Vytěžování dat, přednáška 1: Úvod

Filip Železný

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Vytěžování dat: základní myšlenky

$$1, 2, 4, 8, ?$$

Co následuje?

$$16$$

Odpovídá vzoru

$$\begin{aligned} x_1 &= 1 \\ x_k &= 2x_{k-1} \ (k \geq 2) \end{aligned}$$

Vytěžování dat = hledání srozumitelných vzorů v datech

Vzor vs. Data

Odvozování dat ze vzorů:

$x_k = 2x_{k-1}$ ⟶  ⟶ $1, 2, 4, 8, \ldots$

Vytěžování dat

$1, 2, 4, 8, \ldots$ ⟶  ⟶ $x_k = 2x_{k-1}$

vzor $\approx$ předpis $\approx$ model $\approx$ hypotéza $\approx$ teorie $\approx \ldots$

Využití vzorů

$$x_k = 2x_{k-1}$$

- Predikce ($x_5 = 16, x_6 = 32, \ldots$)
- Zlepšení rozhodování

- Interpretace člověkem (vzor vyjadřuje znalost)
- Porozumění procesům

Vytěžování dat (Data Mining)

„Data Mining je netriviální proces identifikace pravdivých, dosud neznámých, potenciálně využitelných a zcela srozumitelných vzorů v datech" (Fayyad)

Používá techniky oborů

- statistika
- strojové učení (umělá inteligence)
- databázové technologie
- vizualizace dat

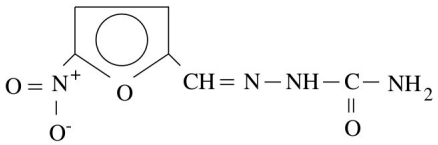
Reálné příklady vytěžování: Asociace v nákupních košících

pivo	párky	horčice	pleny	...
+	−	−	+	
+	+	+	−	
−	+	−	−	

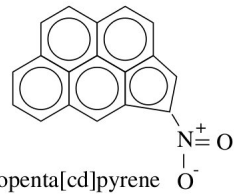
(atd.)

pleny $\rightarrow$ pivo

Reálné příklady vytěžování: Predikce karcinogenity

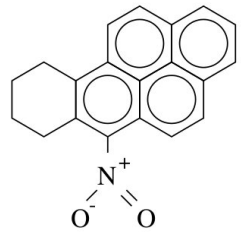


nitrofurazone

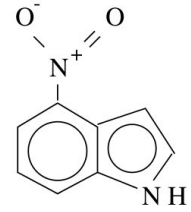


4-nitropenta[cd]pyrene

karcinogenní

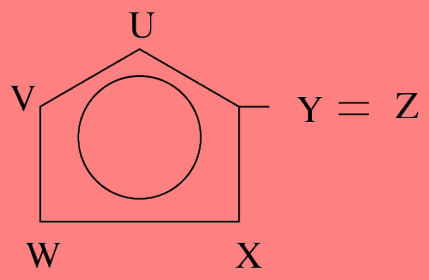


6-nitro-7,8,9,10-tetrahydrobenzo[a]pyrene



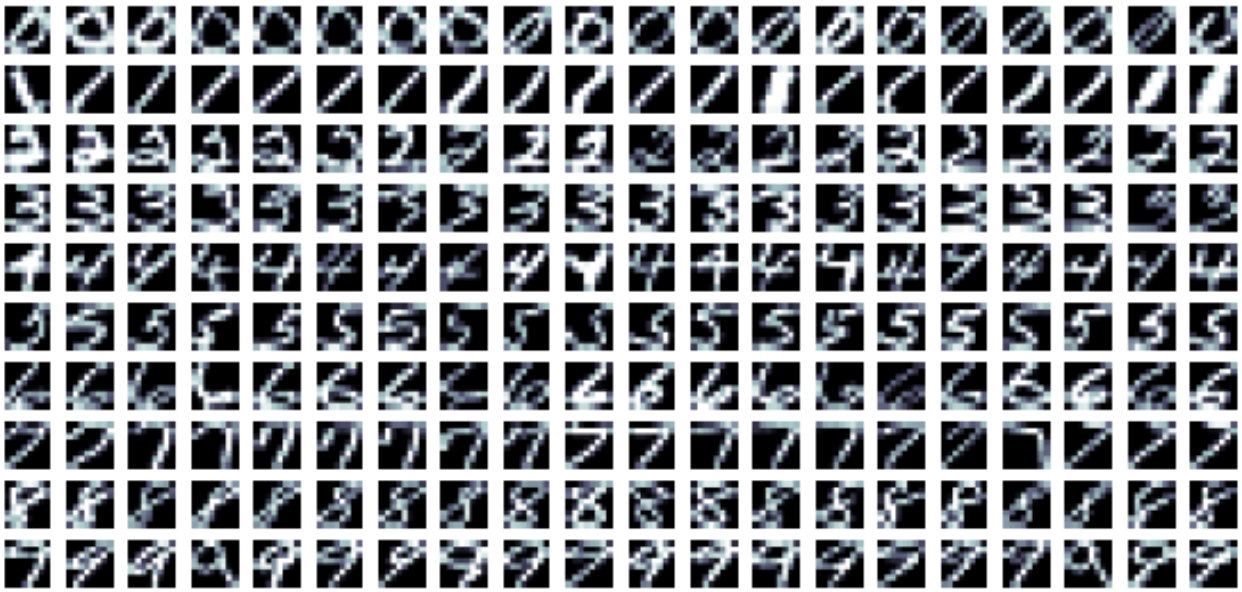
4-nitroindole

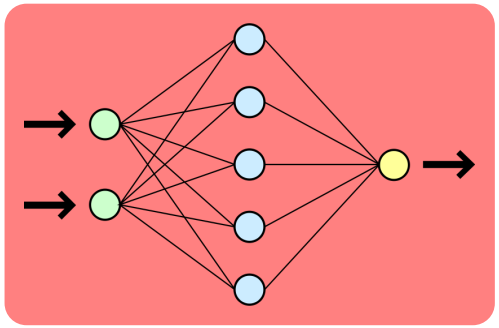
kontrolní



v karcinogenních

Reálné příklady vytěžování: Rozpoznávání obrazu





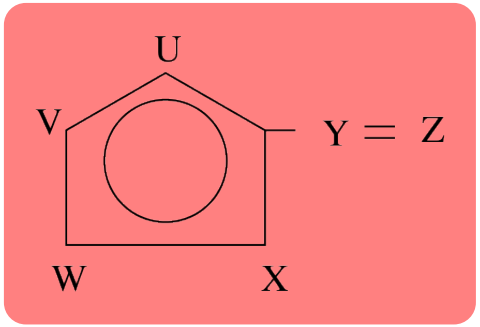
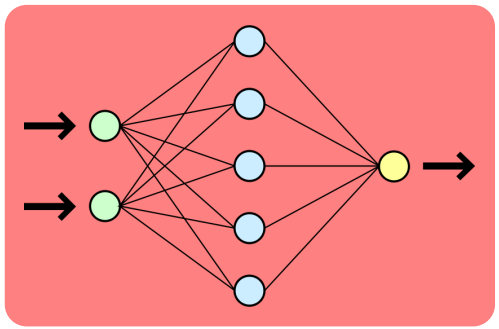
6

Struktura a parametry vzoru

Vzory jsou rozličných druhů (rovnice, pravidla, grafy, ...).

Rozlišujeme jejich

- diskrétní *strukturu*
- reálné *parametry*

struktura	parametry
$x_k = ax_{k-1}$	a
pleny $\rightarrow$ pivo	(žádné)
	(žádné)
	váhy synapsí $\mathbf{W}$

Parametrické vs. neparametrické metody

Parametrické metody:

- Struktura je zadána předem
- Úkolem vytěžování je vyhledat hodnoty parametrů $\vec{a} \in R^n$ maximalizující soulad s daty

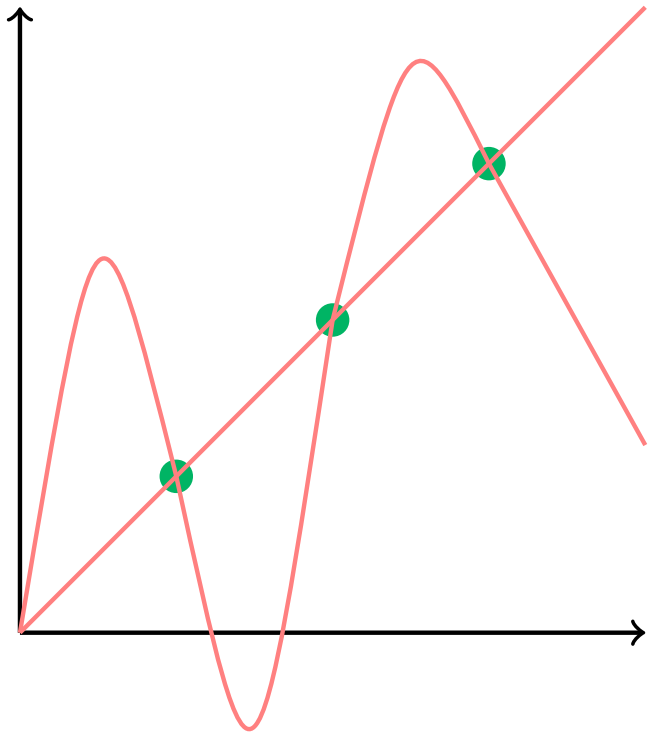
Neparametrické metody:

- Hledá se struktura i případné parametry
- Struktura se hledá v nějaké předepsané konečné množině struktur $\mathcal{S}$

Zaujetí algoritmu

Čím menší n resp. $\mathcal{S}$, tím větší *zaujetí* vytěžovacího algoritmu. Větší zaujetí = menší flexibilita = menší možnost adaptace na data.

Zaujetí

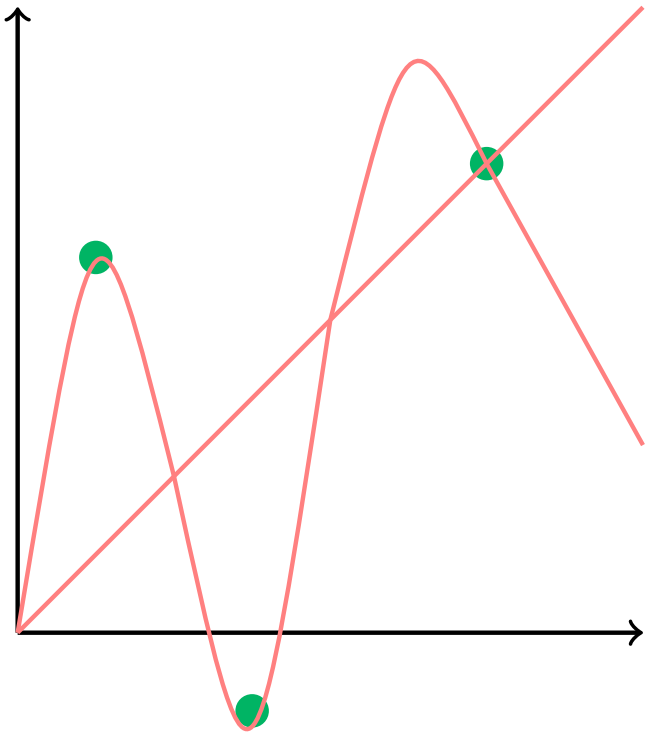


$y = a_1 x$

$y = a_1 x + a_2 x^2 + a_3 x^3$

příliš slabé zaujetí, velká flexibilita

Zaujetí



$y = a_1 x$

příliš silné zaujetí, malá flexibilita

$y = a_1 x + a_2 x^2 + a_3 x^3$

Vytěžování dat: definice 1

Úloha nalezení množiny vzorů ϕ

Vstupní data

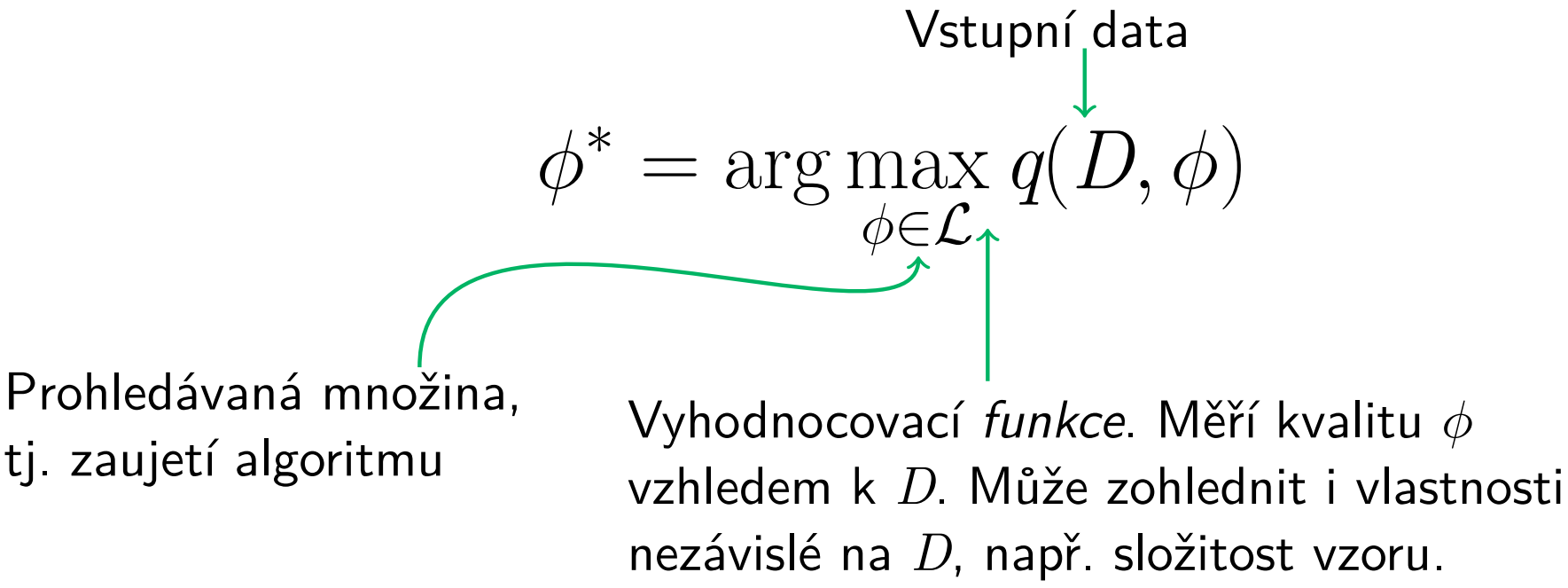
$$\{\phi \in \mathcal{L} \mid \text{platí } \mathrm{q}(D, \phi)\}$$

Prohledávaná množina, tj. zaujetí algoritmu. Obsahuje všechny přípustné vzory.

Vyhodnocovací predikát. Platí, pokud ϕ je v souladu s D

Vytěžování dat: definice 2

Úloha nalezení nejlepšího vzoru ϕ^*



Pro parametrické metody: hledáme optimální parametry pro zadanou strukturu S. Zde tedy $\mathcal{L} = \{(S, \vec{a}) \mid \vec{a} \in R^n\}$

Jak hledat vzory

Konkrétní techniky: v dalších přednáškách

Obecně:

- Parametrické metody: optimální parametry lze někdy vyjádřit a spočítat analyticky, jindy prohledávání $\mathcal{L}$ ("pokus-omyl")
- Neparametrické metody: téměř vždy prohledávání

Volba zaujetí

Konkrétní techniky: v dalších přednáškách

Obecně:

- Volíme dle
  - typu dat (grafy, vektory reálných čísel, ...)
  - toho, co se chceme dozvědět
  - dostupnosti algoritmů (různá zaujetí - různé algoritmy)
- Čím více o datech předem víme, tím silnější zaujetí můžeme stanovit
  - Např. data leží na přímce $\Rightarrow$ $y = ax$, hledáme jen a
- *Occamova břitva*: "fungují-li" dva vzory stejně dobře na datech, preferujeme ten jednodušší

Generativní modely

$$x_1 = 1$$
$$x_k = 2x_{k-1} \ (k \geq 2)$$

pleny $\rightarrow$ pivo

- *Generativní (též globální)* model
- Vzor jednoznačně určující, jak generovat data

- Ostatní (lokální) vzory fungují jako omezující podmínky
- Nejsou předpisem pro generování dat

Data: statistické předpoklady

Budou nalezené vzory platit i v budoucích datech? Neplatí ve vytěžených datech jen náhodou?

Je třeba zavést statistické předpoklady na data:

- Uvažujeme množinu všech možných instancí X
  - obsahy nákupních košíků, grafy molekul, řádky v relační tabulce, ...
- Pravděpodobnostní rozdělení P_X na X
- Data: D je multimnožina

$$D = \{x_1, x_2, \ldots x_m\} \ (m \in N)$$

  prvky vybrány náhodně a navzájem nezávisle z P_X
- Vzor se bude používat na tomtéž X a P_X.

Data: statistické předpoklady (pokr.)

- Nalezené *generativní* vzory pak aproximují P_X, např.

$$P_X(x) = N(\mu, \sigma)$$

- Jiné typy nalezených vzorů zachycují určité vlastnosti P_X, např.

$$\text{pivo} \rightarrow \text{pleny}$$

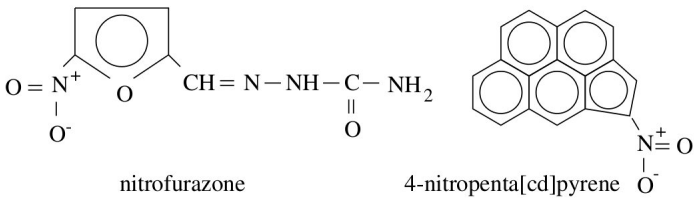
  říká, že $P_X(x)$ je malá pro instance x v nichž implikace neplatí.
- Takto definované úloze říkáme učení (vytěžování) *bez učitele* (terminologie strojového učení)

Cvičení

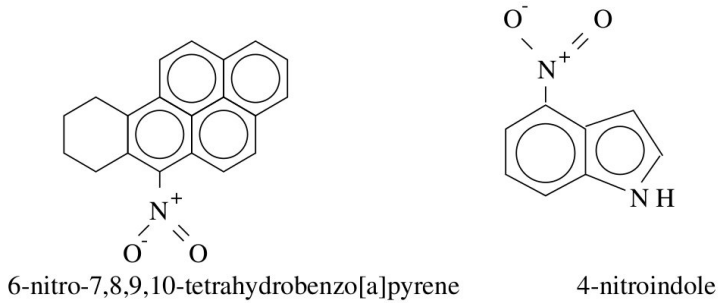
Jak za právě definovaných statistických předpokladů zformulovat úlohu vytěžování z úvodu přednášky (str. 2)?

Učení s učitelem

Speciální, zvláště častá úloha vytěžování: najít vzor pro předpovídání *cílové veličiny* (třídy) instance ze zbylého popisu instance. Příklad:



karcinogenní



kontrolní

Vedle X uvažujeme ještě Y: množinu hodnot cílové veličiny. Zde

- X: struktury molekul (grafy)
- $Y = \{\text{karcinogenní}, \text{ kontrolní}\}$

Učení s učitelem (pokr.)

Předpoklady na data:

- Pravděpodobnostní rozdělení P_{XY} na $X \times Y$
- Data: D je multimnožina

$$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_2)\} \ (m \in N)$$

  prvky vybrány náhodně a navzájem nezávisle z P_{XY}
- Vzor se bude používat na tomtéž X, Y a P_{XY}

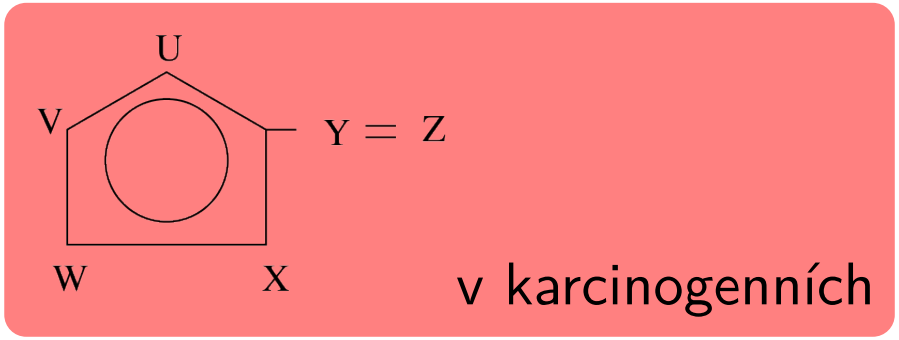
Obvykle hledáme vzory aproximující

- podmíněnou pravděpodobnost $P_{Y|X}(y|x) = P_{XY}(x, y)/P_X(x)$
- nebo nejpravděpodobnější hodnotu $\arg\max_{y \in Y} P_{Y|X}(y|x)$

pro zadané $x \in X$.

Učení s učitelem: příklad

Např. nalezený vzor



interpretujeme jako

$$P_{Y|X}(y, x) = \begin{cases} 1 \text{ je-li tato struktura v } x \\ 0 \text{ jinak} \end{cases}$$

resp.

$$y = \begin{cases} \text{karcinogenní, je-li tato struktura v } x \\ \text{jinak kontrolní} \end{cases}$$

Učení s učitelem (pokr.)

Proč "učení s učitelem"? Terminologie strojového učení.

- "Učitel" poskytuje x i y prostřednictvím dat D (trénovací data)
- "Žák" (algoritmus) se učí odvozovat y z x (trénovací fáze)
- Potom učitel zadává pouze x a žák odhaduje y pomocí naučených vzorů

Příznakový popis

Většina známých vytěžovacích algoritmů umí pracovat pouze s daty v *příznakovém* popisu. Předpokládají, že

$$X = X_1 \times X_2 \times \ldots \times X_n \ (n \in N)$$

kde každé X_i je obor hodnot příznaku i, např.

- $X_i = R$ (reálná čísla)
- $X_i = \{\text{muž}, \text{žena}\}$ (kategorie)

tedy pouze "jednoduché" datové typy, ne struktury, grafy apod.

Některé algoritmy dále omezují přípustné typy příznaků, např. pouze numerické, pouze kategorické ('nominální'), pouze binární ...

Příznakový popis (pokr.)

Data $D = \{x_1, x_2, \ldots x_m\}$ $(m \in N)$, $x_i \in X$ $(1 \leq i \leq m)$ jsou tedy *n-ticemi* hodnot příznaků.

- Příklad:

	věk	**pohlaví**	**kuřák**	**rakovina**
x_1:	56	muž	$+$	$+$
x_2:	32	žena	$-$	$-$
x_3:	48	žena	$+$	$+$
x_4:	60	muž	$+$	$+$

Příznaková data tedy tvoří matici odpovídající jedné tabulce relační databáze.

Jednotlivé instance jsou její řádky.

Příznakový popis (pokr.)

Co s daty, která nejsou v příznakovém popisu?

- grafy
- relační struktury
- signály (např. zvuk)
- obrazy
- číselné řady
- texty

Použít specializovaných algoritmů

- graph mining, text mining, induktivní logické programování, počítačové vidění, ...

Převést na příznakovou reprezentaci

- nelehký úkol, je třeba zachovat podstatnou informaci.

(mimo rozsah tohoto kursu)

Přehled přednášek

1. Úvod
2. Odhad parametrů Gaussovské směsi, EM algoritmus
3. Grafické pravděpodobností modely
4. Shluková analýza
5. Samoorganizující se mapy
6. Časté množiny a asociační pravidla
7. Klasifikační úloha, klasifikace dle podobnosti a Bayesovská
8. Rozhodovací stromy a pravidla
9. Lineární a polynomiální klasifikace
10. Perceptron a neuronové sítě s dopřednou strukturou
11. Testování modelů
12. Kombinování modelů a výběr příznaků
13. Rezerva (aplikace vytěžování dat)

Bez učitele S učitelem Parametrické Neparametrické