
- Zahrnuje zejména:
  - Odstraňování šumu
  - Konstrukce atributů
  - Agregace
  - Normalizace
  - Diskretizace hodnot kvantitativního atributu
  - Transformace kategorických dat.
- Velký překryv s jinými úlohami přípravy dat (čištění, redukce dat).

Normalizace

- Min-max normalizace: na [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Příklad: Plat v rozsahu 12 tis. Kč až 98 tis. Kč je normalizován na interval <0.0, 1.0>. Hodnotě 73.6 tis. Kč odpovídá:

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Normalizace

- Z-score normalizace/standardizace (μ: průměr, σ: směr. odchylka):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Příklad: Nechť μ = 54 tis. Kč, σ = 16 tis. Kč. Pak: $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

  - Normalizace může výrazně změnit data → uchovat parametry (zde μ a σ)

- Normalizace změnou dekadického měřítka

$$v' = \frac{v}{10^j}$$ kde *j* je nejmenší celé číslo takové, že max(|v'|) < 1

Normalizace

- Kvantilová normalizace – založená na kvantilech (pořadí uspořádaných hodnot pozorování)

Příklad: Normalizace hodnot exprese mezi mikročipy (převzato i s obrázky ze [6])

1. Uspořádání podle velikosti

hodnoty

Gén	čip1	čip2	čip3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

→

pořadí

Gén	čip1	čip2	čip3
A	iv	iii	i
B	i	i	ii
C	ii	iii	iii
D	iii	ii	iv

→

seřazené

Pořadí	čip1	čip2	čip3
i	B(2)	B(1)	A(3)
ii	C(3)	D(2)	B(4)
iii	D(4)	A(4)	C(6)
iv	A(5)	C(4)	D(8)

Normalizace

- Kvantilová normalizace
- *Příklad* (pokračování):

2. Stanovení referenčního rozložení

seřazené

Pořadí	čip1	čip2	čip3
i	B(2)	B(1)	A(3)
ii	C(3)	D(2)	B(4)
iii	D(4)	A(4)	C(6)
iv	A(5)	C(4)	D(8)

→

referenční mikročip

Pořadí	Ref. hodnota
i	(2 1 3)/3 = 2.00
ii	(3 2 4)/3 = 3.00
iii	(4 4 6)/3 = 4.67
iv	(5 4 8)/3 = 5.67

3. Vlastní normalizace referenčními hodnotami

hodnoty

Gén	čip1	čip2	čip3
A	5(iv)	4(iii)	3(i)
B	2(i)	1(i)	4(ii)
C	3(ii)	4(iii)	6(iii)
D	4(iii)	2(ii)	8(iv)

referenční mikročip

Pořadí	Ref. hodnota
i	(2 1 3)/3 = 2.00
ii	(3 2 4)/3 = 3.00
iii	(4 4 6)/3 = 4.67
iv	(5 4 8)/3 = 5.67

→

normalizované hodnoty

Gén	čip1	čip2	čip3
A	5.67	4.67	2.00
B	2.00	2.00	3.00
C	3.00	4.67	4.67
D	4.67	3.00	5.67

Diskretizace

- Podstata:
  - Rozdělení rozsahu kvantitativního atributu na intervaly, přiřazení návěští intervalu
- Důvody/přínosy:
  - Některé klasifikační algoritmy vyžadují pouze kategorické atributy.
  - Dochází k redukci dat.
  - Příprava pro další analýzu
- Přístupy:
  - S učitelem/bez učitele – využívá informaci o třídě?
  - Shora-dolů/zdola-nahoru.

Diskretizace

- Metody:
  - Plnění
  - Analýza histogramu – případně rekurzivní dělení/slučování → konceptuální hierarchie.
  - Shlukování - shora-dolů/zdola-nahoru → konceptuální hierarchie
  - Založené na rozhodovacím stromu – hierarchie podle hodnot atributu pro štěpení (minimální entropie) – informace o třídě.

Transformace kategorických dat

- Kódování – převod na kvantitativní atribut
  - Důvod: některé algoritmy získávání znalostí (např. klasifikační SVM, KNN) pracují jen s numerickými hodnotami.
  - Metody:
    - *ručně* – typicky pro ordinální atributy (známe uspořádání)
    - automaticky řetězec → celé číslo – nahrazuje v pořadí výskytu ve sloupci (LabelEncoder, OrdinalEncoder)
    - 1 z N – kategorická hodnota → sloupec; pro nominální, případně ordinální; nárůst dimenzí.
    - Binární – (řetězec) → celé číslo → dvojkové vyjádření → sloupec pro každou pozici dvojkového vyjádření; pro nominální, případně ordinální; nárůst dimenzí menší za cenu ztráty informace (překryv).
    - Hašované – podobné jako 1 z N, ale mapované na M<N.
    - Jiné.

Transformace kategorických dat

- Redukce - např. *N* nejčetnějších hodnot + zbytek do jedné kategorie

  Příklad: Oracle Data Mining (ODM) – *N* a jméno pro ostatní.
- Generování konceptuálních hierarchií
  - Specifikace množiny atributů a jejich částečného uspořádání z hlediska hierarchie uživatelem/expertem

    Příklad: Adresa: ulice < město < kraj < stát
  - Manuálně zařazením hodnot do hierarchie

    Příklad: Adresa: Brno → Jihomoravský kraj
  - Specifikace pouze množiny atributů bez jejich částečného uspořádání + automatická odvození částečného uspořádání

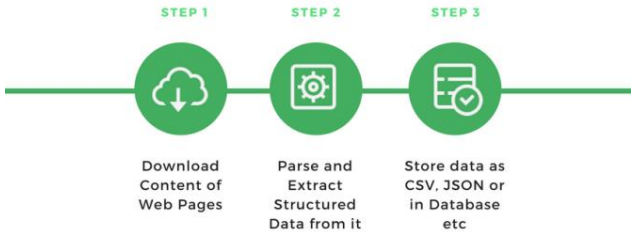
    Příklad: {ulice, město} ?

Obsah

- Krok přípravy dat
- Čištění dat
- Integrace
- Úprava datové sady
- Transformace dat
- Další úlohy předzpracování

Specifické předzpracování

- Příklady
  - Dolování v textu
    - Techniky zpracování přirozeného jazyka → vektor rysů reprezentující dokument
    - Vkládání slov (word embedding) využitím NN
  - Zpracování obrazu
    - Extrakce lokálních rysů (SIFT, SURF) → vektor rysů reprezentující obrázek
  - Extrakce dat z webu (Web scraping)
    - Extrakce strukturovaných dat z webových stránek (viz přednáška dr. Burgeta)



Literatura (1)

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Third Edition. Elsevier Inc., 2012, ISBN 978-0-12-381479-1, s. 83 – 123.
2. Zendulka a kol.: Získávání znalostí z databází. Studijní opora, FIT VUT v Brně, 160 s., 2009. (elektronicky) Dostupná mezi soubory k předmětu. (s. 45 – 57).
3. Kumar, R.: Understanding Principal Component Analysis. Dostupné na https://medium.com/@aptrishu/understanding-principle-component-analysis-e32be0253ef0.
4. What is web scraping - Part 1 - Beginner's guide. Dostupné na https://www.scrapehero.com/a-beginners-guide-to-web-scraping-part-1-the-basics/ .

Literatura (2)

5. Guo, X. et al.: On the Class Imbalance Problem. Fourth International Conference on Natural Computation, 2008. Dostupné na https://www.researchgate.net/publication/228612392_On_the_Class_Imbalance_Problem.
6. Budínská, E.: Analýza genomických a proteomických dat. Dostupné na https://portal.matematickabiologie.cz/index.php?pg=analyza-genomickych-a-proteomickych-dat--analyza-genomickych-a-proteomickych-dat.