Истраживање података

Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

Увод

- Енглески термин: *Data mining*
- Уводни део
  - Садржај курса
  - Процес истраживања података
  - Подаци - врсте и типови
  - Основне технике ИП

Садржај курса

- Увод у истраживање података
  - Основни подаци и дефиниције
  - Типови и квалитет података
  - Мере сличности и различитости
- Процес припреме података
- Визуелизација резултата
- Класификација података
  - Основни концепти и алгоритми
  - Алтерантивни методи класификације
  - Алгоритми и перформансе
  - СВМ (машине са потпорним векторима)

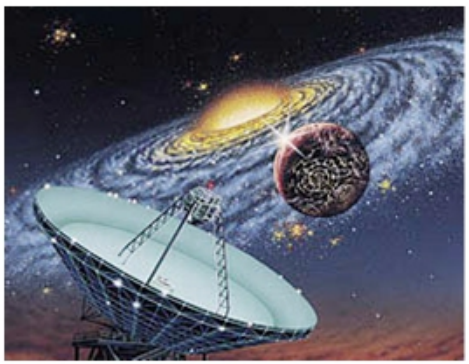
Садржај курса (наставак)

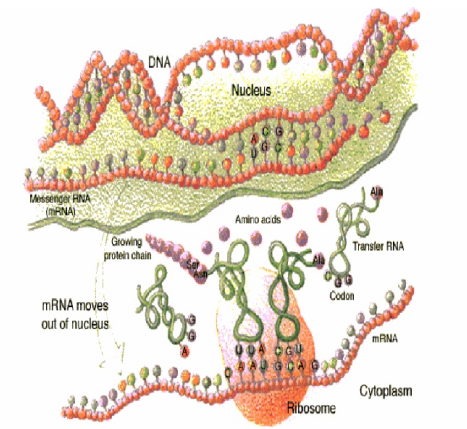
- Кластер анализа
  - Основне технике
  - Напредне технике
- Правила придруживања
  - Основни концепти и алгоритми
  - Напредни концепти и алгоритми
- Додаци - кратке напомене
  - Димензиона редукција
  - Откривање аномалија
  - Истраживање текстуалних података
  - Примена истраживања података
  - ПММЛ (енг. *Predictive Model Markup Language)*

Зашто истраживање података

Потреба:

- стално се прикупљају велике количине података
- различите области: наука, инжењерство, медицина, пословне примене, ...
- велика количина *равних* података за обраду





Зашто истраживање података (наставак)

- снажнији рачунари
- добијање концизних и употребљивих информација за одређени циљ
- пословне апликације - квалитетније информације - предност у односу на конкуренцију

Зашто истраживање података (наставак)

- због количине и просторно-временске природе података традиционалне методе за анализу нису погодне за употребу
- постоје 'сакривене' информације које нису одмах (или лако) уочљиве
- традиционалне методе - велики део података никада и не стиже до анализе
- ...

Дефиниција истраживања података

Не постоји строга дефиниција. Најчешће се користе

- процес који укључује прикупљање података, њихово чишћење, обраду, анализу и добијање корисних сазнања о њима
- проналажење скривених информација у бази података

Дефиниција истраживања података

- нетривијално издвајање имплицитних, претходно непознатих и потенцијално корисних информација из база података
- интегрални део откривања знања у базама података (енг. *Knowledge Discovery in Databases, KDD*)

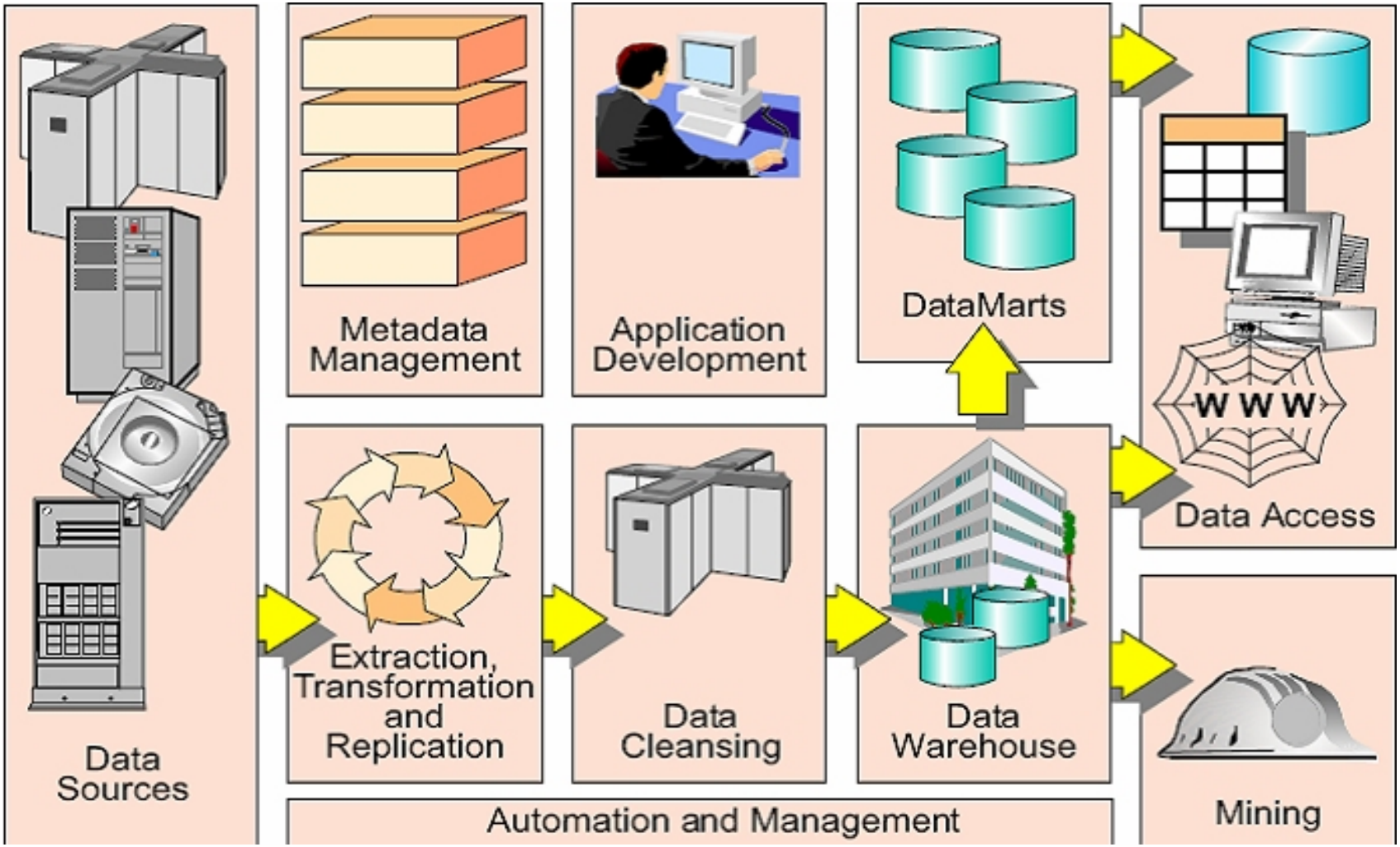
Порекло истраживања података

- Статистика
- Вештачка интелигенција
- Машинско учење
- Препознавање образаца
- Технологије база података
- Паралелно и дистрибуирано рачунарство

Карактеристике истраживања података

- Алгоритамска заснованост
- Сваки алгоритам покашава да укалупи податке у неки модел
- Бира се модел који је најближи карактеристикама података
- Подаци су у базама података
- Потребна знања из области база података, машинског учења, вероватноће и статистике, вештачке интелигенције, препознавања образаца, програмирања, ...

Фазе процеса ИП



Изазови истраживања података

- Велики број проблема и могућих решења
- Нема готових рецепата за добијање резултата
- Велики број различитих формата/типова података који утичу на могуће решење
- Интерпретација резултата
- Визуелизација резултата
- Скалабилност (велика количина улазног материјала)
- Димензионалност (велики број атрибута)
- Сложени и хетерогени подаци (мултимедија,...)
- Квалитет података

Изазови истраживања података

- Нарушавање приватности
- Профилисање (нпр. шта ако је неко погрешно квалификован?)
- Неауторизовано коришћење - заштита тајности информација
- Потреба за обрадом података који се непрекидно прикупљају - поток података (енг. *data stream*)
- Однос према великим подацима (енг.*big data*)

Шта су подаци

- Подаци (енг. *data set*)
- Скуп објеката и њихових атрибута
- Атрибути су својство или карактеристика објекта
- Пример: температура, боја аута, величина екрана, итд.
- Атрибути су познати и као променљиве, поља, особине, карактеристике, ...
- Скуп атрибута описује објекат
- Објекат - користе се и термини слог, пример, ентитет, инстанца, ...

Вредност атрибута

Вредности атрибута су бројеви или симболи који су придружени атрибуту

- Разлика између атрибута и њихових вредности
  - Исти атрибути могу да буду пресликани у различите вредности атрибута
  - Пример: дужина може да се мери у метрима или километрима
- Различити атрибути могу да буду пресликани у исти скуп вредности, при чему особине вредност атрибута могу да буду различите
  - Пример: вредности за број година и тежину су целобројне, али број година не може да се смањује док тежина може

Дискретни атрибути

Атрибути могу да буду описани и преко броја вредности које садрже

Дискретни атрибути

- Имају коначан број или пребројиво бесконачан скуп вредности
- Пример: поштански бројеви, рачуни, скуп речи у неком документу
- Често се приказују као целобројне променљиве
- Бинарни атрибути су специјалан случај дискретних атрибута

Континуирани атрибути

Континуирани (непрекидни, континуални) атрибути

- Скуп вредности ових атрибута чине реални бројеви
- Пример: температура, висина, тежина, притисак, брзина
- Реалне вредности могу да се мере и представљају само преко коначног броја цифара
- Уобичајено - реални бројеви у покретном зарезу

Асиметрични атрибути

Једино се присуство не-нула вредности сматра значајним

- На пример, нека је објекат студент чији су атрибути информација да ли је слушао неки од курсева који се држе на факултету (1-слушао, 0-није слушао)
- Када су два студента слична по курсевима које су слушали?

Бинарни атрибути код којих су битне не-нула вредности се зову асиметрични бинарни атрибути

Типови атрибута према операцијама

Тип атрибута може да се одреди према операцији која може да се примени на атрибут

Врста операције	Рбр	Операција	Тип атрибута
Различитост	1	$=$ i $\neq$	Именски (1)
Уређење	2	$<, \leq, >$ i $\geq$	Редни (1,2)
Адитивност	3	$+$ i $-$	Интервални (1,2,3)
Мултипликативност	4	$\times$ i $\setminus$	Размерни (1,2,3,4)

Категорички атрибути: именски и редни

Непрекидни атрибути: интервални и размерни

Независни подаци

Међусобно независни подаци - најчешће мултидимензионални или текстуални

Индекс	Име	Презиме	Датум уписа	Датум рођења	Место рођења
20140021	Милош	Перић	06.07.2014	20.01.1995	Београд
20140022	Маријана	Савковић	05.07.2014	11.03.1995	Краљево
20130023	Сања	Терзић	04.07.2013	09.11.1994	Београд
20130024	Никола	Вуковић	04.07.2013	17.09.1994	
20140025	Маријана	Савковић	06.07.2014	04.02.1995	Краљево
20140026	Зорица	Миладиновић	06.07.2014	08.10.1995	Врање
20130027	Милена	Станковић			

Мултидимензионални подаци

- Најједноставнији облик независних података
- Састоје се од слогова (инстанци, трансакција, ентитета, торки, објекта, вектора-особина,...)
- Слог се састоји од поља (атрибути, димензије, ...)
- Примери
  - Квантитативни (температура, висина, тежина, притисак, брзина)
  - Категорички (ЈМБГ, боја очију, поштански број, радно место)
  - Бинарни
  - Текстуални

Мултидимензионални подаци

x_1^1	x_1^2	x_1^3	x_1^4	...	x_1^d
x_2^1	x_2^2	x_2^3	x_2^4	...	x_2^d
x_3^1	x_3^2	x_3^3	x_3^4	...	x_3^d
x_4^1	x_4^2	x_4^3	x_4^4	...	x_4^d
x_5^1	x_5^2	x_5^3	x_5^4	...	x_5^d
x_6^1	x_6^2	x_6^3	x_6^4	...	x_6^d
...	...	...	...	...	...
x_n^1	x_n^2	x_n^3	x_n^4	...	x_n^d

Дефиниција 1. Скуп мултидимензионалних података $\mathscr{D}$ представља скуп од n слогова $\overline{X_1}, ..., \overline{X_n}$ таквих да сваки од слогова $\overline{X_i}$ садржи скуп од d особина означених са $(x_i^1, ..., x_i^d)$

Ретки подаци

x_1^1		x_1^3	x_1^4	...	
x_2^1				...	
	x_3^2			...	x_3^d
			x_4^4	...	
x_5^1		x_5^3		...	
	x_6^2			...	
...	...	...	...	...	...
	x_n^2			...	x_n^d

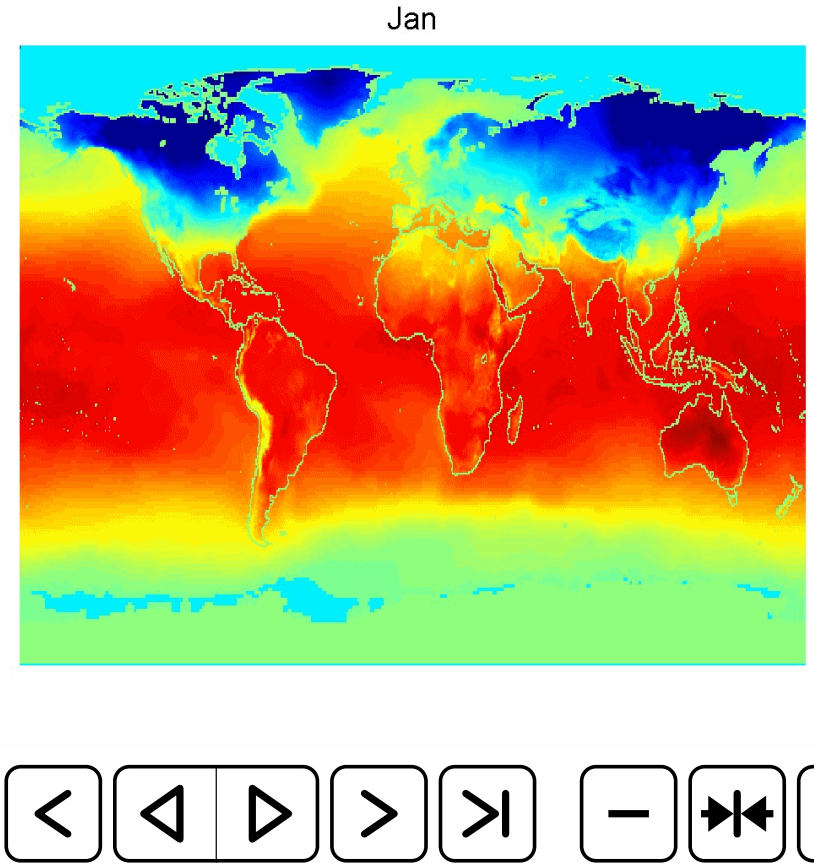
Међусобно зависни подаци

Међусобно зависни подаци - имплицитна или експлицитна зависност између података

Индекс	Име	Презиме	Датум уписа	Датум рођења	Место рођења
20140021	Милош	Перић	06.07.2014	20.01.1995	Београд
20140022	Маријана	Савковић	05.07.2014	11.03.1995	Краљево
20130023	Сања	Терзић	04.07.2013	09.11.1994	Београд
20130024	Никола	Вуковић	04.07.2013	17.09.1994	
20140025	Маријана	Савковић	06.07.2014	04.02.1995	Краљево
20140026	Зорица	Миладиновић	06.07.2014	08.10.1995	Врање
20130027	Милена	Станковић			

Подаци везани за временске серије

- Садрже вредности добијене непрекидним мерењем кроз време
- Имплицитна зависност од претходних мерења (npr. ЕКГ, мерење температуре, ...



Просторни подаци

- Одређују просторне локације
- Атрибути (најчешће два) који одређују простор су контекстуални
- Други атрибути (којих може да буде више) могу да моделирају понашање

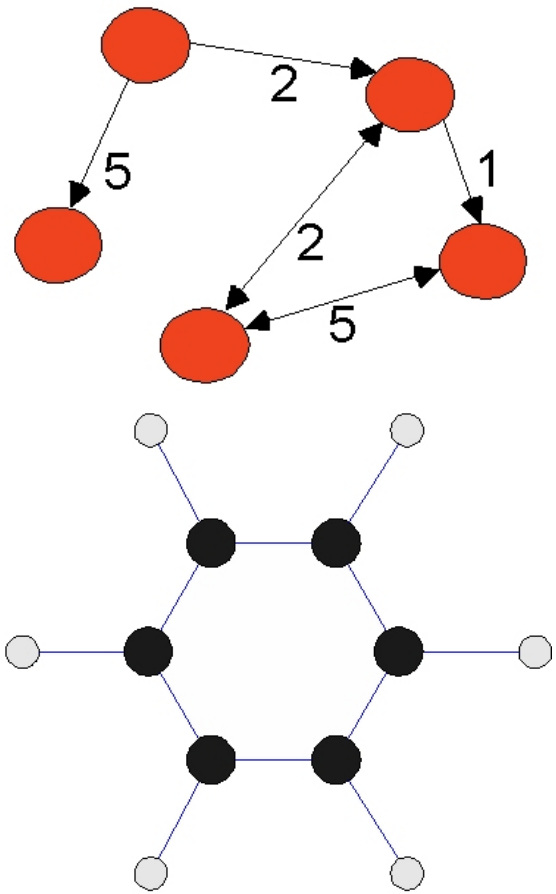
Просторно-временски подаци

Два типа просторно-временских података

- И просторни и временски атрибути могу да буду контекстуални (нпр. мерење варијације температуре мора кроз време)
- Временски атрибут је контекстуалан, а просторни моделира понашање (нпр. анализа трајекторија)

Графовски и мрежни подаци

- Вредности могу да буду придружене чворовима у мрежи
- Однос између вредности података је приказан преко грана
- У неким случајевима атрибути могу да буду додељени чворовима



Најзначајнији градивни блокови у ИП

- Подаци
- Правила придруживања
- Класификација
- Кластеровање
- Анализа и визуелизација резултата

Подаци у процесу ИП

- Мултидимензионална база $\mathscr{D}$ са n слогова и d атрибута
- Претходна припрема података (препроцесирање)
  - Редукција
  - Откривање аномалија
  - ...
- Представљање у облику матрице података D димензије $n \times d$
- Релације између колона
- Релације између врста
- Релације између група атрибута у врстама

Аномалије и елементи ван граница

Дефиниција 6. (Откривање аномалија) За дату матрицу података D одредити редове у матрици који су јако различити од остатка редова

- Податак је елемент ван граница (енг. *outlier*) ако је у значајној мери различит од осталих података
- Синоними: аномалија, абнормалност, дискоординација, ...
- Пример: откривање упада у рачунарски систем, идентификација СПАМ порука, злоупотреба кредитних картица, медицинска дијагностика, спровођење закона, ...

Правила придруживања

- Најједноставнији облик: ретка бинарна база (0/1, са највећим бројем 0)
- Колоне - ставке, врста трансакције:
  $(i,j) = 1 \longrightarrow$ трансакција i садржи ставку j

Бр. транс.	хлеб	млеко	пелене	пиво	јаја	кола
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Правила придруживања

Дефиниција 7. У датој бинарној матрици D величине $n \times d$ посматрају се сви подскупови колона A такви да све вредности у тим колонама у одговарајућој врсти имају вредност 1. Тада важе следеће ознаке:

- A је скуп ставки
- $\#(A)$ означава број појављивања скупа ставки A у комплетном скупу
- N представља број редова у комплетном скупу
- $A \Rightarrow B$ означава да је скуп ставки B придружен скупу ставки A

Правила придруживања

Дефиниција 8. Нека су A и B два скупа ставки. Тада се подршка (енг. *support*) правила придруживања $A \Rightarrow B$ у ознаци *sup* дефинише као количник броја трансакција које садрже A и B у односу на укупан број трансакција

$$sup(A \Rightarrow B) = \frac{\#(A \cup B)}{N}$$

Дефиниција 9. Нека су A и B два скупа ставки. Тада се поузданост (поверење, енг. *confidence*) правила придруживања $A \Rightarrow B$ у ознаци *conf* дефинише као количник броја трансакција које садрже A и B у односу на број трансакција које садрже A

$$conf(A \Rightarrow B) = \frac{\#(A \cup B)}{\#(A)}$$

Правила придруживања

- Задатак је одредити правила придруживања (енг. *association rules*) која повезују атрибуте у истој инстанци
- *Интересантна* су правила која имају одређен ниво подршке и поузданости
- За налажење интересантних правила често се не користе апсолутне фреквенције већ χ^2 мера
- Елементи матрице не морају да буду бинарне вредности

Правила придруживања

Бр. транс.	хлеб	млеко	пелене	пиво	јаја	кола
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

#({млеко, хлеб, пелене})=2

за {млеко, пелене}⇒{пиво}

sup({млеко, хлеб, пелене})=2/5

$$\text{sup}=\frac{\#(\{\text{млеко, пелене, пиво}\})}{N}=2/5$$

$$\text{conf}=\frac{\#(\{\text{млеко, пелене, пиво}\})}{\#(\{\text{млеко, пелене}\})}=2/3$$

Кластеровање

- Груписање врста по 'сличности'
- Кластеровање се у литератури јавља и као *класификација без надзора*

x_1^1	x_1^2	x_1^3	x_1^4	...	x_1^d
x_2^1	x_2^2	x_2^3	x_2^4	...	x_2^d
x_3^1	x_3^2	x_3^3	x_3^4	...	x_3^d
x_4^1	x_4^2	x_4^3	x_4^4	...	x_4^d
x_5^1	x_5^2	x_5^3	x_5^4	...	x_5^d
x_6^1	x_6^2	x_6^3	x_6^4	...	x_6^d
...	...	...	...	...	...
x_n^1	x_n^2	x_n^3	x_n^4	...	x_n^d

Класификација

- Релације између колона - нека колона је значајнија од других
- Класификација се у литератури јавља и као *Класификација под назором*

x_1^1	x_1^2	x_1^3	x_1^4	...	x_1^d
x_2^1	x_2^2	x_2^3	x_2^4	...	x_2^d
x_3^1	x_3^2	x_3^3	x_3^4	...	x_3^d
x_4^1	x_4^2	x_4^3	x_4^4	...	x_4^d
x_5^1	x_5^2	x_5^3	x_5^4	...	x_5^d
x_6^1	x_6^2	x_6^3	x_6^4	...	x_6^d
...	...	...	...	...	...
x_n^1	x_n^2	x_n^3	x_n^4	...	x_n^d

Класификација (наставак)

Дефиниција 10. (Класификација података) За дату матрицу података D из базе $\mathscr{D}$ величине $n \times d$ и вредност ознака класа у интервалу $\{1, ..., k\}$ које су придружене сваком од n слогова (редова) формира се модел за тренирање $\mathscr{M}$, који може да се користи за предвиђање ознаке класа d-димензионалног слога $\overline{Y} \notin \mathscr{D}$.

- Модел се формира над *подацима за тренинг*
- *Тест подаци* $\notin \mathscr{D}$ (скуп који се користи за формирање модела)

Неки примери ИП апликација

- Распоређивање производа у радњама
- Препоруке купцима
- Аномалије у веб логовима
- ...