# Strojové učení a dolování dat – přehled

**Jiří Kléma**

Katedra kybernetiky,
FEL, ČVUT v Praze

http://ida.felk.cvut.cz

# Osnova přednášek

| Přednáška | Učitel | Obsah |
|---|---|---|
| 1. | J. Kléma | Úvod do předmětu, učení s a bez učitele. Shluková analýza, formalizace. |
| 2. | J. Kléma | Shluková analýza, EM algoritmus, k-means, hierarchické shlukování. |
| 3. | J. Kléma | Spektrální, konceptuální, fuzzy shlukování. Dvojshlukování. |
| 4. | J. Kléma | Časté množiny položek, algoritmus Apriori, asociační pravidla. |
| 5. | J. Kléma | Časté posloupnosti, epizodální pravidla, modely posloupností. |
| 6. | J. Kléma | Časté podstromy/podgrafy. |
| 7. | J. Kléma | Učení z textů a webu, aplikace. |
| 8. | F. Železný | Výpočetní teorie učení, konceptový prostor, PAC učení. |
| 9. | F. Železný | PAC učení logických forem. |
| 10. | F. Železný | Nekonečné konceptové prostory. |
| 11. | F. Železný | Empirické odhady rizika. |
| 12. | F. Železný | Induktivní logické programování, nejmenší zobecnění, inverze důsledku. |
| 13. | F. Železný | Učení z logických interpretací, relační rozhodovací stromy, relační rysy. |
| 14. | F. Železný | Statistické relační učení, markovská logika. |

Učení bez učitele. Deskriptivní modely.

Symbolické učení – koncepty.

Induktivní a statistické učení logických forem.

## Základní pravděpodobnostní značení

| | |
|---|---|
| $P_X$ | Rozdělení pravděpodobnosti (hustota) na spočetné (resp. nespočetné) množině $X$. |
| $P_X(x)$ | Hodnota $P_X$ pro konkrétní prvek $x \in X$. |
| $P_{X,Y}$ | Rozdělení sdružené pravděpodobnosti (hustota) na $X \times Y$. |
| $P_{X,Y}(x, y)$ | Hodnota $P_{X,Y}$ pro konkrétní prvky $x$ a $y$. |
| $P_{X\|Y}$ | Rozdělení podmíněné pravděpodobnosti, tj. $P_{X\|Y} = P_{X,Y}/P_Y$. |
| $P_{X\|Y}(x\|y)$ | Hodnota $P_{X\|Y}$ pro konkrétní prvky $x$ and $y$. |
| $\Pr(\text{expression})$ | Pravděpodobnost události určené výrazem, například $a = 1 \wedge b = 2$, typicky vyčíslena z příslušných rozdělení $P_A(1)P_{B\|A}(1\|2)$. |

# Učení bez učitele

:: Předpoklady:

- Existuje instanční prostor $X$
  - reálné vektory, grafy, sekvence, relační struktury, ...
- Existuje pravděpodobnostní hustota $P_X$ na $X$

:: Vstup:

- Konečný *vzorek* ($m \in N$)

$$S = \{x_1, x_2, \ldots, x_m\}$$

  generovaný i.i.d. z $P_X$.
  - $S$ je multimnožina, prvky nazýváme *příklady*.

:: Cíle:

- Obecný: naučit se $P_X$: *úloha odhadu hustoty*, nebo
- Speciální: nauč se něco o $P_X$: *učení variety (manifoldu)*

# Odhad hustoty pravděpodobnosti

:: Neparametrický

- Nemáme k dispozici apriorní znalost o $P_X$
- Obecně nezvládnutelný problém
  - lze jen v případech, že $P_X$ je jednoduchá a/nebo $m$ je velmi velké.

:: Parametrický, např.

- Směs multivariátních gaussovských rozdělení
  - $X = R^n$
  - počet gaussiánů je předem známý
  - učíme se parametry: středy $\vec{\mu}$ a kovarianční matice $\Sigma$
- bayesovské sítě
  - obvykle $X = \{0, 1\}^n$ (tj., náhodné události)
  - známy vztahy podmíněné nezávislosti mezi proměnnými (graf)
  - učíme se parametry: tabulky podmíněných pstí v uzlech grafu (CPT's)
- etc.

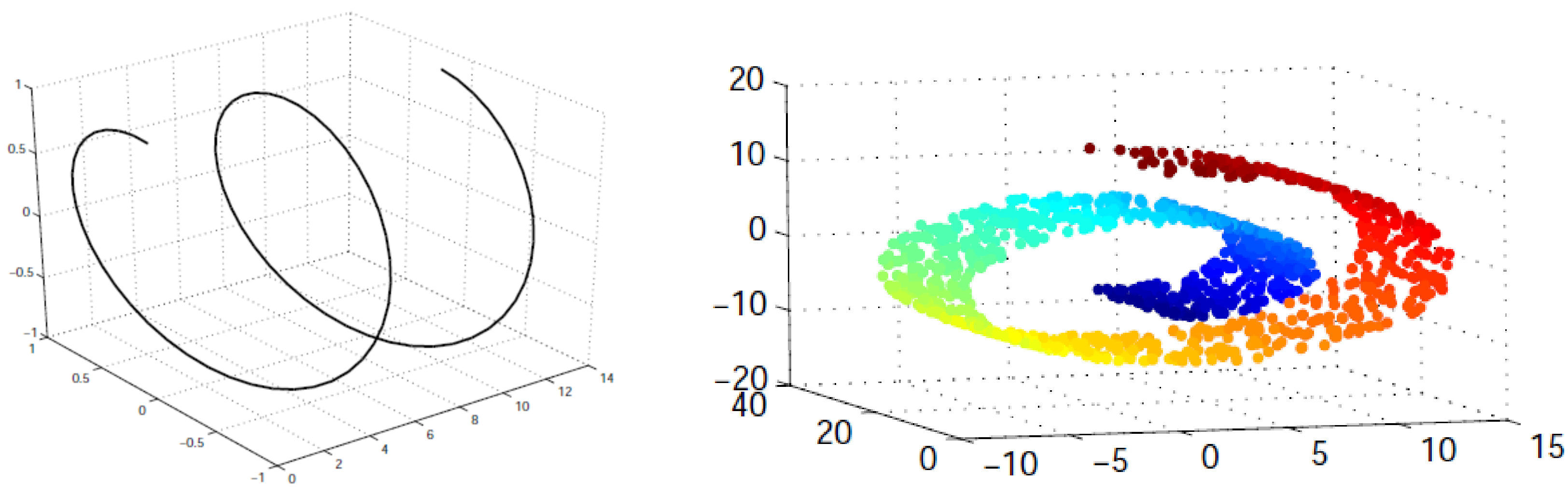
## Učení variety (manifoldu)

:: Varieta (manifold)

- topologický prostor lokálně podobný euklidovskému, globálně typicky nelineární,

:: Učení

- Identifikace topologického prostoru nižší dimenze zanořeného v prostoru dimenze větší,
- s následnou možnou projekcí do prostoru dimenze manifoldu – nelineární redukce dimenze,
- lineární analogií je PCA nebo vícedimenzionální škálování.



Cayton: Algorithms for Manifold Learning.

# Učení variety (manifoldu) – příklady

:: Redukce dimenzionality

- Lineární – PCA, vícedimenzionální škálování
- Nelineární – kernel PCA, locally linear embedding
- V čem spočívá učení? Zjednodušení problému, transformace odkrývá strukturu manifoldu.

:: Shlukování

- Hledáme oblasti s vysokou $P_X$
- Oblasti jsou vyjádřeny explicitně (přiřazením příkladů)

:: Hledání vzorů

- Vzory definují manifoldy v $X$ s nečekaně vysokou $P_X$
- Časté množiny položek, podgrafy, podsekvence, ...
- *Jak* vzory definují manifoldy?

# Učení s učitelem

:: Předpoklady:

- Existuje instanční prostor $X$
  - reálné vektory, grafy, sekvence, relační struktury, ...
- Existuje stavový prostor $Y$
  - také různé druhy, ale obvykle podmnožina $R$
  - Existuje pravděpodobnostní hustota $P_{XY}$ na $X \times Y$

:: Vstupy:

- Konečný *vzorek* ($m \in N$)

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

generovaný i.i.d. z $P_{XY}$. $S$ je multimnožina, prvky nazýváme *příklady*.

:: Cíle?

## Učení s učitelem: cíle

:: Nejobecnější cíl, chci umět odpovídat na libovolnou otázku

- učení $P_{XY}$
  - v zásadě shodná třída metod jako pro učení $P_X$

:: Nejčastější cíl, chci umět usuzovat na skrytý stav $y$ na základě pozorování $x$

- učení $P_{Y|X}$
  - Jde o speciálnější úlohu než učení $P_{XY}$. Proč?

:: Můj odhad stavu nemusí mít charakter distribuce, stačí mi odhad nejpravděpdoobnějšího stavu

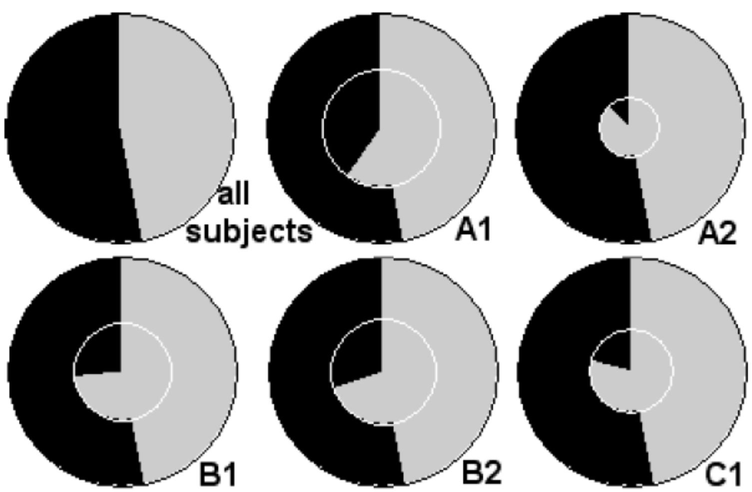
- $f : X \to Y$ such that
$$f(x) = \arg\max_{y \in Y} P_{Y|X}(y|x)$$
  - Jde o speciálnější úlohu než učení $P_{Y|X}$. Proč?

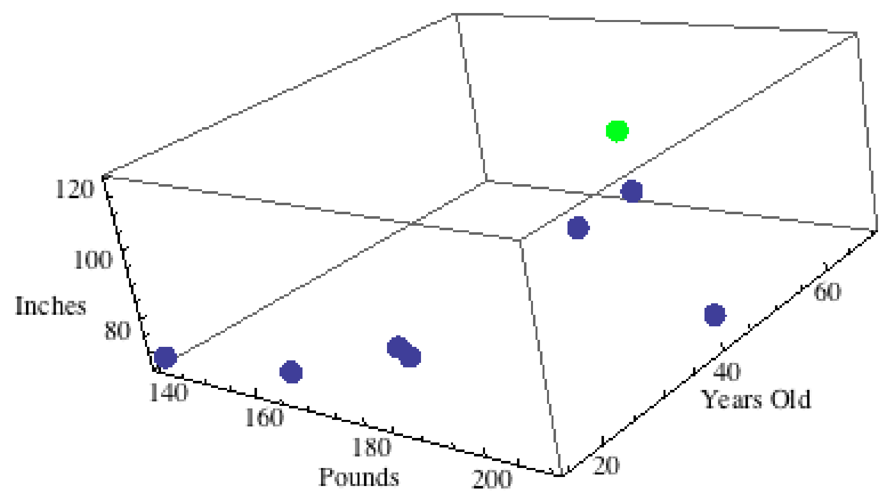
## Dolování dat

- Co to je?
  - Velkým úkolem dnešního (biologického) bádání je dospět od **dat** ke **znalostem**. Znám lidi, kteří si myslí, že data už jsou znalosti; ti se však pro změnu pídí po tom, jak dospět od znalostí k pochopení (Sydney Brenner – The Scientist, 2002).
  - Dolování dat je aplikací algoritmů pro extrakci **smysluplných vzorů**.
- Srovnání s učením
  - používají v zásadě podobné postupy,
  - důraz na srozumitelnost, originalitu a praktickou využitelnost,
  - blíže praxi (technologie více než věda).
- Jednotící teorie
  - $T = \{\phi \in \mathcal{L} \mid \mathrm{q}(D, \phi) \text{ je pravda}\}$
  - $\mathcal{L}$ . . . formální jazyk (spočetná množina formulí),
  - predikát $q$ vyhodnocuje kvalitu formule $\phi \in \mathcal{L}$ vzhledem ke vstupním datům $D \subseteq X$,
  - $T$ reprezentuje znalost získanou z $D$, formulím $\phi \in T$ říkáme vzory.

# Deskriptivní modely

- slouží ke **zhuštěnému popisu dat**, zjednodušeně zachycují obecné závislosti,
- kategorizace popisných modelů
  - na co se soustředí – tvoří globální model dat?
    - vyhledávání dominantních struktur
      - detekce podskupin, segmentace, shlukování, asociace,
    - vyhledávání nugetů, detekce odchylek
      - podvodné operace, síťové útoky, závadné www stránky,



# Deskriptivní modely

- kategorizace popisných modelů (pokračování)
  - jaký typ modelů využívají?
    - pravděpodbnostní modely – popis dat pomocí pravděpodobnostního rozdělení,
      - parametrické, neparametrické, směsi rozdělení,
    - **symbolické** modely – data interpretují konceptuálně na základě pojmů a jejich vztahů,
      - grafy, pravidla, taxonomie, logické vazby,
      - charakteristika: zřetelně a lidsky srozumitelně vyjadřují znalost,
    - kombinované modely
      - mj. grafické pstní modely – bayesovské sítě, markovské modely,
  - s jakými vstupními daty pracují?
    - číselná data, symbolická data, texty,
    - atributová reprezentace, relační databáze,
    - časová a sekvenční data.

# Použití deskriptivních modelů

- privátní sektor
  - banky, pojišťovny, obchodní firmy,
  - snížení nákladů, zvýšení prodejů, průzkum trhu, odhalení podvodů,
- veřejný sektor
  - veřejná správa, lékařství, zpravodajské služby,
  - efektivita, zamezení ztrát a podvodům.

## Interakce s jinými předměty

- přímá prerekvizita
  - A4B33RPZ – Rozpoznávání a strojové učení,
    - (ne)bayesovské rozhodování, minimalizace ztráty z rozhodnutí za neurčitosti,
    - statistické učení – lineární, kNN, SVM, neuronové sítě,
    - odhady parametrů z dat – věrohodnost, EM algoritmus,
- souvislosti
  - A4B33ZUI – Základy umělé inteligence,
  - A0B01LGR – Logika a grafy,
  - A4B33FLP – Funkcionální a logické programování,

  - AE4M33GMM – Boris Flach, Graphical Markov Models.