

Извличане на знания от текст

Преслав Наков

19 февруари 2025 г.

(some slides adapted from Marti Hearst, Gert Lanckriet,
Chris Manning, Barbara Rosario, and others)

Класификация



Въведение

План

- Класификация и примери:
 - Разпознаване на автор
 - Разпознаване на език
 - Категоризиране в семантична категория
- Класификация и клъстеризация
- Типове атрибути
- По-важни класификатори
- Манипулация с атрибути
 - селекция
 - претегляне
 - трансформация: ЛСА

Класификация

Цел: Разпределяне на "обекти" от някакво множество в два или повече *класа* или *категории*

Примери:

Задача

тагиране с част на речта

определяне на значение

information retrieval

класификация на текст

разпознаване на автор

Обект

дума

дума

документ

документ

документ

Категории

части на речта

значения

подходящ/неподходящ

категория (тема)

автор(и)

Пример



Разпознаване на автор

Разпознаване на автор

- Тая прохладна майска вечер чорбаджи Марко, гологлав, по халат, вечеряше с челядта си на двора. Господарската трапеза беше сложена, както по обикновение, под лозата, между бистрия и студен чучур на барата, който като лястовичка пееше, деня и нощя, и между високите бухлати чемшири, що се тъмнееха край зида, зиме и лете все зелени. Фенерът светеше, окачен на клончето на едно люлеково дръвче, което приятелски надвисваше миризливите си люлеки над главите на челядта. А тя беше многобройна.
- Като на всеки празник, кръчмата се пълнеше с хора. През отворените прозорци можеше да се види как идат и ония селяни, които бяха позакъснели. Вървяха бавно, тежко, като че умората беше ги налегнала едвам сега, когато бяха останали без работа. Всички бяха си турили чисти ризи с широки бели ръкави, спираха се и гледаха насам, гледаха нататък. И как няма да гледат? Трева е поникнало и на камък. Такава зеленина е навън, че и в кръчмата като си седи човек, пред очите му играят зелени кръгове.

Разпознаване на автор

- Като заваля дъжд, та цяла неделя! Тихо, кротко, ден и нощ. Вали, вали, вали – напои хубаво майката земя, па духна тих ветрец, очисти небето и пекна топло есенно слънце. Засъхнаха нивята. Оправи се време – само за оране. Боне Крайненецът впрегна пак Сивушка и Белчо и тръгна след ралото. Нивата му е в един хубав широк валог! От всички страни гора и завет.
- Това беше отговорът, който получих на 9 септември т. г. от г. Стоилова на телеграмата ми, с която молих да се спрат явно беззаконните „морални влияния“ на свищовските преставители на властта. Окръжният управител с председателя на постоянната комисия и околийския началник няколко дни преди изборите бяха тръгнали от село на село из Свищовската околия и упражняваха „морално влияние“ по такъв начин: „Негово царско височество пред няколко дни, като мина край Свищов, поръча да изберете тези, недейте избира другите, защото те ще изпъдят княза.“

Разпознаване на автор

- Иван Вазов: “Под игото”
- Йордан Йовков: “Другоселец”
- Елин Пелин: “На браздата”
- Алеко Константинов: “По изборите в Свищов”

Разпознаване на автор (стилометрия)



?



Разпознаване на автор

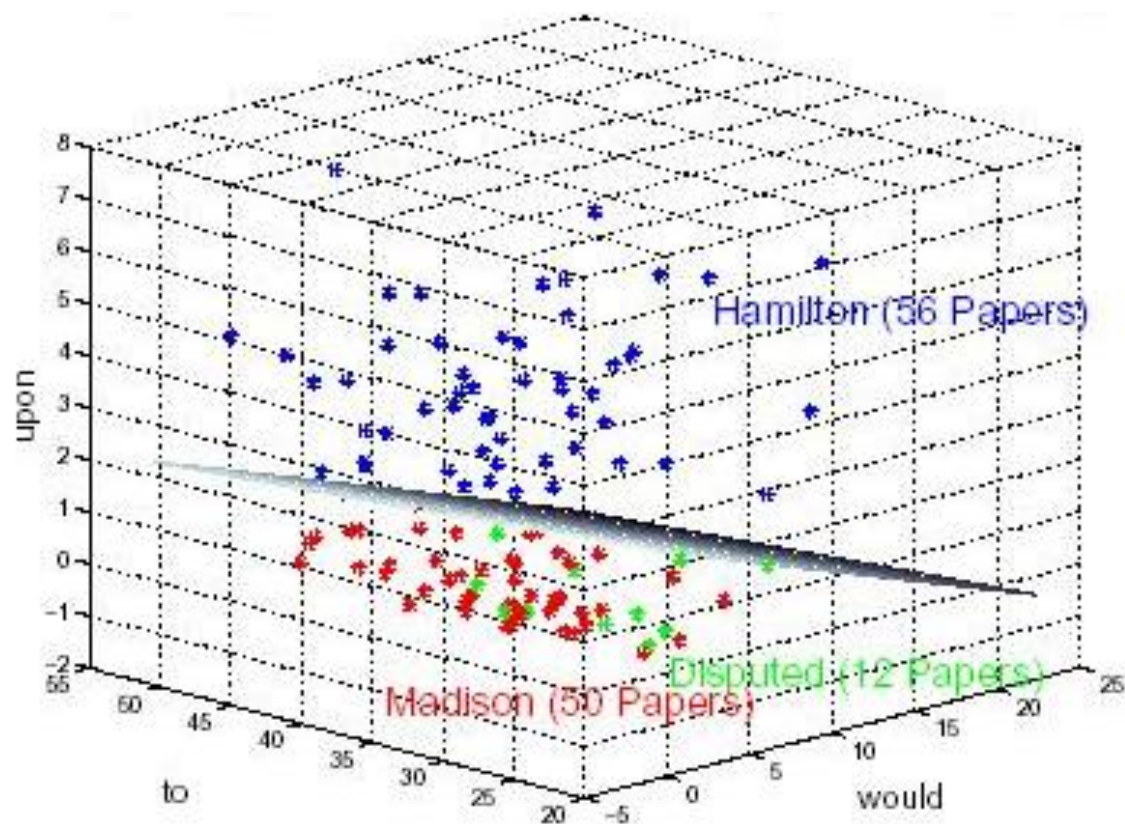
□ Federalist papers

- 77 кратки есета, 1787-1788 г.,
- автори: Хамилтън и Мадисън
 - публикувани под псевдоним
- за да убедят Ню Йорк да ратифицира американската конституция
- Авторството на 12 е оспорвано
- Решение (*Мадисън*): 1964 г., Мостълър и Уолъс
 - 70 функционални думи
 - статистически техники

Функционални думи за разпознаване на автор

1	<i>a</i>	15	<i>do</i>	29	<i>is</i>	43	<i>or</i>	57	<i>this</i>
2	<i>all</i>	16	<i>down</i>	30	<i>it</i>	44	<i>our</i>	58	<i>to</i>
3	<i>also</i>	17	<i>even</i>	31	<i>its</i>	45	<i>shall</i>	59	<i>up</i>
4	<i>an</i>	18	<i>every</i>	32	<i>may</i>	46	<i>should</i>	60	<i>upon</i>
5	<i>and</i>	19	<i>for</i>	33	<i>more</i>	47	<i>so</i>	61	<i>was</i>
6	<i>any</i>	20	<i>from</i>	34	<i>must</i>	48	<i>some</i>	62	<i>were</i>
7	<i>are</i>	21	<i>had</i>	35	<i>my</i>	49	<i>such</i>	63	<i>what</i>
8	<i>as</i>	22	<i>has</i>	36	<i>no</i>	50	<i>than</i>	64	<i>when</i>
9	<i>at</i>	23	<i>have</i>	37	<i>not</i>	51	<i>that</i>	65	<i>which</i>
10	<i>be</i>	24	<i>her</i>	38	<i>now</i>	52	<i>the</i>	66	<i>who</i>
11	<i>been</i>	25	<i>his</i>	39	<i>of</i>	53	<i>their</i>	67	<i>will</i>
12	<i>but</i>	26	<i>if</i>	40	<i>on</i>	54	<i>then</i>	68	<i>with</i>
13	<i>by</i>	27	<i>in</i>	41	<i>one</i>	55	<i>there</i>	69	<i>would</i>
14	<i>can</i>	28	<i>into</i>	42	<i>only</i>	56	<i>things</i>	70	<i>your</i>

Разпознаване на автор



Пример



Разпознаване на език

Класификация

Цел: Разпределяне на “обекти” от някакво множество в два или повече *класа* или *категории*

Примери:

Задача

Разпознаване на автор

Разпознаване на език

Обект

документ

документ

Категории

автор(и)

език

Разпознаване на език

- Tutti gli esseri umani nascono liberi ed eguali in dignità e diritti. Essi sono dotati di ragione e di coscienza e devono agire gli uni verso gli altri in spirito di fratellanza.
- Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geist der Brüderlichkeit begegnen.
- Сите човечки суштества се раѓаат слободни и еднакви по достоинство и права. Тие се обдарени со разум и совест и треба да се однесуваат еден кон друг во духот на општо човечката припадност.

ООН: из "Универсална декларација на човешките права" (на 363
езика)

Разпознаване на език

- Sva ljudska bića radjaju se slobodna i jednaka u dostojanstvu i pravima. Ona su obdarena razumom i svešću i treba jedni prema drugima da postupaju u duhu bratstva.
- Všichni lidé rodí se svobodní a sobě rovní co do důstojnosti a práv. Jsou nadáni rozumem a svědomím a mají spolu jednat v duchu bratrství.
- Wszyscy ludzie rodzą się wolni i równi pod względem swej godności i swych praw. Są oni obdarzeni rozumem i sumieniem i powinni postępować wobec innych w duchu braterstwa.

ООН: из "Универсална декларация на човешките права" (на 363 езика)

Разпознаване на език

□ Думи

- égaux
- uguali
- iguales
- edistämään

□ Знаци

- ü
- ě
- Я, Щ, Ю

Категоризиране на текст



Класифициране в
семантични категории
(определяне на **тема**)

Класификация

Цел: Разпределяне на "обекти" от някакво множество в два или повече *класа* или *категории*

Примери:

Задача

разпознаване на автор

разпознаване на език

категоризиране на текст

Обект

документ

документ

документ

Категории

автор(и)

език

тема

Категоризиране на текст

▣ Категоризиране: класифициране в **семантични категории**

Същевременно ЦСКА уговори още 2 контроли, които ще са в София. На 29 януари ще има мач с "Пирин 1922", а на 19 февруари с "Конелиано". За лагера в Агия Напа (Кип) вече са ясни 4 срещи. Съперници ще са румънският "Динамо", чешкият "Сигма", полският "Легия" и местният "Неа Саламина".

Резките движения на пазара отчасти бяха причинени от почивните дни около празниците, когато валутната търговия не е особено активна. Инвеститорите спекулираха с това, че администрацията на Буш няма да направи нищо, за да спре понижаването на долара.

Категории в Yahoo News

Business

AP Reuters | AFP | BusinessWeek Online | FT.com | NPR | USATODAY.com | My Sources

- McDonald's stocks jump on 3Q forecast AP - 17 minutes ago
- Dow passes 11,900 for 1st time AP - 10 minutes ago
- PepsiCo 3Q profit climbs 71 percent AP - 2 hours, 51 minutes ago
- Oil drives trade deficit to new high AP - 2 hours, 55 minutes ago
- Costco 4th-quarter earnings edge up AP - 1 hour, 46 minutes ago

» All Business from AP

Science

AP Reuters | AFP | SPACE.com / LiveScience.com | NPR | My Sources

- New type of mouse discovered in Cyprus AP - Thu Oct 12, 7:17 AM ET
- Jupiter tiny spot goes from white to red AP - Wed Oct 11, 9:47 PM ET
- Cat-cloning company to close its doors AP - Wed Oct 11, 8:44 PM ET
- 2 scientists name asteroid for Nev. town AP - Wed Oct 11, 8:44 PM ET
- Wasps released in La. to combat bugs AP - Wed Oct 11, 10:31 PM ET

» All Science from AP

Technology

AP Reuters | USATODAY.com | PC World | PC Magazine | AFP | My Sources

- REVIEW: TiVo extras add little luster AP - Wed Oct 11, 10:50 PM ET
- Computers may translate in war settings AP - Wed Oct 11, 6:42 PM ET
- YouTube community worried by Google deal AP - Wed Oct 11, 6:56 PM ET
- Sun hosts news conference in Second Life AP - Wed Oct 11, 6:46 PM ET
- Microsoft releases 6 patches for flaws AP - Wed Oct 11, 6:53 PM ET

» All Technology from AP

Health

AP Reuters | HealthDay | AFP | NPR | ACS News Today | My Sources

Категоризиране на текст: някои приложения

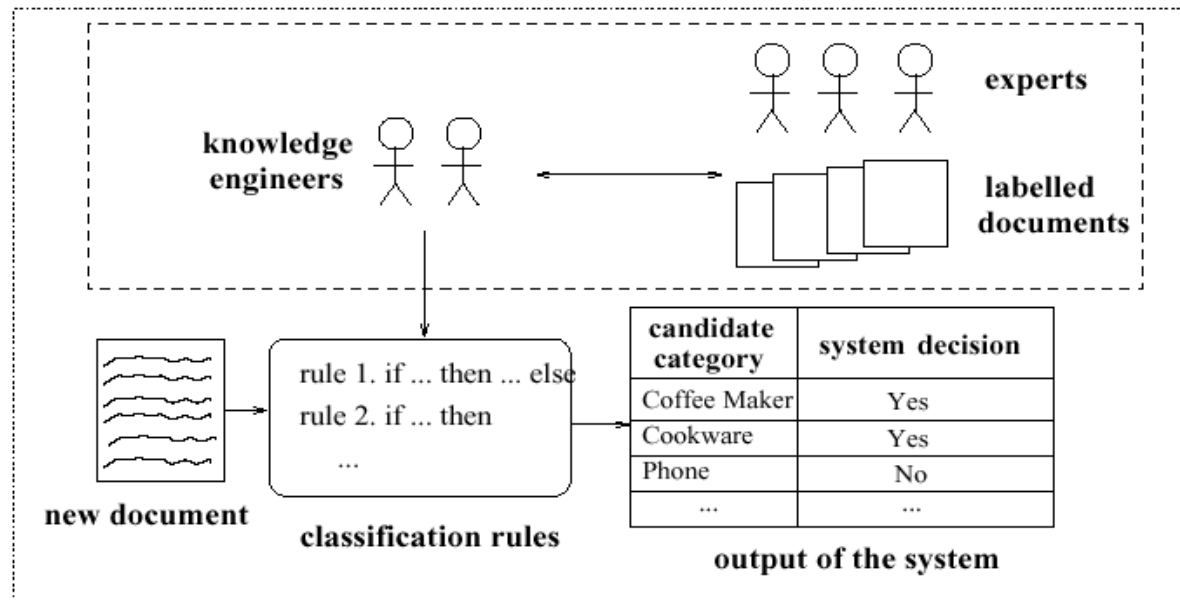
- Филтриране на новини
- Организиране на Уеб страници в йерархични категории
- Филтриране на спам
- Автоматично категоризиране на имейл по папки
- Библиографско индексирание на статии: (Library of Congress, MEDLINE, и др.)

Защо не полуавтоматично категоризиране на текст?

- Хората могат да кодират съответно експертно знание за категоризиране.
- Това знание може да бъде кодирано и използвано за автоматично категоризиране на нови примери.
- Например...

Экспертни системи (края на 80-те)

Expert system for text categorization (late 1980s)



Системи, основани на правила

□ Текст

“Saeco revolutionized *espresso* brewing a decade ago by introducing Saeco SuperAutomatic *machines*, which go from bean to *coffee* at the touch of a button. The all-new Saeco Vienna Super-Automatic home coffee and *cappucino machine* combines top quality with low price!”

□ Правила

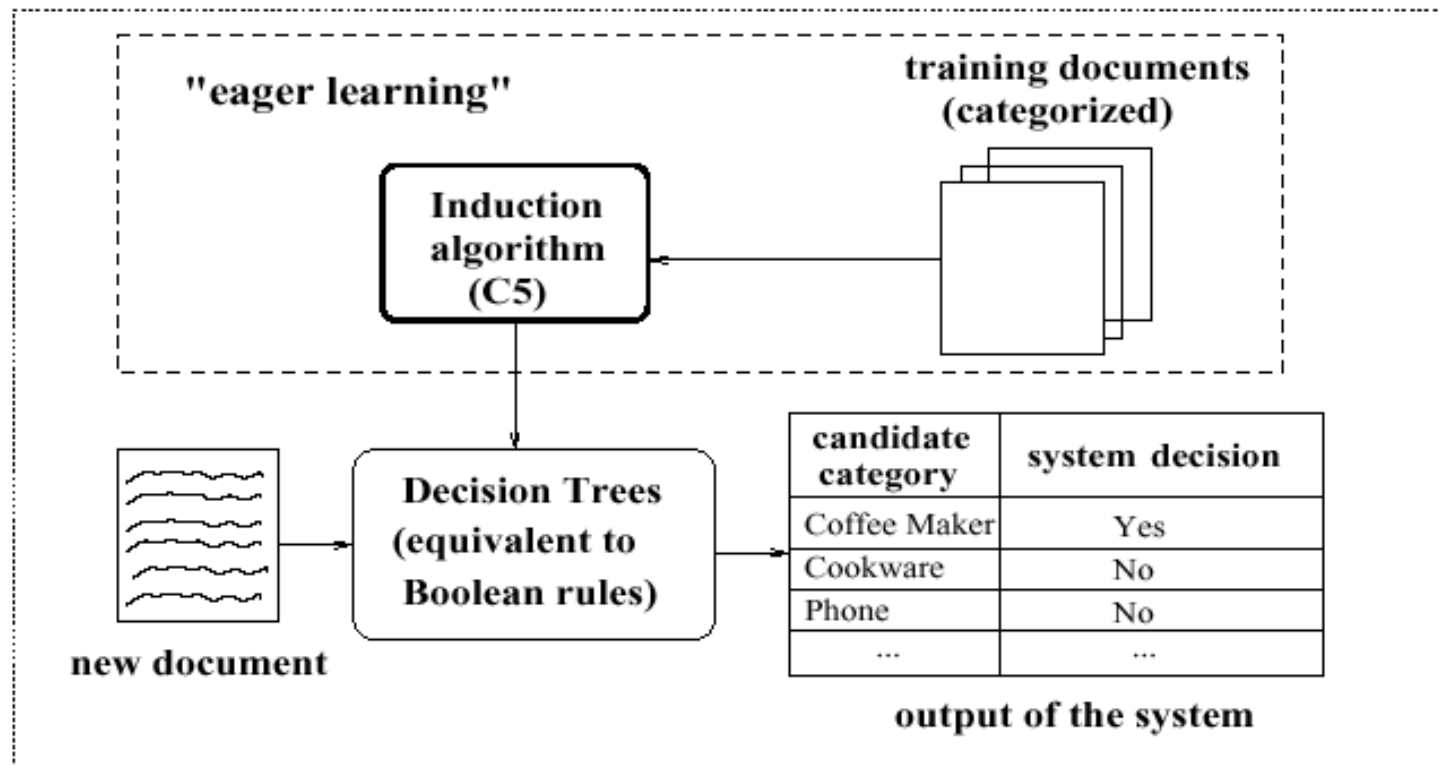
- Правило 1.
(*espresso* **or** *coffee* **or** *cappucino*) **and** *machine** \Rightarrow *Coffee Maker*
- Правило 2.
*automat** **and** *answering* **and** *machine** \Rightarrow *Phone*
- Правило *n*...

Ръчно кодирани правила

- Това е достатъчно за прости приложения
 - Google и Yahoo! alerts позволяват на потребителите да задават ключови думи, по които да получават новини
 - На англ. "filtering" или "routing"
 - Работи добре, ако е допустимо да изпуснем някои неща
- Но ако е нужна висока точност, практиката показва че
 - много времеемко
 - много трудно
 - проблеми със съвместимостта на правилата (особено при голям набор правила)

Използване на машинно самообучение

DTree induction for text categorization (since 1994)



Цена на ръчната категоризация

■ Yahoo!

- 200 (?) души ръчно аотират Уеб страници
- йерархия с 500,000 категории!

■ MEDLINE (National Library of Medicine)

- \$2 млн. годишно за ръчно индексирание на статии от списания
- MeSH: MEdical Subject Headings (18,000 категории)

■ Mayo Clinic

- \$1.4 млн. годишно за кодиране на събития в записи на пациенти
- International Classification of Diseases (ICD) за застрахователни компании

■ US Census Bureau (1990: 22 млн. отговора)

- 232 индустриални категории и 504 категории заетост
- \$15 млн., ако се прави изцяло ръчно

Ръчно кодиране
на знание

vs.

Машинно
самообучение

- US Census Bureau Decennial Census за 1990
 - 232 индустриални категории и 504 категории заетост
 - \$15 млн., ако се прави изцяло ръчно

- Ръчно дефинирани правила
 - Експертна система AIOCS
 - Време: 192 човекомесеца (2 души, 8 години)
 - Точност: 47%

- Машинно самообучение
 - Метод на най-близкия съсед (Creesy '92: 1-NN)
 - Време: 4 човекомесеца (Thinking Machine)
 - Точност: 60%

Категоризиране на текст: колекции

□ Reuters

- колекция от 21,578 новини
- 135 категории

□ 20 newsgroups

- приблизително 20,000 документа
- 20 групи – равномерно разпределени

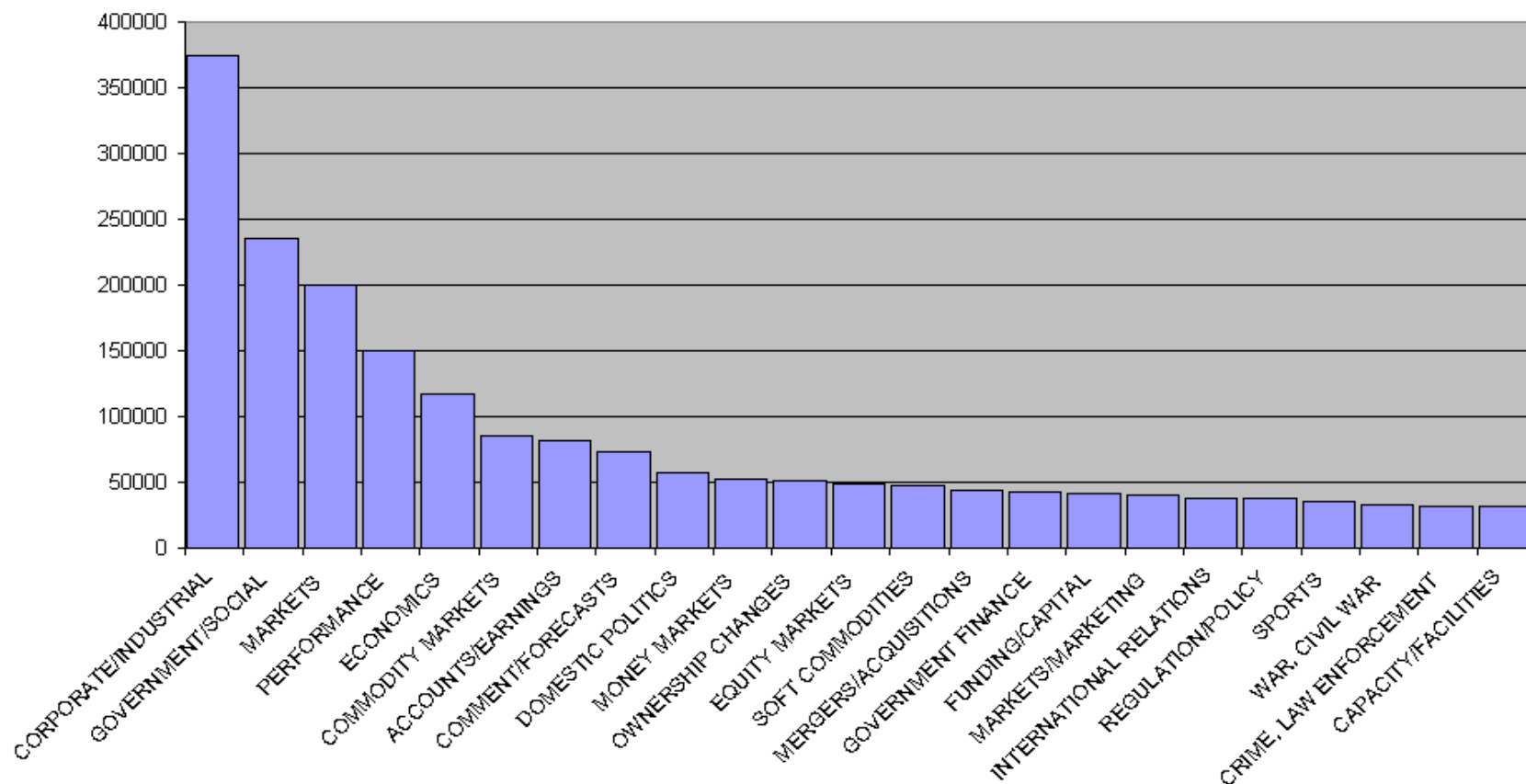
□ MEDLINE

- могат да се ползват MeSH (**M**edical **S**ubject **H**eadings)

□ LINGSPAM – част от LINGUIST list

- филтриране на спам

□ Основни категории



Reuters

<http://trec.nist.gov/data/reuters/reuters.html>

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981"
NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off

tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

</BODY></TEXT></REUTERS>

20 Newsgroups

http://people.csail.mit.edu/u/j/jrennie/public_html/20Newsgroups/

- Источник: събрани от Кен Ланг
- Съдържание и структура:
 - приблизително 20,000 новинарски групи
 - 19,997 общо
 - 18,828 без дубликатите
 - *равномерно* разпределени в 20 категории
- Някои категории са близки, т.е. трудни:

computers



comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Пример от "talk.politics.guns"

From: cdt@sw.stratus.com (C. D. Tavares)

Subject: Re: Congress to review ATF's status

In article <C5vzHF.D5K@cbnews.cb.att.com>, lvc@cbnews.cb.att.com (Larry Cipriani) writes:

> WASHINGTON (UPI) -- As part of its investigation of the deadly
> confrontation with a Texas cult, Congress will consider whether the
> Bureau of Alcohol, Tobacco and Firearms should be moved from the
> Treasury Department to the Justice Department, senators said Wednesday.
> The idea will be considered because of the violent and fatal events
> at the beginning and end of the agency's confrontation with the Branch
> Davidian cult.

Of course. When the catbox begins to smell, simply transfer its
contents into the potted plant in the foyer.

"Why Hillary! Your government smells so... FRESH!"

--

cdt@rocket.sw.stratus.com --If you believe that I speak for my company,
OR cdt@vos.stratus.com write today for my special Investors' Packet...

Класификация и клъстеризация



Обучение **с** и **без** учител

Класификация и клъстеризация

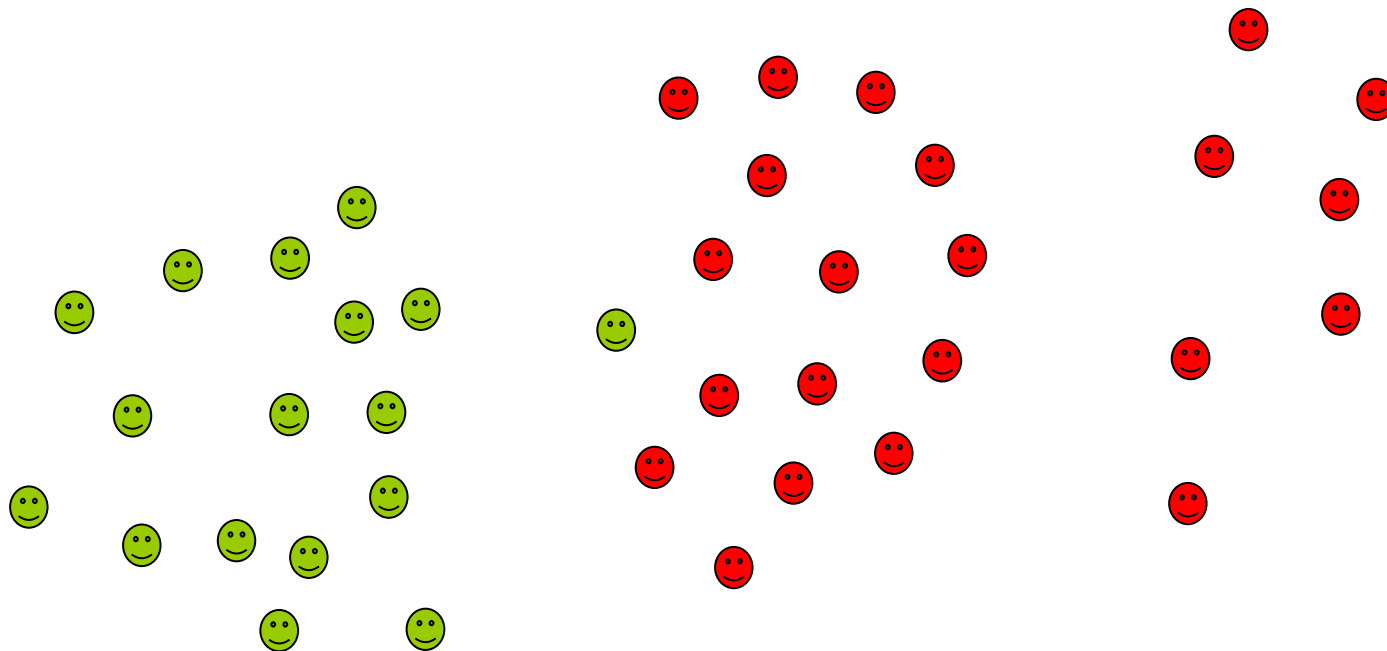
□ Класификация

- **обучение с учител**
- аотирани примери (с техния клас)
- знаем кои са класовете
- **искаме да предскажем класа на **нови** примери**

□ Клъстеризация

- **обучение без учител**
- примерите не са аотирани
- може би не знаем броя на класовете
- **търсим вътрешна структура в **дадените** примери**

Класификация

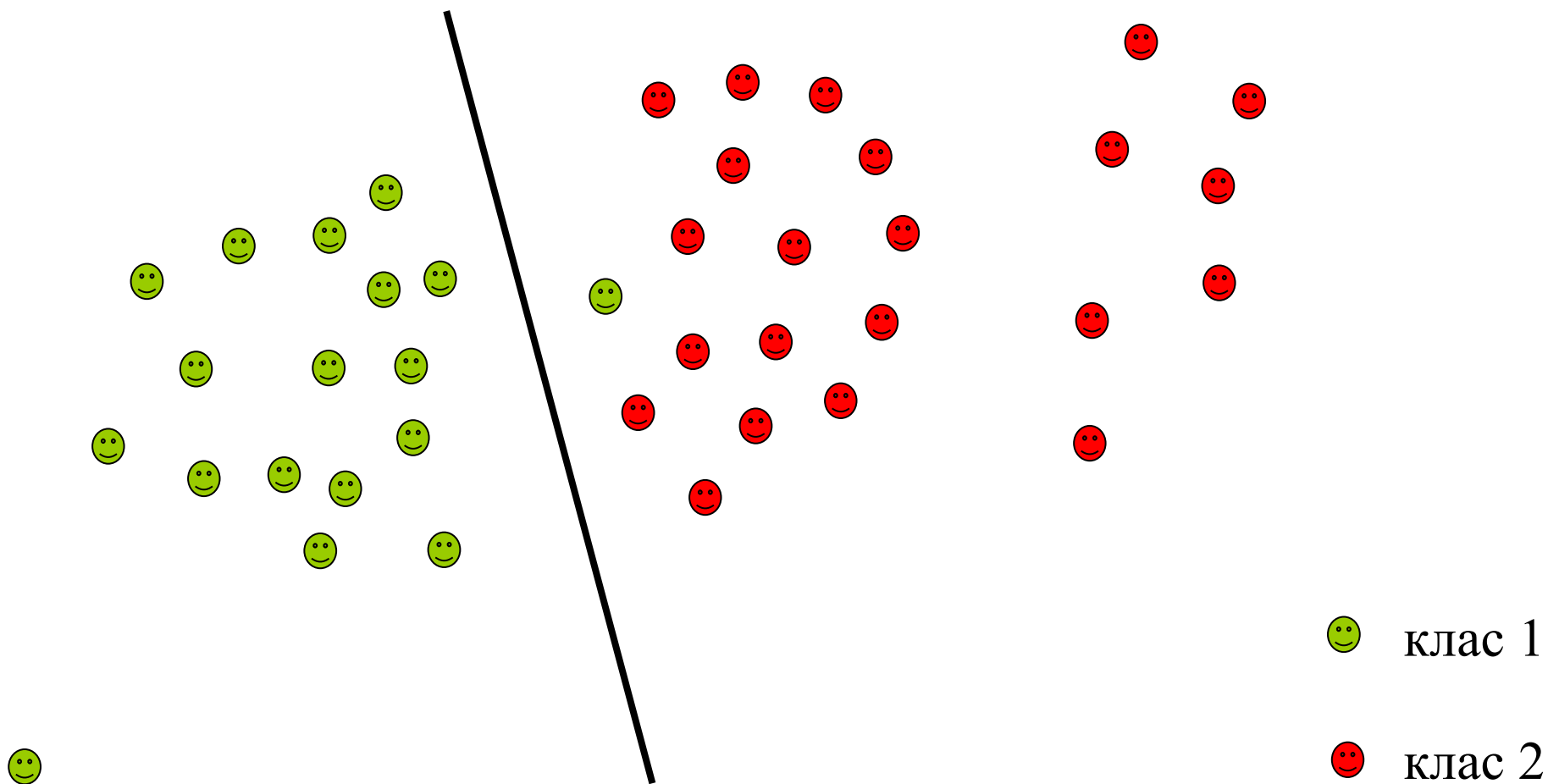


😊 клас 1

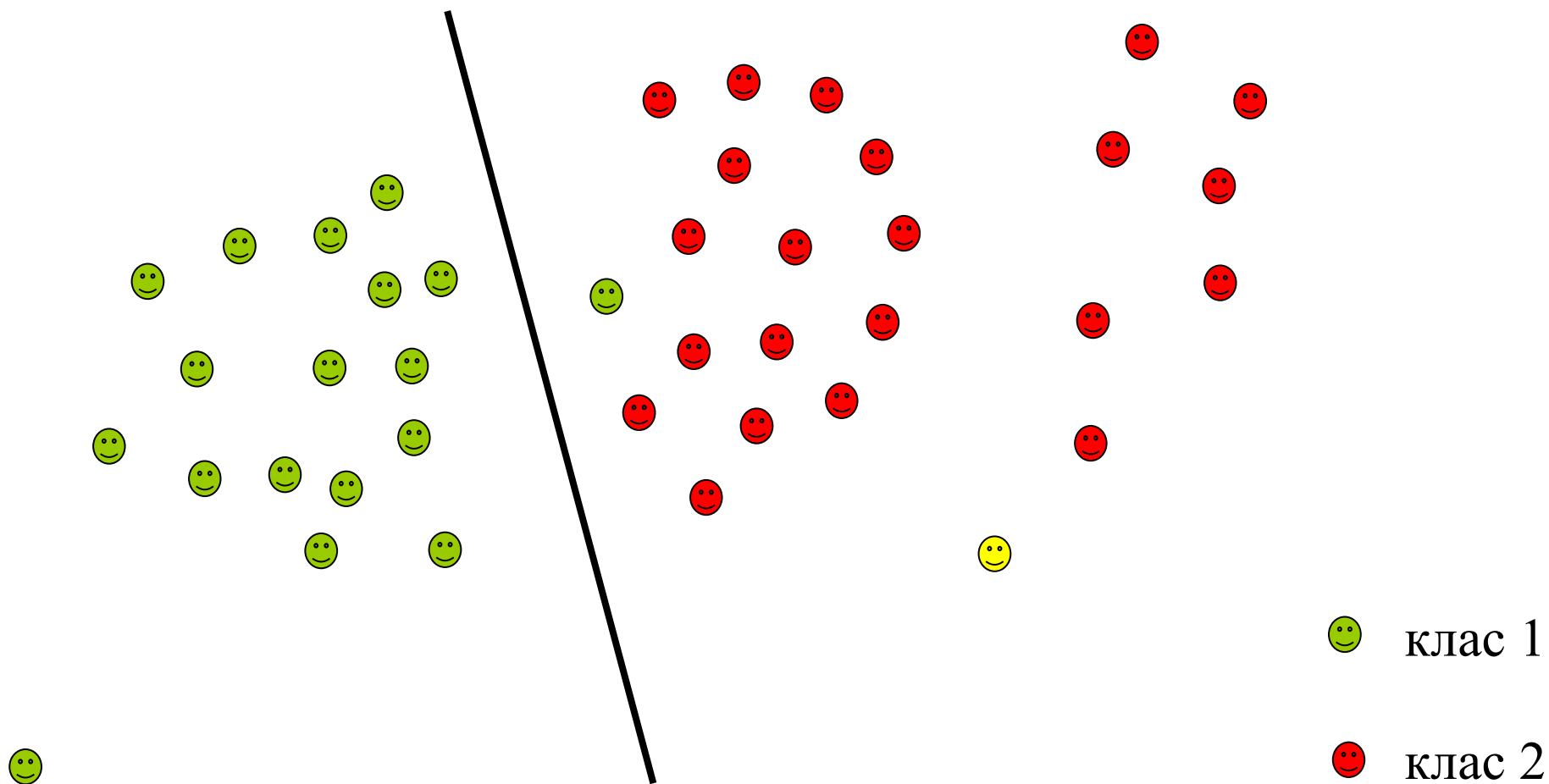
😬 клас 2



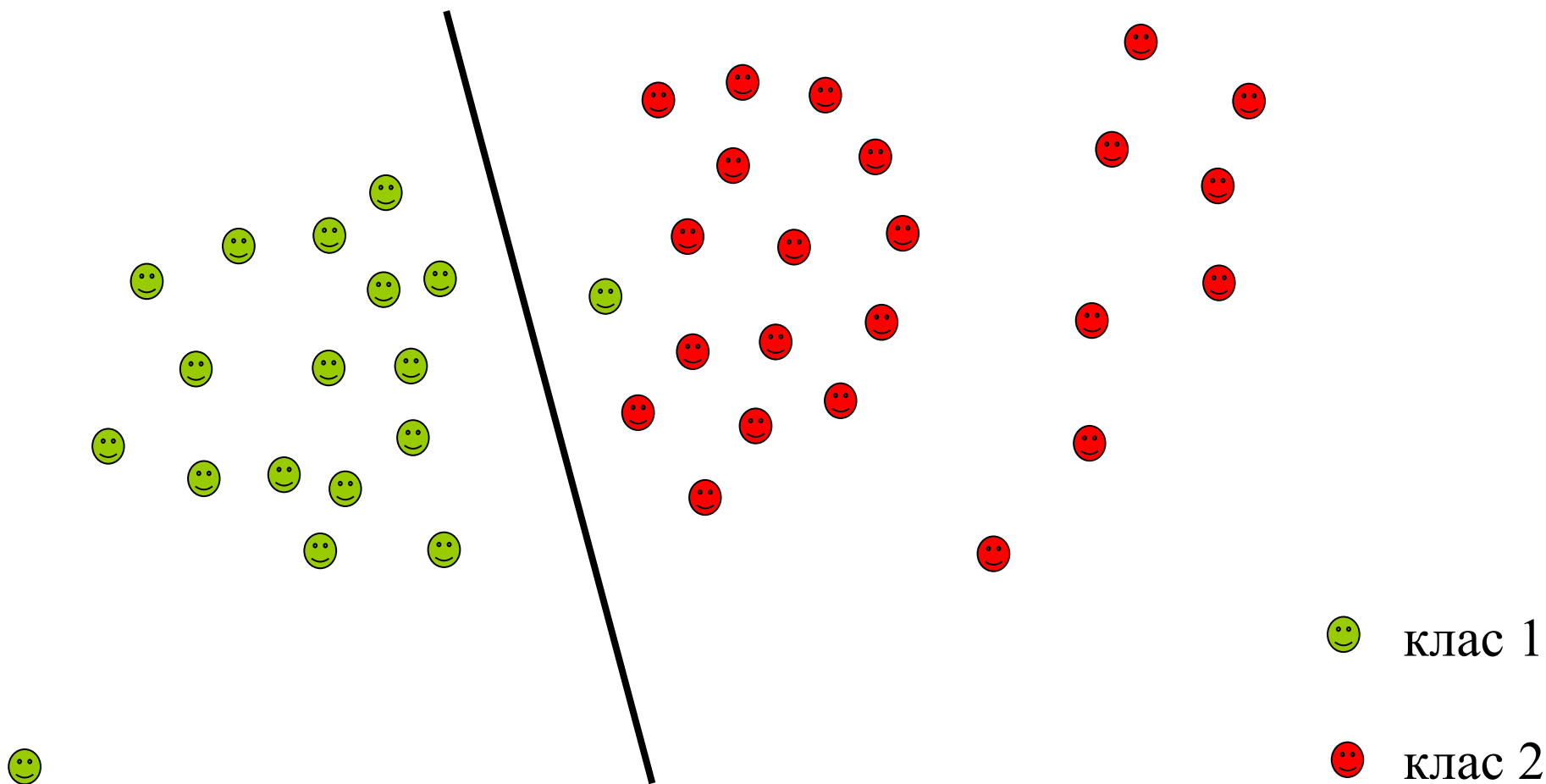
Класификация



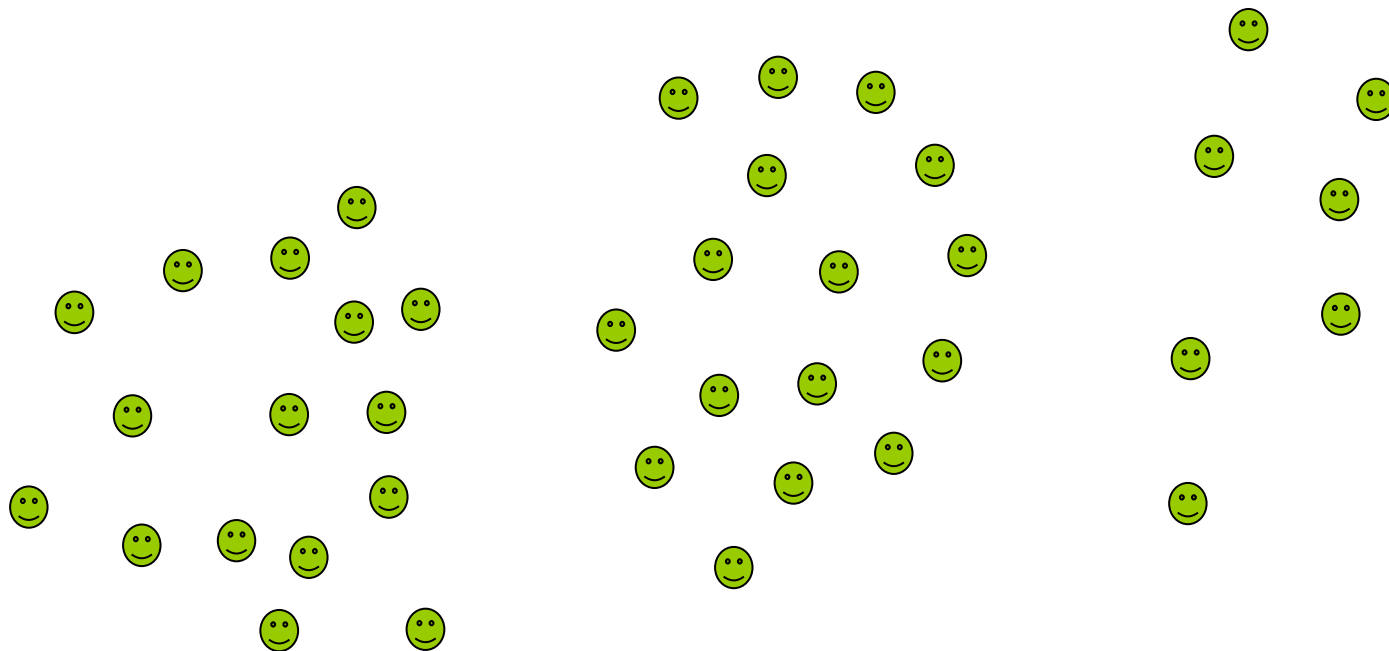
Класификация



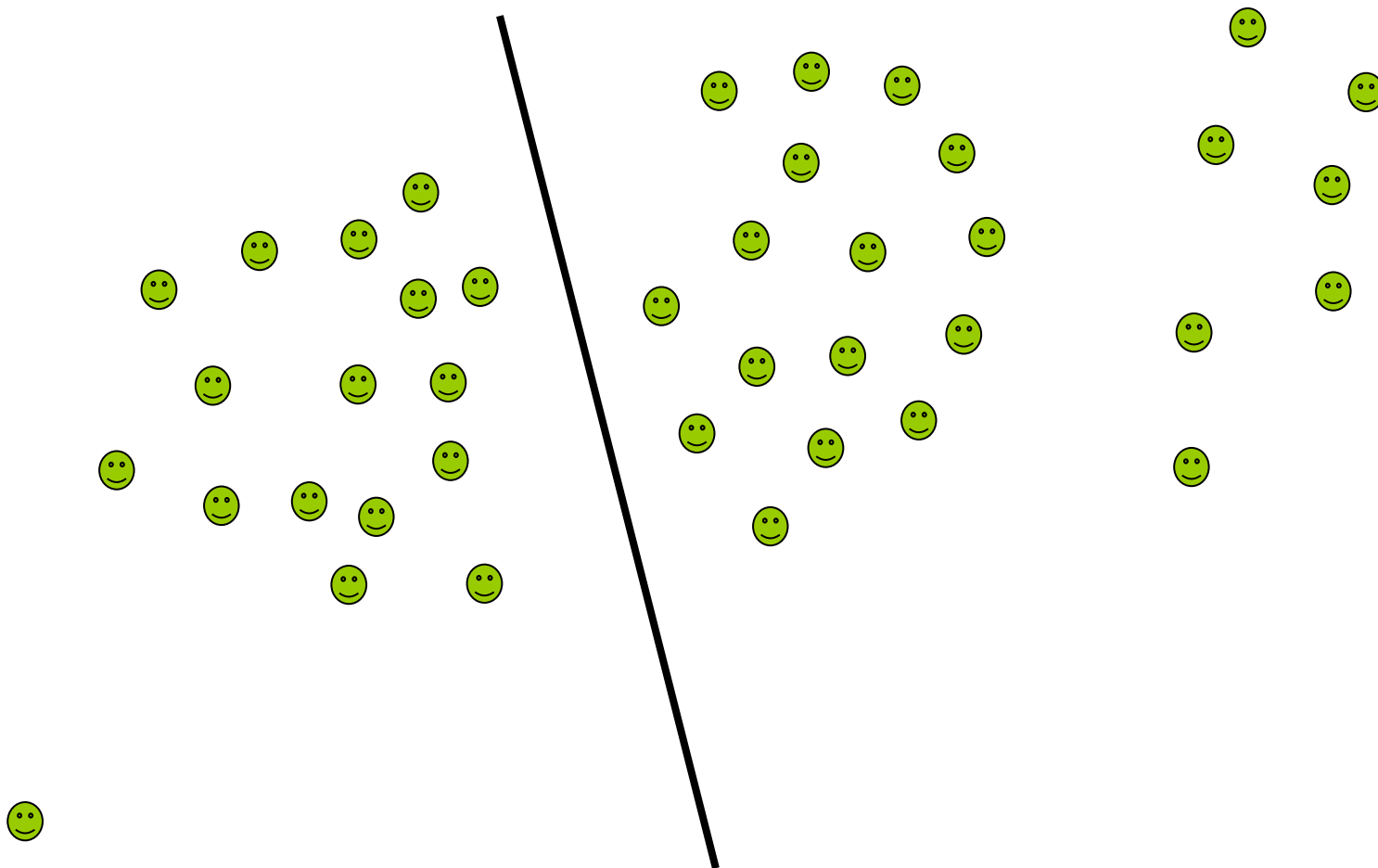
Класификация



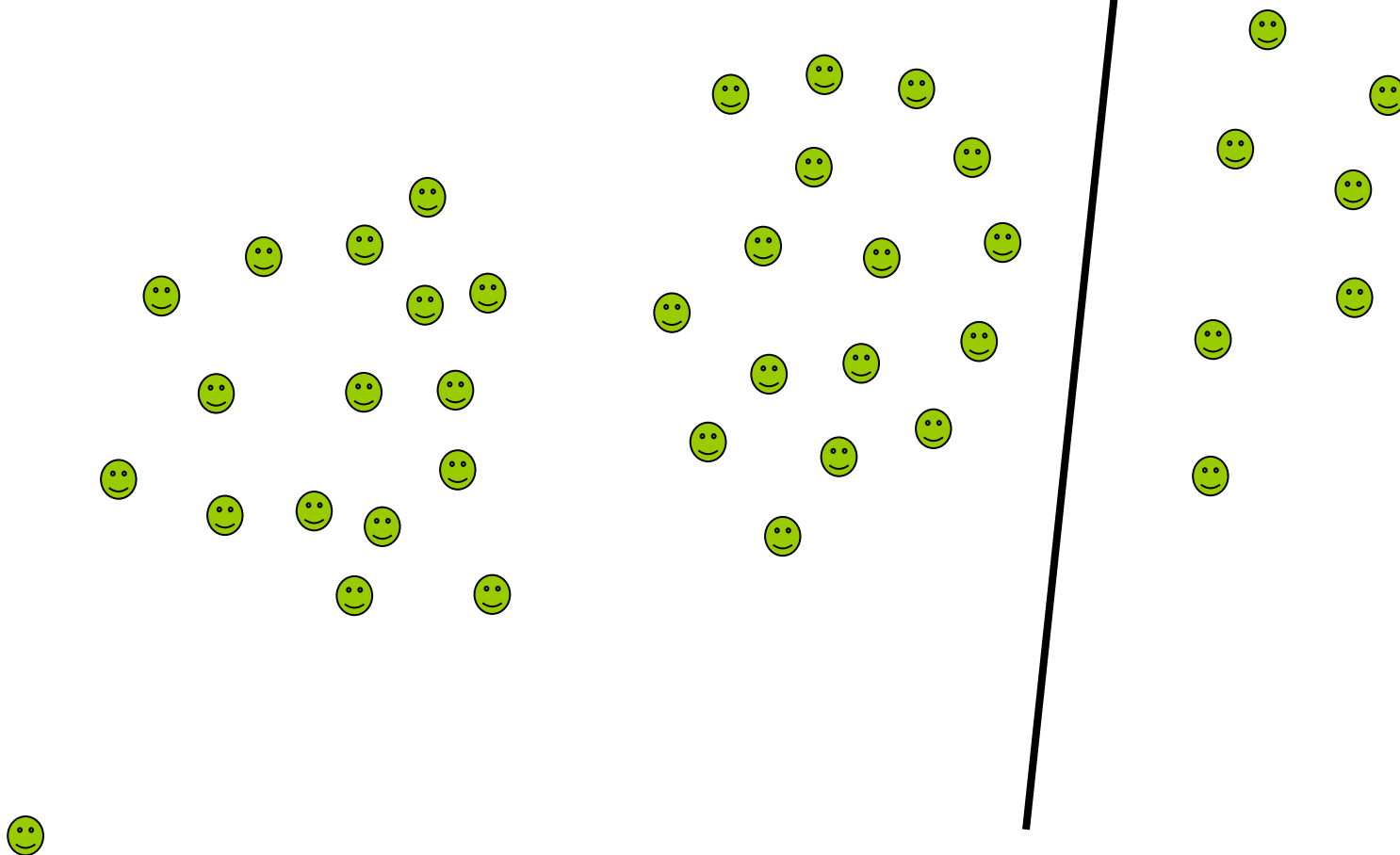
Клъстеризация



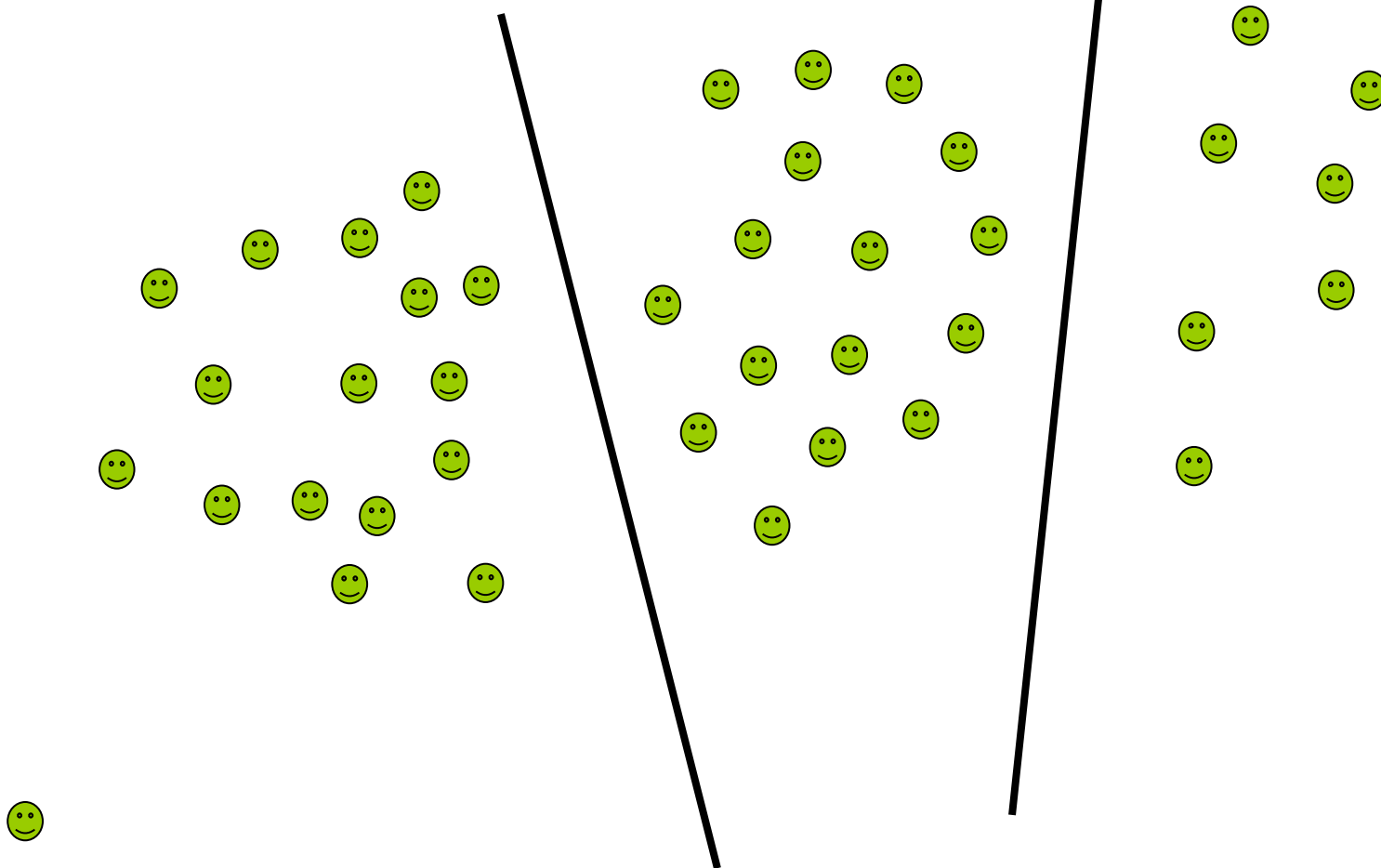
Клъстеризация



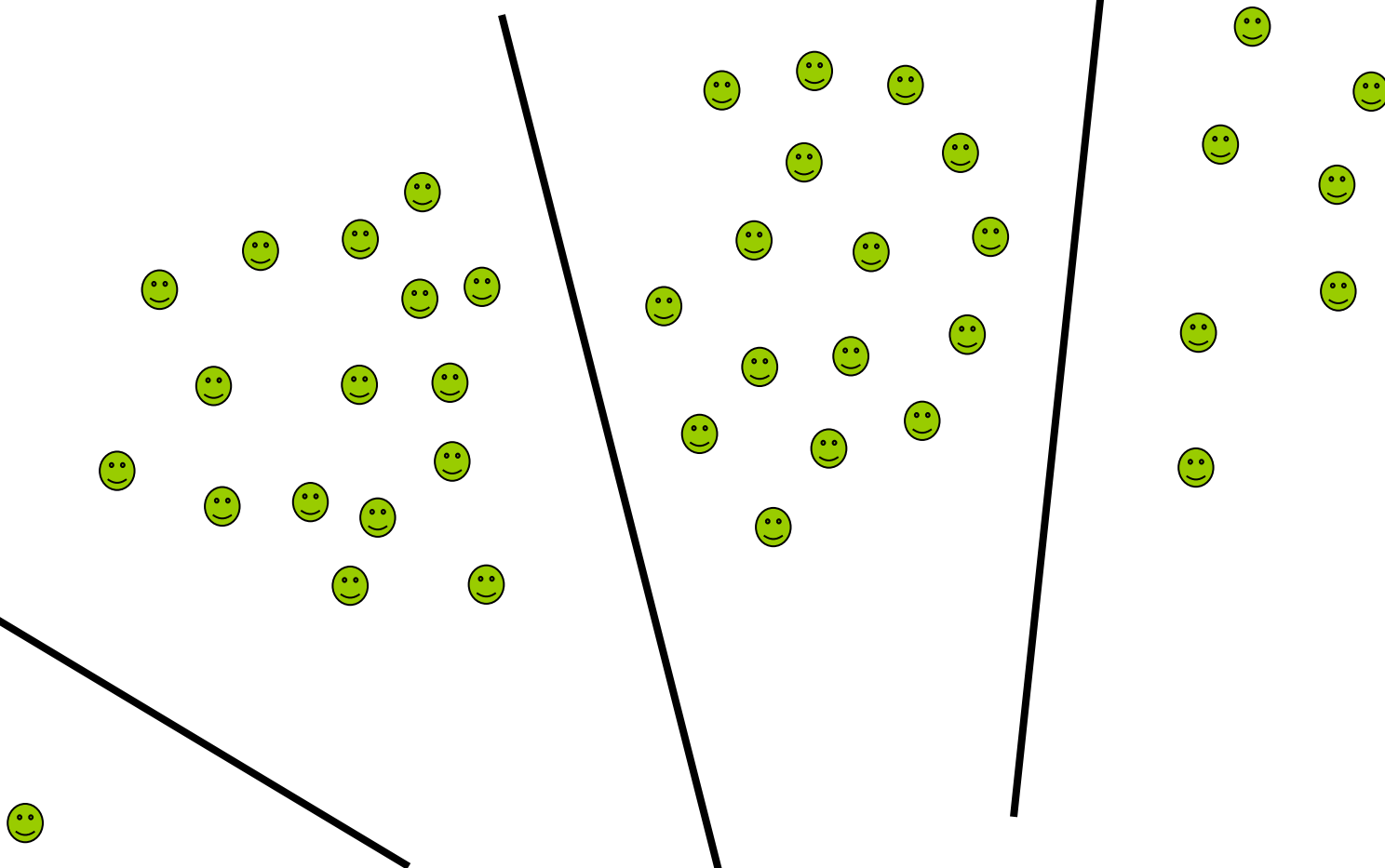
Клъстеризация



Клъстеризация



Клъстеризация



Класификация



Типове

Анотиране на данни

- Избор на класове (кои? колко?)

- в зависимост от:

- типа текст
 - приложението

- Анотиране на текста

- трудно

- бавно

- несъответствие между различни хора

- *капа статистика*

Reuters

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

Защо не **topic = policy?**

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off

tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the **National Pork Producers Council**, NPPC.

Delegates to the three day **Congress** will be considering 26 resolutions concerning various issues, including the future direction of **farm policy** and the **tax law** as it applies to the **agriculture sector**. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

</BODY></TEXT></REUTERS>

Два и повече класа

- Дихотомична класификация
 - два класа
- Повече от два класа
 - често се третира като дихотомична класификация
 - един клас срещу останалите (за всички класове)

Проста и йерархична класификация

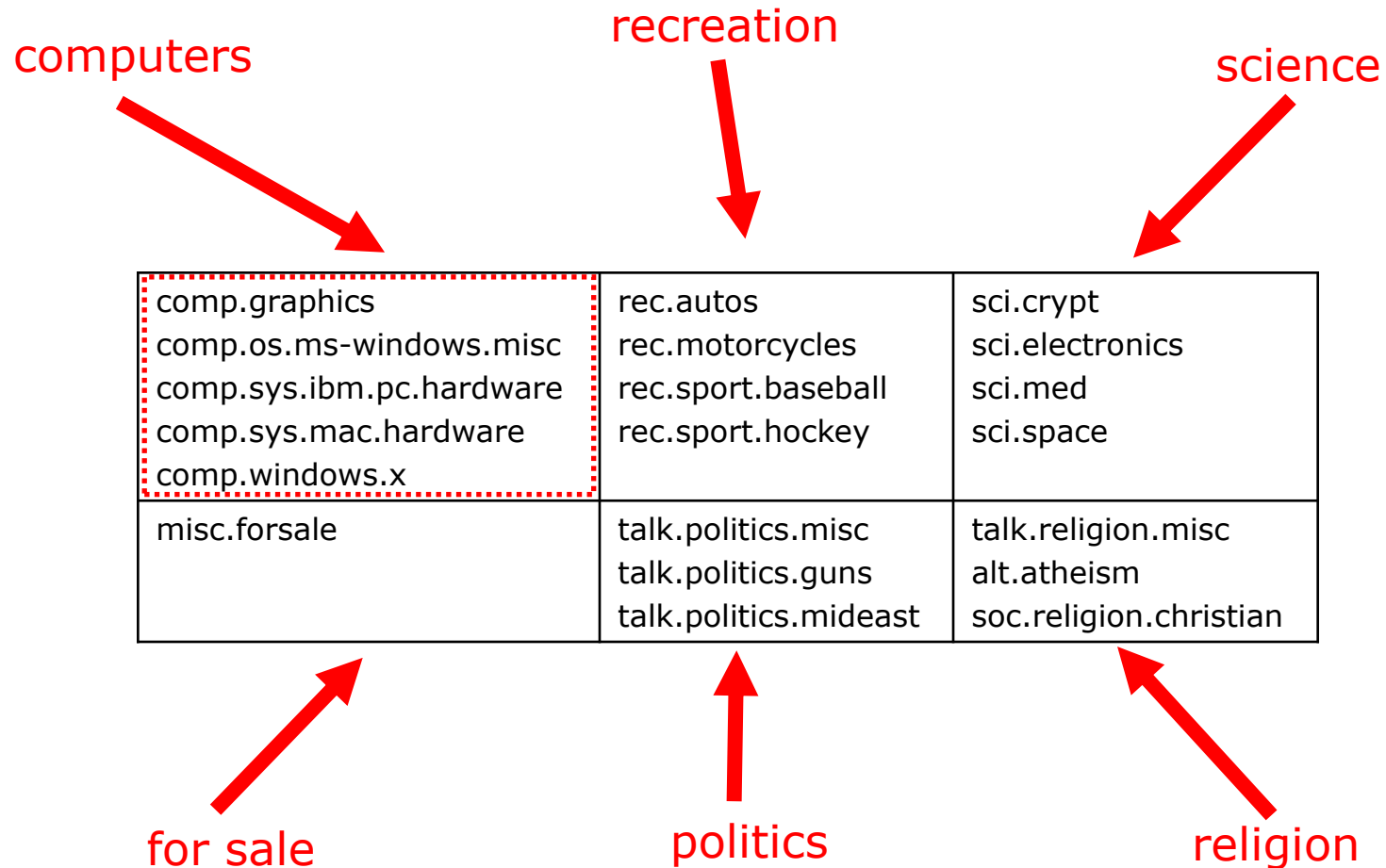
□ Проста класификация:

- точно един клас за всеки пример
- без отношение между класовете

□ Йерархична класификация:

- нула или повече класове за всеки пример
- йерархия
- всеки клас е подклас на своите родители

Иерархия в "20 Newsgroups"



Атрибути



Типове, избор и др.

Атрибути

- Текст: "Seven-time Formula One champion Michael Schumacher took on the Shanghai circuit Saturday in qualifying for the first Chinese Grand Prix."
- Категория?
 - sport
- Атрибути?

Атрибути

- *Текст: "Seven-time **Formula One** champion **Michael Schumacher** took on the **Shanghai** circuit **Saturday** in qualifying for the first **Chinese Grand Prix**."*

- **Атрибути?**

Атрибути

- **Атрибут:** свойство на текста с отношение към задачата
- **Стойност:** реализация на атрибута в текста
 - думи: *Formula, Schumacher, China...*
 - честота на думи: *Formula (10), Schumacher (1)...*
 - съдържа ли дати? да/не
 - съдържа ли имена на хора? да/не
 - съдържа ли имена на организации? да/не
 - WordNet:
 - холоними (China е част от Asia),
 - синоними (China, People's Republic of China, mainland China)

Типове атрибути

Булеви

- приемат две стойности: *истина* и *лъжа*
- най-простите атрибути

■ $f_1(\text{text}) = 1$ ако текстът съдържа "Schumacher"
0 иначе

■ $f_2(\text{text}) = 1$ ако текстът съдържа име на човек
0 иначе

Типове атрибути

Целочислени

- приемат целочислени стойности
- носят повече информация

■ $f_1(\text{text})$ = **КОЛКО ПЪТИ** текстът съдържа **"Schumacher"**

■ $f_2(\text{text})$ = **КОЛКО ПЪТИ** текстът съдържа **име на човек**

Секции в "talk.politics.guns"

ОТ КОГО
From: cdt@sw.stratus.com (C. D. Tavares)
Subject: Re: Congress to review ATF's status

заглавие

тема

In article <C5vzHF.D5K@cbnews.cb.att.com>, lvc@cbnews.cb.att.com (Larry Cipriani) writes:

> WASHINGTON (UPI) -- As part of its investigation of the deadly
> confrontation with a Texas cult, Congress will consider whether the
> Bureau of Alcohol, Tobacco and Firearms should be moved from the
> Treasury Department to the Justice Department, senators said Wednesday.
> The idea will be considered because of the violent and fatal events
> at the beginning and end of the agency's confrontation with the Branch
> Davidian cult.

отговор

Of course. When the catbox begins to smell, simply transfer its contents into the potted plant in the foyer.

анекдот

"Why Hillary! Your government smells so... FRESH!"

--

сигнатура

cdt@rocket.sw.stratus.com --If you believe that I speak for my company,
OR cdt@vos.stratus.com write today for my special Investors' Packet...

Класификатори



Обучение и тестване на
класификатор

Класификация

- Дефиниране на категории
- Анотиране на текст
- Извличане на атрибути
- Избор на класификатор
 - наивен Бейсов класификатор
 - невронна мрежа (персептрон)
 - SVM и др.
- Обучение и тестване
- Приложение върху нови примери

Обучение

- Адаптиране на класификатора към данните
 - търсене на “добри” стойности на параметрите
 - критерий за оптимизация:
 - ✓ *брой грешки*
 - ✓ *разстояние между класовете*
 - ✓ *други*
- Някои класификатори гарантирано намират оптимални стойности на параметрите си

Набори документи

□ Учебен

- настройка на *основните* параметри на класификатора

□ Валидационен

- настройка на *допълнителни* параметри на класификатора

□ Тестов

- Различен от учебния
- Не се използва за промяна на параметрите
- Дава оценка на качеството на класификатора

Оценка на класификатори

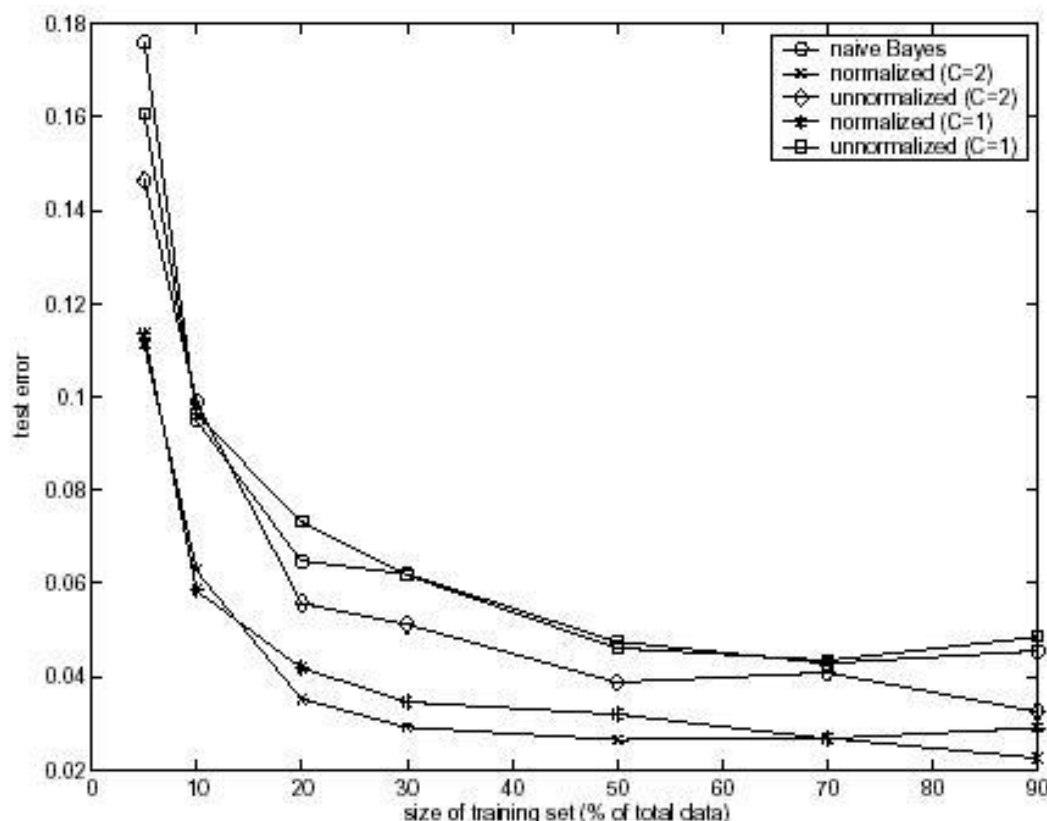
- Таблица за оценка на двоични класификатори

	Зелено (вярно)	Червено (вярно)
Зелено предсказано	a	b
Червено предсказано	c	d

- Accuracy = $(a+d)/(a+b+c+d)$
- Precision: $P_{\text{зелен}} = a/(a+b)$, $P_{\text{черв}} = d/(c+d)$
- Recall: $R_{\text{зелен}} = a/(a+c)$, $R_{\text{черв}} = d/(b+d)$
- $F1 = 2PR/(P+R)$ $F1_{\text{зелен}}$ / $F1_{\text{черв}}$

Размер на учебните данни

- Повече данни (по-принцип) е по-добре!
- 20 newsgroups: rec.sport.baseball и rec.sport.hockey



*From: Improving the Performance of Naive Bayes for Text Classification, Shen and Yang

Размер на учебните данни

□ Идентификация на автор

“Насищане”



Размерът има значение!

□ От Banko & Brill '01

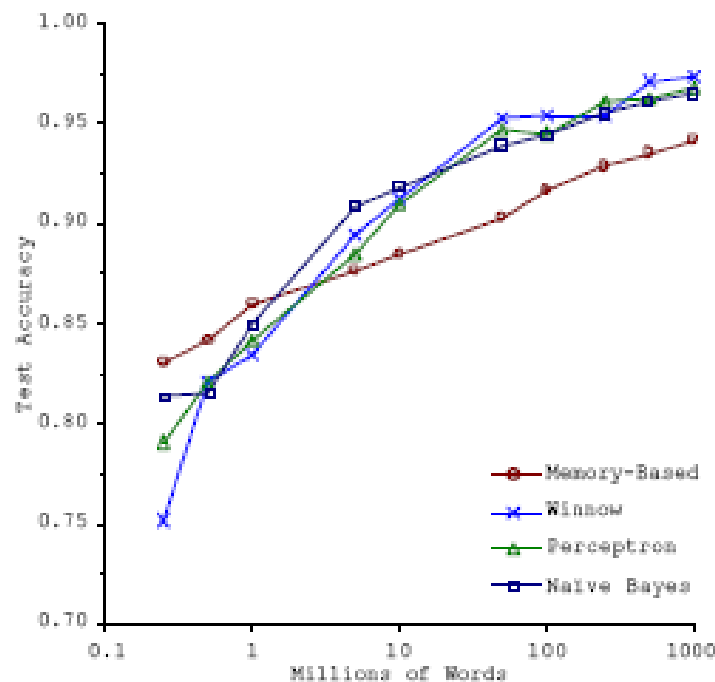


Figure 1. Learning Curves for Confusion Set Disambiguation

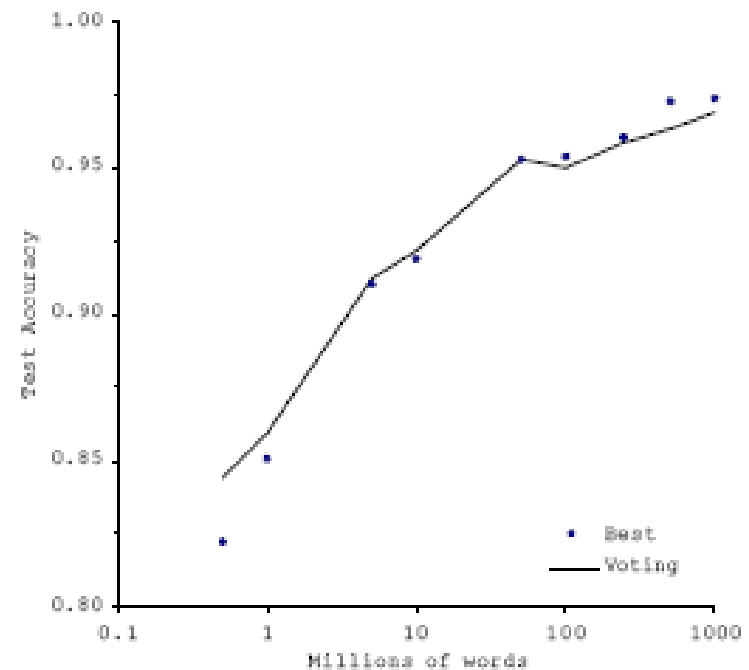


Figure 3. Voting Among Classifiers

Класификатори



Някои по-важни
класификатори

По-важни алгоритми

□ Алгоритми за класификация

□ **Двоична класификация**

- Персептрон/Winnow
- Support Vector Machines (SVM)

□ **Множествена класификация**

- **Multiclass:** избор на един измежду N класа?

- Дърво на решенията
- Наивен Бейсов класификатор
- Метод на най-близките съседи

- **Multilabel:** избор на $0,1,2,\dots,N$ измежду N класа?
- Решава се с N двоични класификатора:
 - всеки клас, срещу останалите $N-1$ класа

Двоична класификация: примери

- ❑ Филтриране на спам (спам, “неспам”)
- ❑ Класифициране на съобщения (спешно vs. неспешно)
- ❑ Класифициране на емоции (положителни, отрицателни)

Понякога е удобно да третираме множествената класификация като двоична: една категория срещу всички останали. **Защо?**

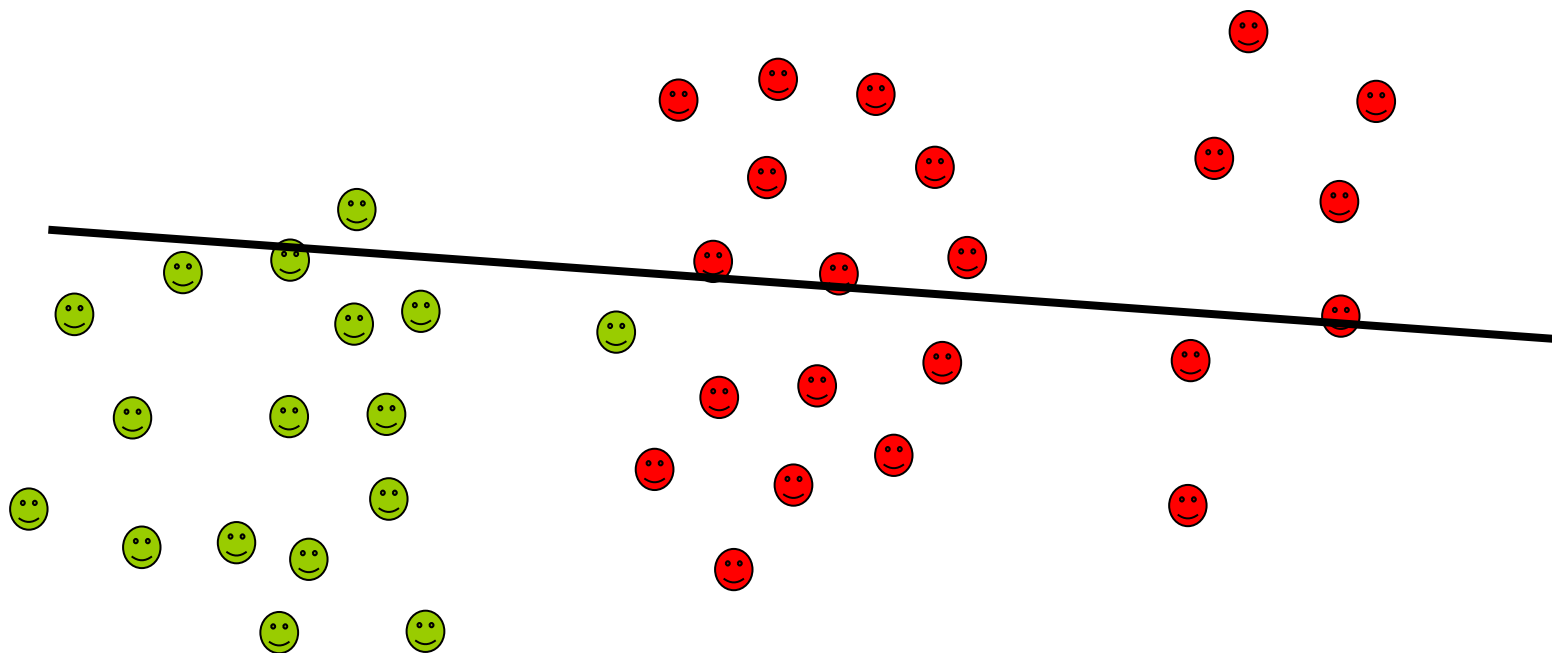
Двоична класификация


- ▣ **Дадено:** някакви данни, принадлежащи към положителен (+1 😊) и отрицателен (-1 😞) клас
- ▣ **Задача:** Да се научи класификатор, който може да предсказва класа на нови данни
- ▣ **Геометрично:** търсим разделител

Линейни и нелинейни алгоритми

- ▣ **Линейно разделими данни:** ако данните могат да се класифицират чрез местоположението си по отношение на някаква права в равнината (хиперравнина)

Линеен класификатор



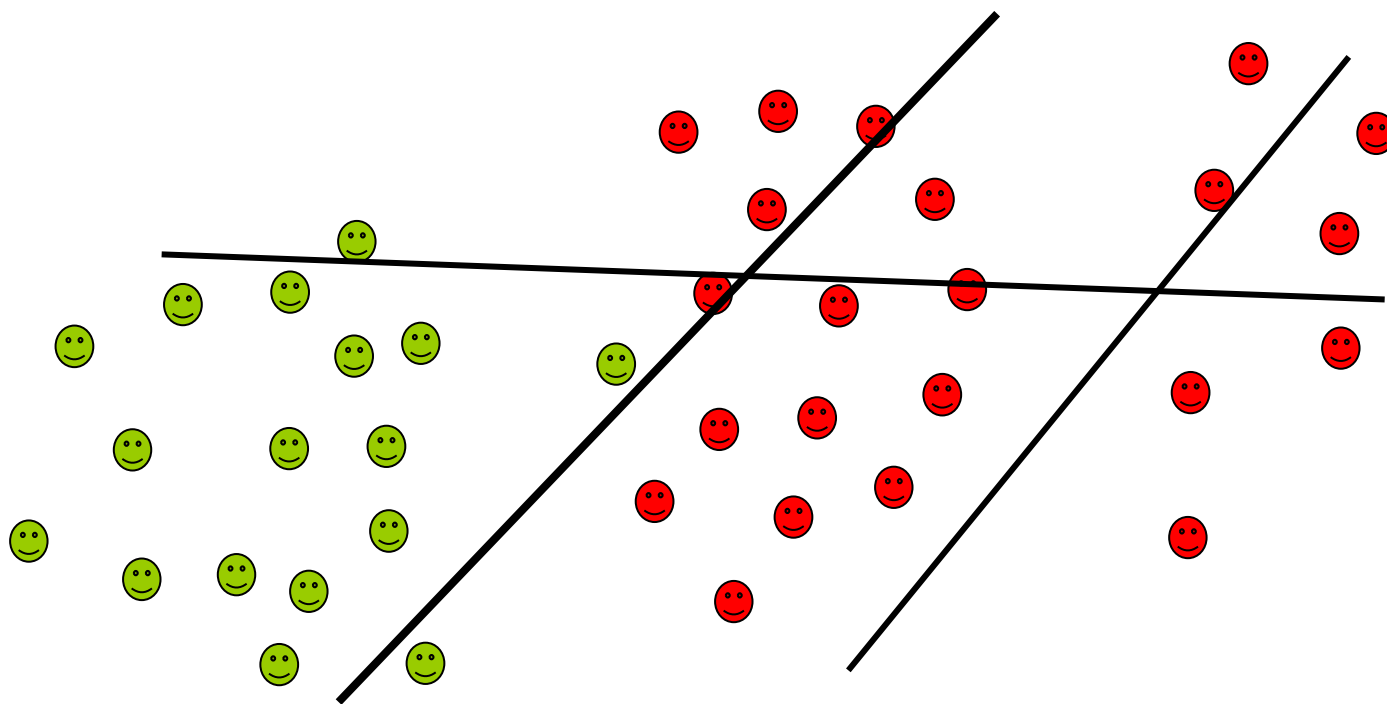
Линеен класификатор: $g(x) = \mathbf{w}x + \mathbf{w}_0$  клас 1

параметри: \mathbf{w} , \mathbf{w}_0

 клас 2



Линеен класификатор



Линеен класификатор: $g(x) = \mathbf{w}x + \mathbf{w}_0$

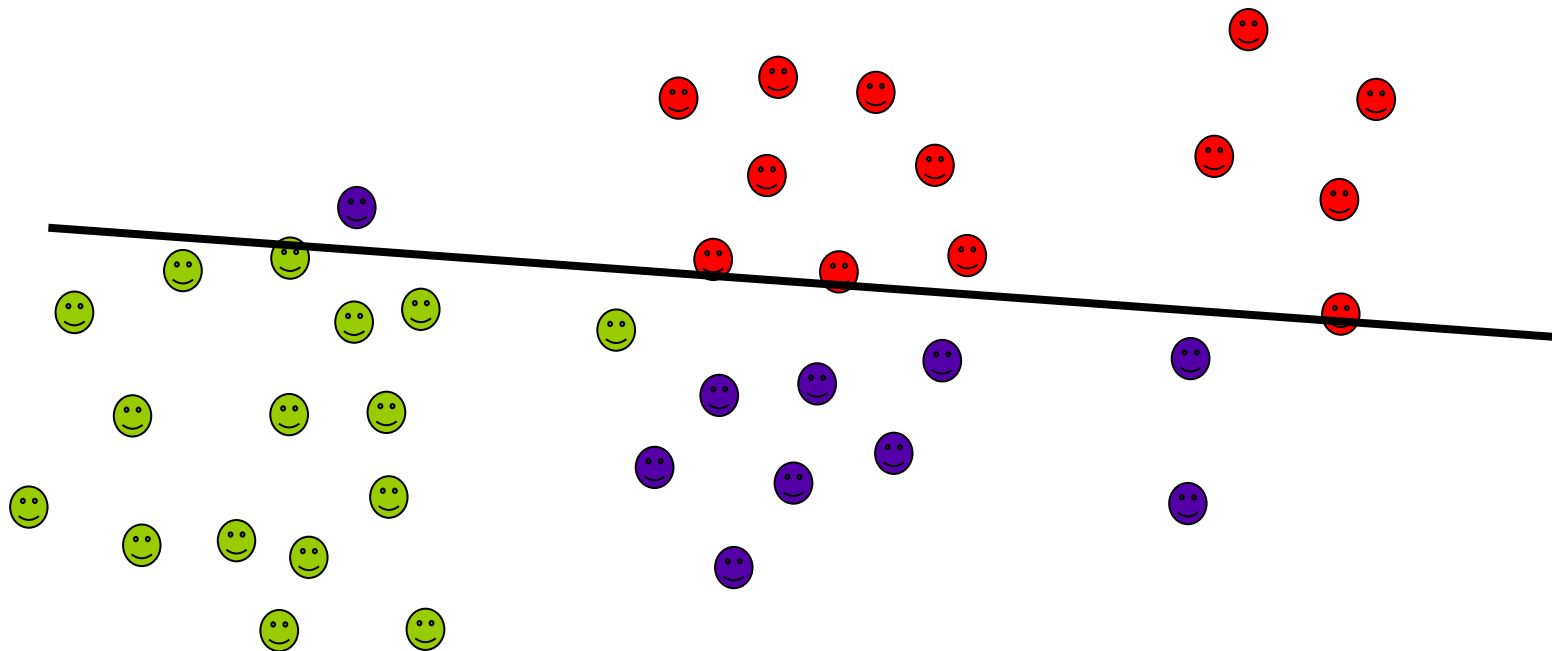
😊 клас 1

различни параметри: \mathbf{w} , \mathbf{w}_0

😬 клас 2



Линеен класификатор



Линеен класификатор: $g(x) = \mathbf{w}x + \mathbf{w}_0$

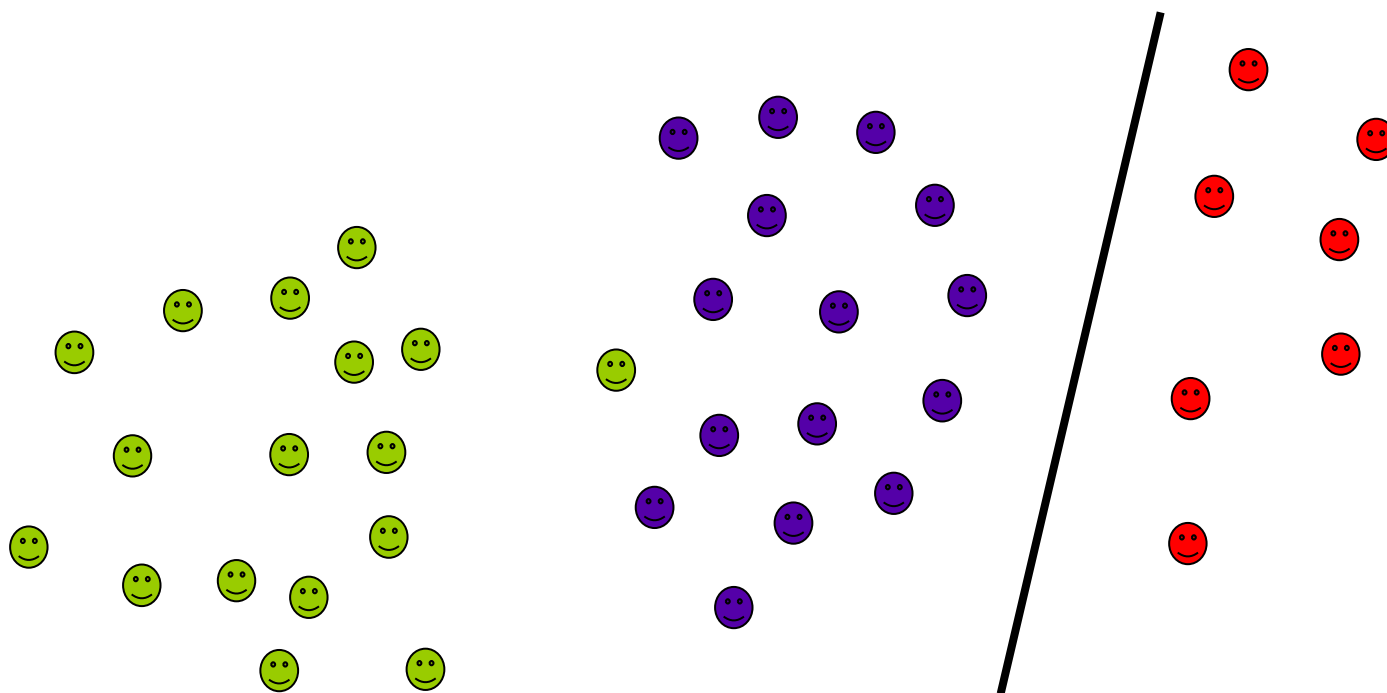
параметри: \mathbf{w} , \mathbf{w}_0 ,
пресмятане на грешката

😊 клас 1

😊 клас 2



Линеен класификатор



Линеен класификатор: $g(x) = \mathbf{w}x + \mathbf{w}_0$

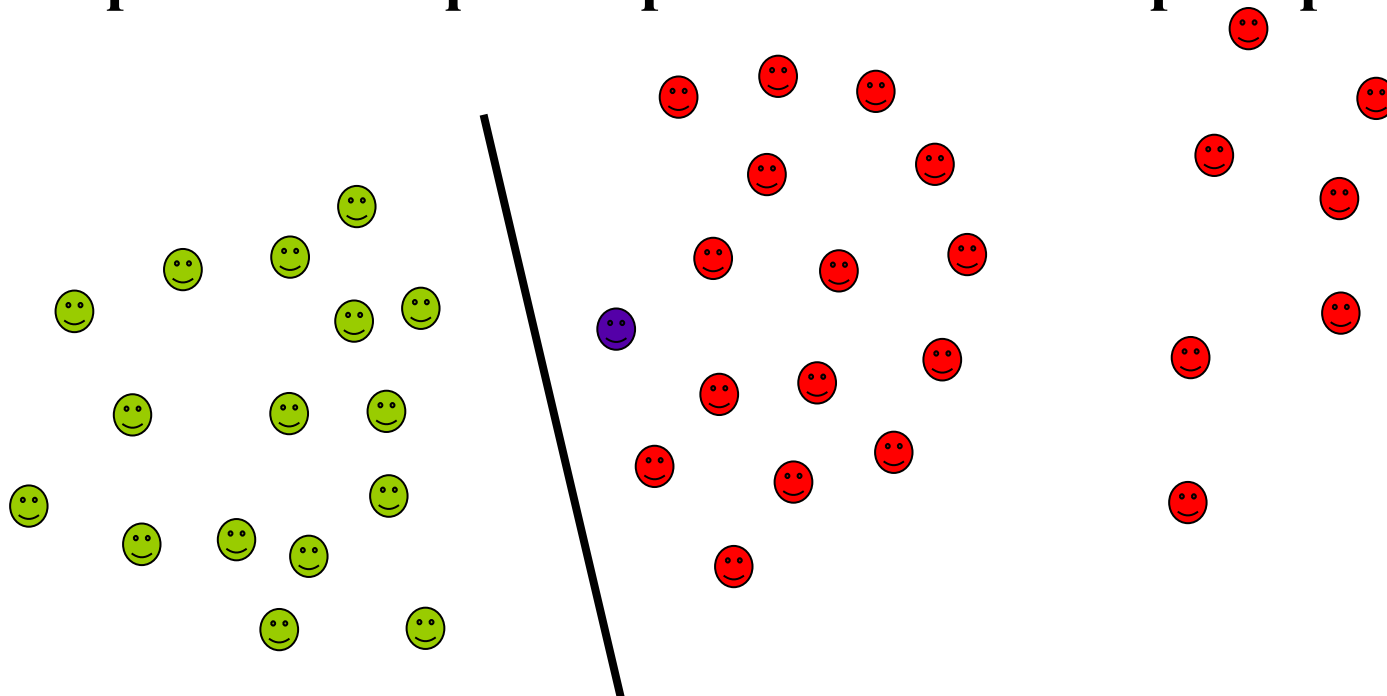
параметри: \mathbf{w} , \mathbf{w}_0 ,
пресмятане на грешката

😊 клас 1

😊 клас 2

Линеен класификатор

Избираме класификатора с минимален брой грешки.



Линеен класификатор: $g(x) = \mathbf{w}x + \mathbf{w}_0$

параметри: \mathbf{w} , \mathbf{w}_0 ,
пресмятане на грешката

😊 клас 1

😊 клас 2



Прости линейни класификатори

□ Персептрон и Winnow

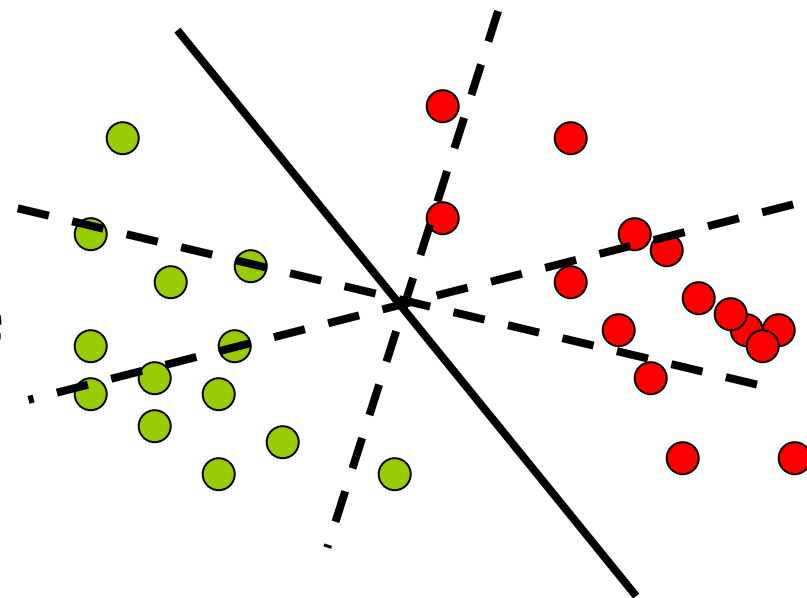
- Двоична класификация
- Online (данните се разглеждат последователно, пример по пример) – подходящо за Big Data
- Ръководят се от грешката

Двоична линейна класификация

- **Данни:** $\{(x_i, y_i)\}_{i=1\dots n}$
 - x от \mathbb{R}^d (x е вектор в d -мерно пространство)
→ feature vector
 - y от $\{-1, +1\}$
→ label (class, category)
- **Задача:**
 - Търсим линейна граница $\mathbf{w}x + \mathbf{b}$ (уравнение на хиперравнина), даващо минимална класификационна грешка
- **Класификационно правило:**
 - $y = \text{sign}(\mathbf{w}x + \mathbf{b})$ т.е.:
 - if $\mathbf{w}x + \mathbf{b} > 0$ then $y = +1$ (положителен пример)
 - if $\mathbf{w}x + \mathbf{b} < 0$ then $y = -1$ (отрицателен пример)

Двоична линейна класификация

- Търсим добра **хиперравнина** (\mathbf{w}, \mathbf{b}) от \mathbf{R}^{d+1} класифицираща данните максимално точно



$$\mathbf{w}\mathbf{x} + \mathbf{b} = 0$$

- Online**: пробваме с примерите един по един и променяме теглата, ако е нужно

Класификационно правило:
 $y = \text{sign}(\mathbf{w}\mathbf{x} + \mathbf{b})$

Линеен класификатор: Персептрон

\mathbf{x}_i - пример

y_i - клас: $\{-1;1\}$

- Инициализираме: $w_1 = 0$
- Актуализация за всеки пример x

- Ако клас(x) \neq избор(x, w)

- тогава

$$w_{k+1} \leftarrow w_k + \alpha y_i x_i$$

- иначе

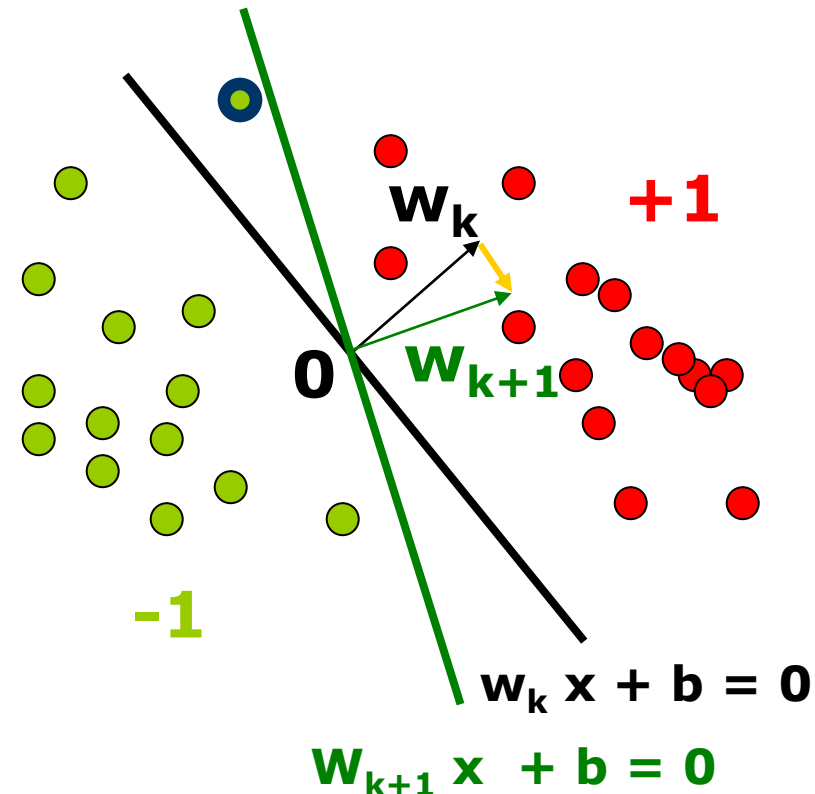
$$w_{k+1} \leftarrow w_k$$

$$k \leftarrow k + 1$$

- Функция **избор(x, w)**

- Ако $w x > 0$ върни +1

- Иначе върни -1



Приключва, щом намери **точно разделяне**

Персептрон

- **Online:** може да се адаптира към модел, който се променя във времето
- **Предимства**
 - Прост и “евтин” за смятане
 - Гарантирано намира решение, ако данните са линейно отделими
- **Недостатъци**
 - Работи само за линейно отделими данни
 - Не особено ефективен при много свойства

Winnow

- Друг online алгоритъм за учене на теглата:

$$\mathbf{f}(\mathbf{x}) = \text{sign}(\mathbf{w}\mathbf{x} + \mathbf{b})$$

- Линеен, двоична класификация
- Актуализация: отново се ръководи от грешката, но ползва **умножение** (вместо събиране)

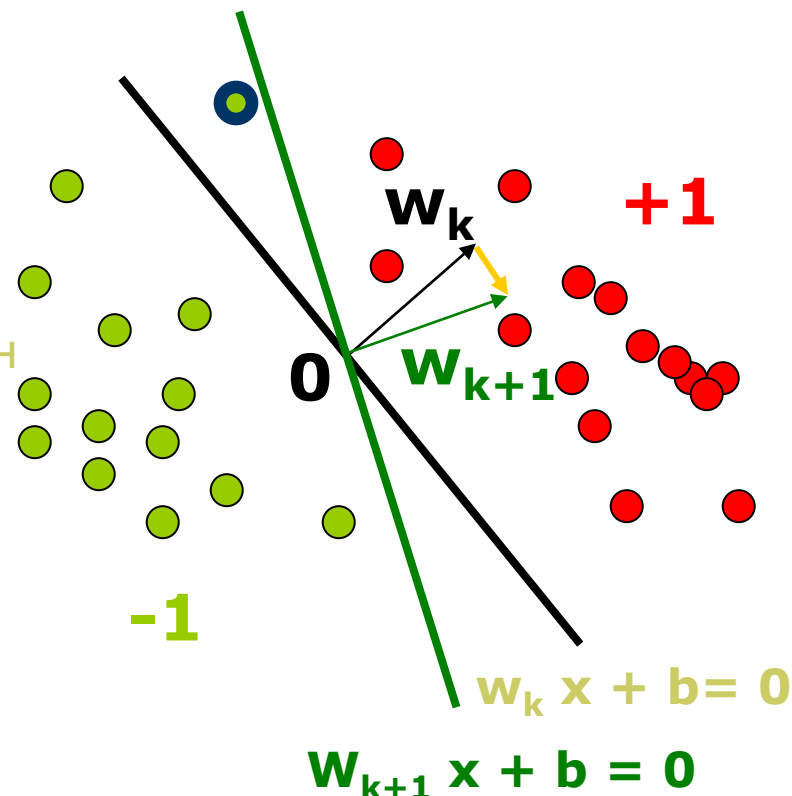
Линеен класификатор: Winnow

x_i - пример

y_i - клас: $\{-1;1\}$

- **Инициализация:** $w_1 = 0$
- **Актуализация** за всеки пример x
 - Ако $\text{клас}(x) \neq \text{избор}(x, w)$
 - тогава
$$w_{k+1} \leftarrow w_k + \alpha y_i x_i \quad \rightarrow \text{персептрон}$$
$$w_{k+1} \leftarrow \alpha w_k * \exp(y_i x_i) \rightarrow \text{Winnow}$$
 - иначе
$$w_{k+1} \leftarrow w_k$$
$$k \leftarrow k + 1$$

- Функция **избор**(x, w)
 - Ако $wx + b > 0$ върни +1
 - Иначе върни -1



Приключва, щом намери **точно разделяне**

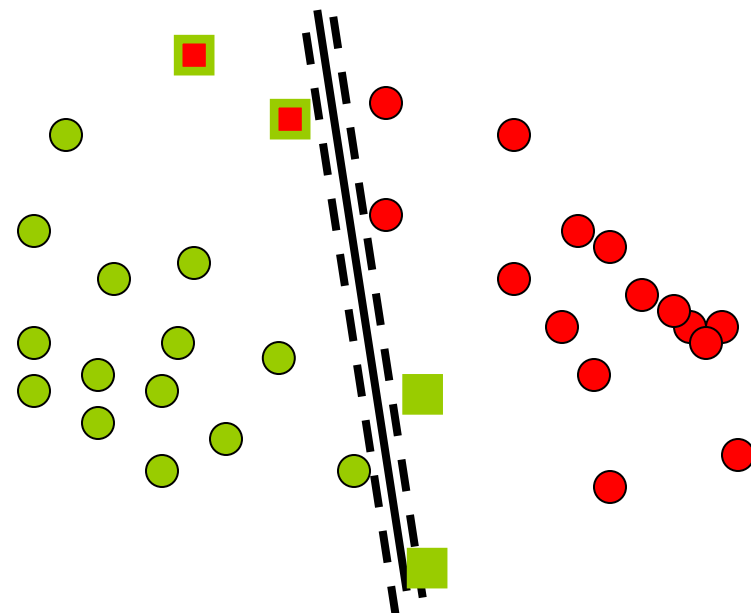
Персептрон и Winnow

- Нека имаме
 - M атрибута
 - само K са важни, $K \ll M$
- персептрон: брой грешки: $O(K N)$
- Winnow: брой грешки: $O(K \log N)$

Winnow се държи по-добре при пространства с голяма размерност.

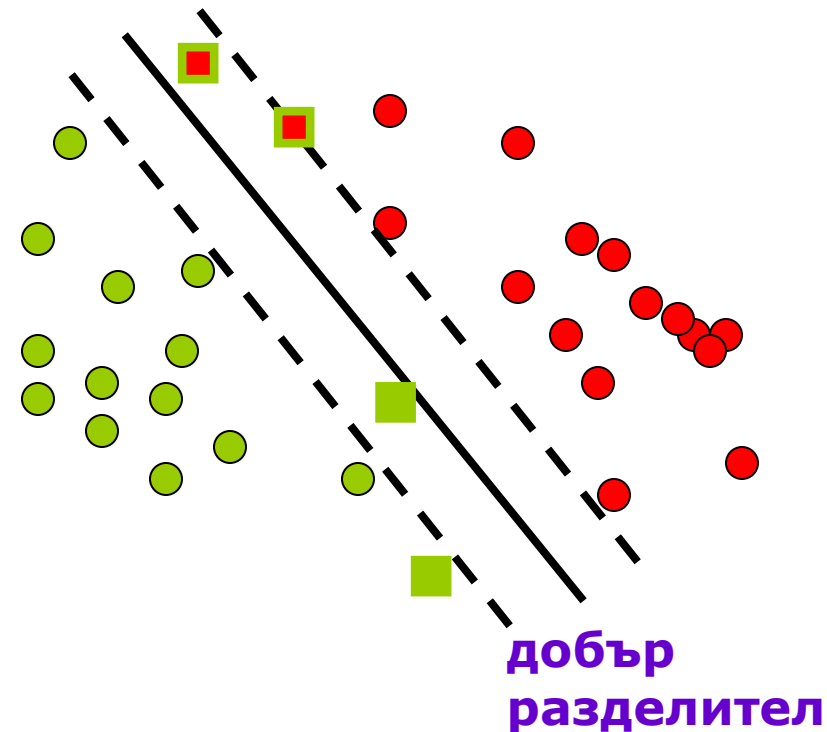
Класификатор с голяма разлика

- Друга фамилия класификатори
- Интуиция (Vapnik, 1965)
- Ако класовете са линейно разделими:
 - Разделяме данните
 - Хипер-равнина “далеч” от данните: **голяма разлика**
 - Статистическите резултати гарантират **добра генерализация**



Класификатор с голяма разлика

- Друга фамилия класификатори
- Интуиция (Vapnik, 1965)
- Ако класовете са линейно разделими:
 - Разделяме данните
 - Хипер-равнина “далеч” от данните: **голяма разлика**
 - Статистическите резултати гарантират **добра генерализация**

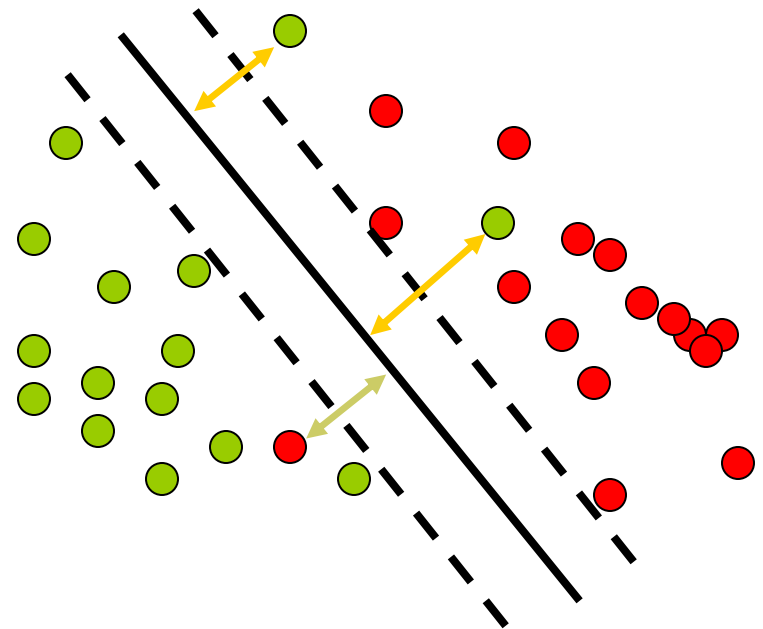


→ класификатор с максимална разлика

Класификатор с голяма разлика

Ако **не са линейно разделими**

- **Допускаме** някои грешки
- Но пак опитваме да прекараме хиперравнина “далеч” от всеки клас



Класификатор с голяма разлика

□ Предимства

- Теоретично по-добри (по-малко грешки)

□ Недостатъци

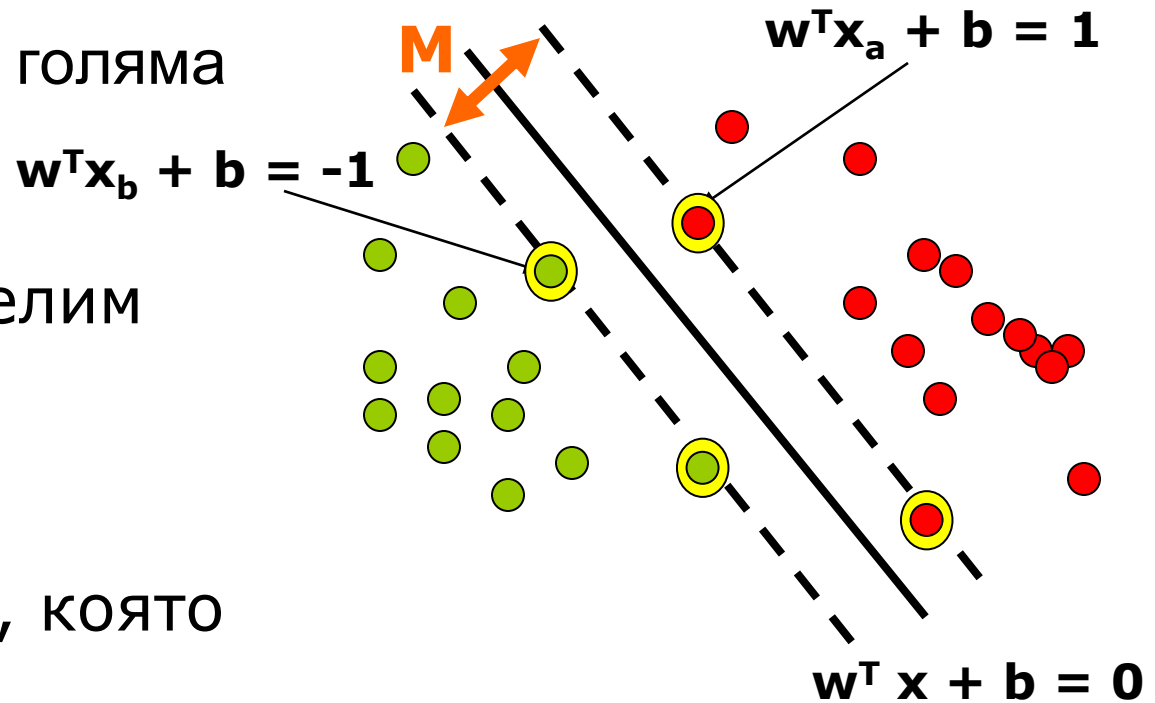
- По-трудни за смятане: квадратично програмиране

Support Vector Machine (SVM)

- Класификатор с голяма разлика

- Линейно разделим случай

- Цел: търсим хиперравнина, която максимизира разстоянието



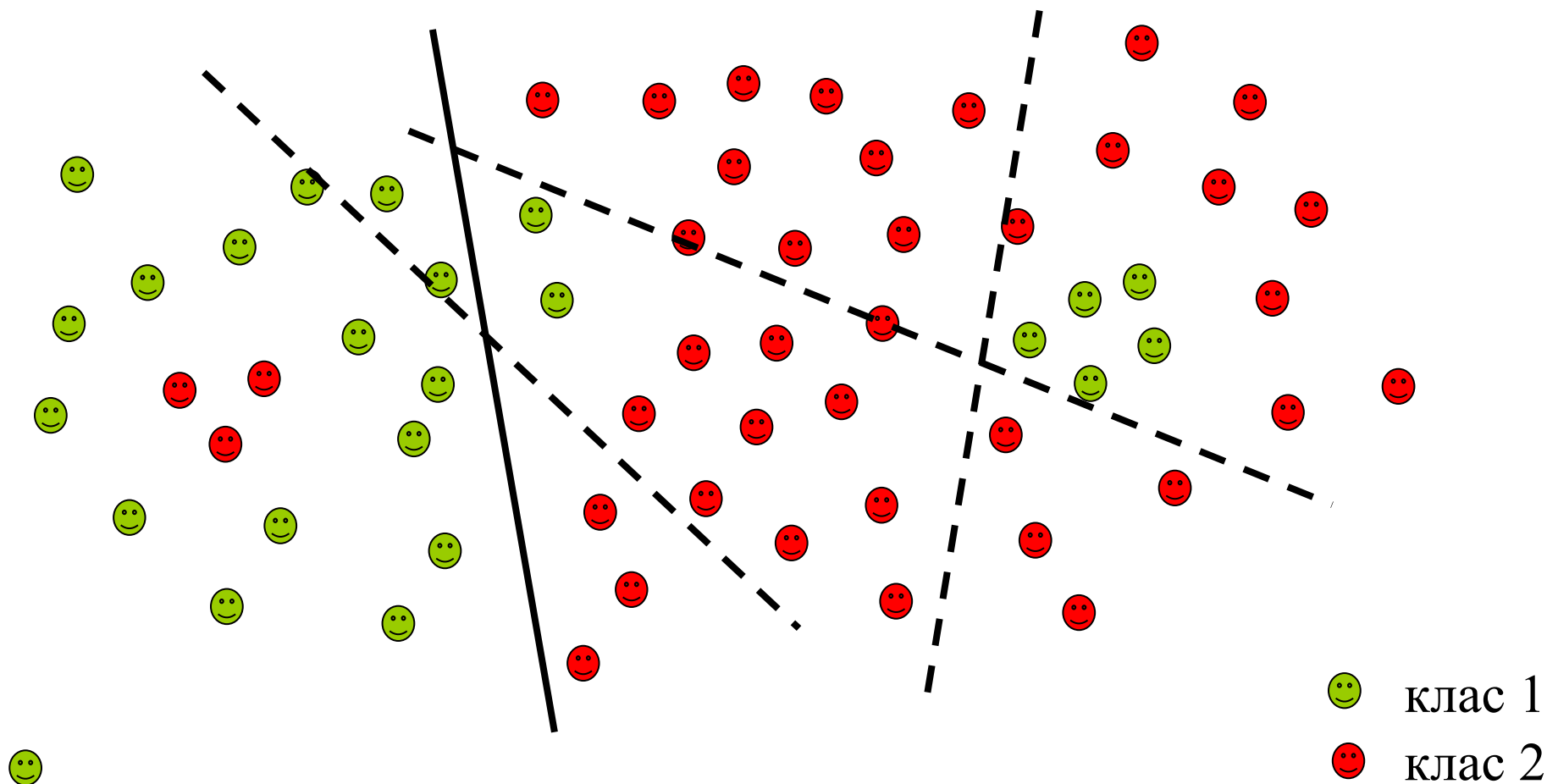
● Support vectors

- Квадратично програмиране

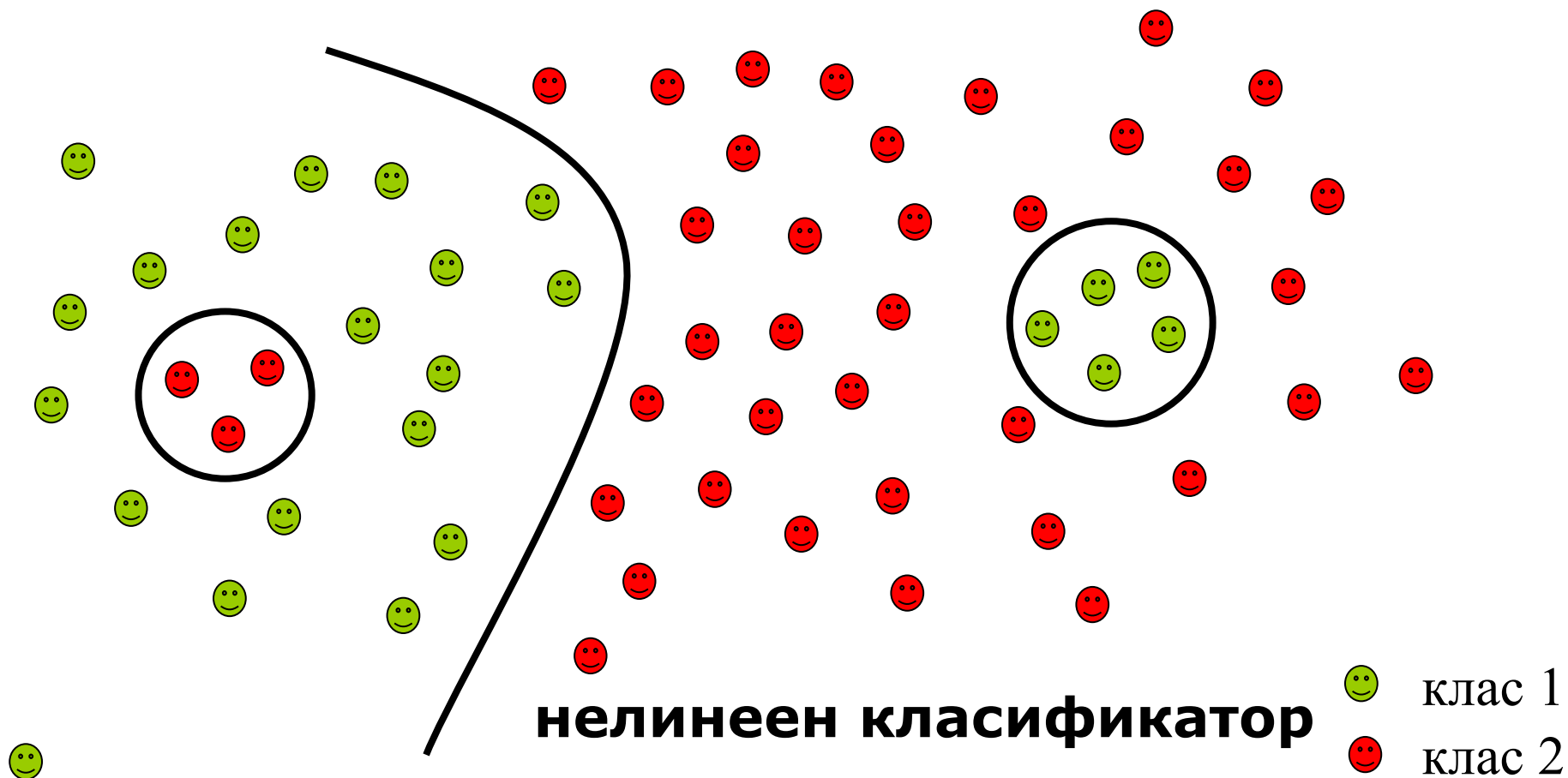
Support Vector Machines (SVM): Някои приложения

- Класифициране на текст
- Разпознаване на ръкописен текст
- Изчислителна биология (напр. анализ на последователности от нуклеотиди)
- Разпознаване на образи
- и др.

Линейно неразделими данни



Линейно неразделими данни



Линейни и нелинейни алгоритми

- Линейни или нелинейно отделими данни?
 - Определя се емпирично
- **Линейни алгоритми** (алгоритми, намиращи линейна граница между класовете)
 - Ако смятаме, че данните ни са линейно отделими
 - **Предимство**
 - По-прости, по-малко параметри
 - **Недостатък**
 - Многомерните данни най-често не са линейно отделими
 - **Все пак:**
 - понякога можем да използваме линейни алгоритми за нелинейни данни
 - Примери: SVM, Winnow, персептрон

Линейни и нелинейни алгоритми

□ Нелинейни алгоритми

- Когато данните не са линейно отделими
- **Предимство**
 - По-точни
- **Недостатък**
 - По-сложни, повече параметри
- Пример:
 - Kernel methods

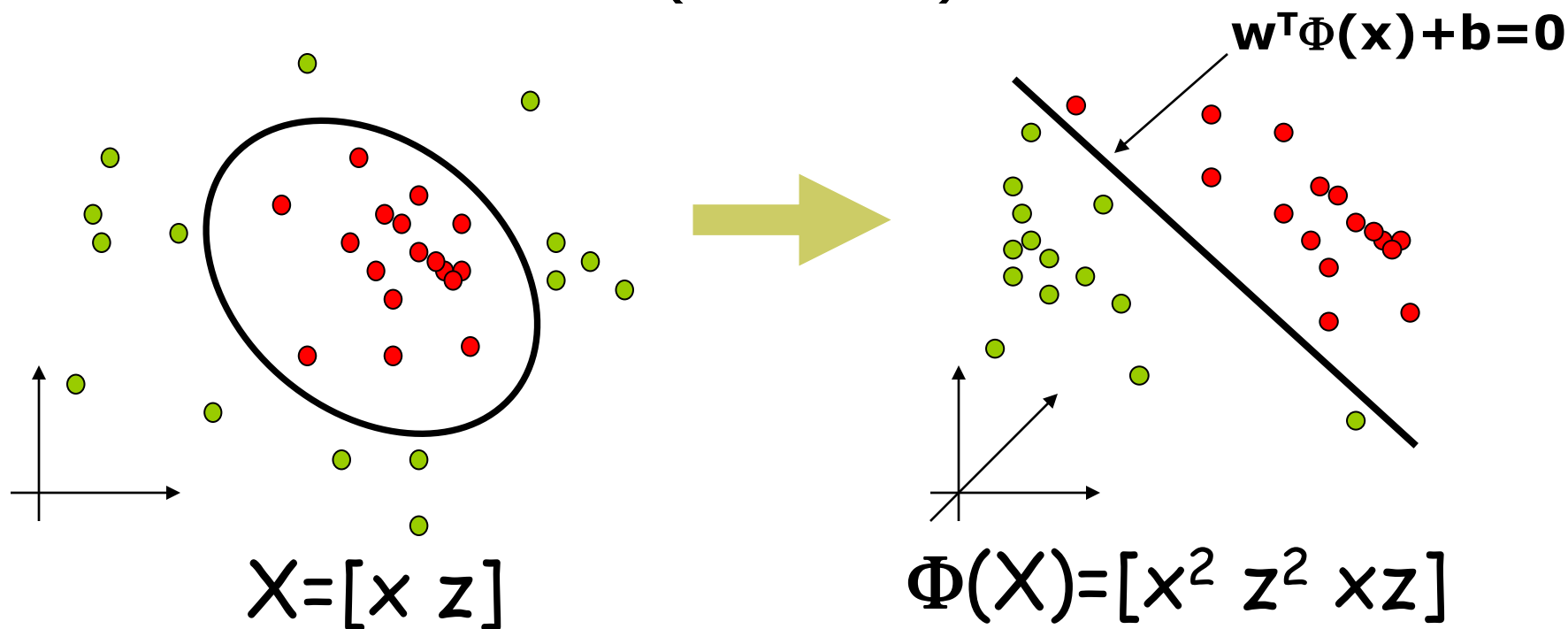
Разграничението между линейни и нелинейни алгоритми може да се прилага и за множествена класификация (ще видим по-късно)

Методи с ядро (Kernel methods)

- Фамилия от **нелинейни алгоритми**
- Трансформират нелинейно пространство в линейно (с други атрибути)
- После използват линейни алгоритми за решаване на задачата в новото пространство.

Основна идея на методите с ядро (**kernel** methods)

$$\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^D \quad (D \gg d)$$



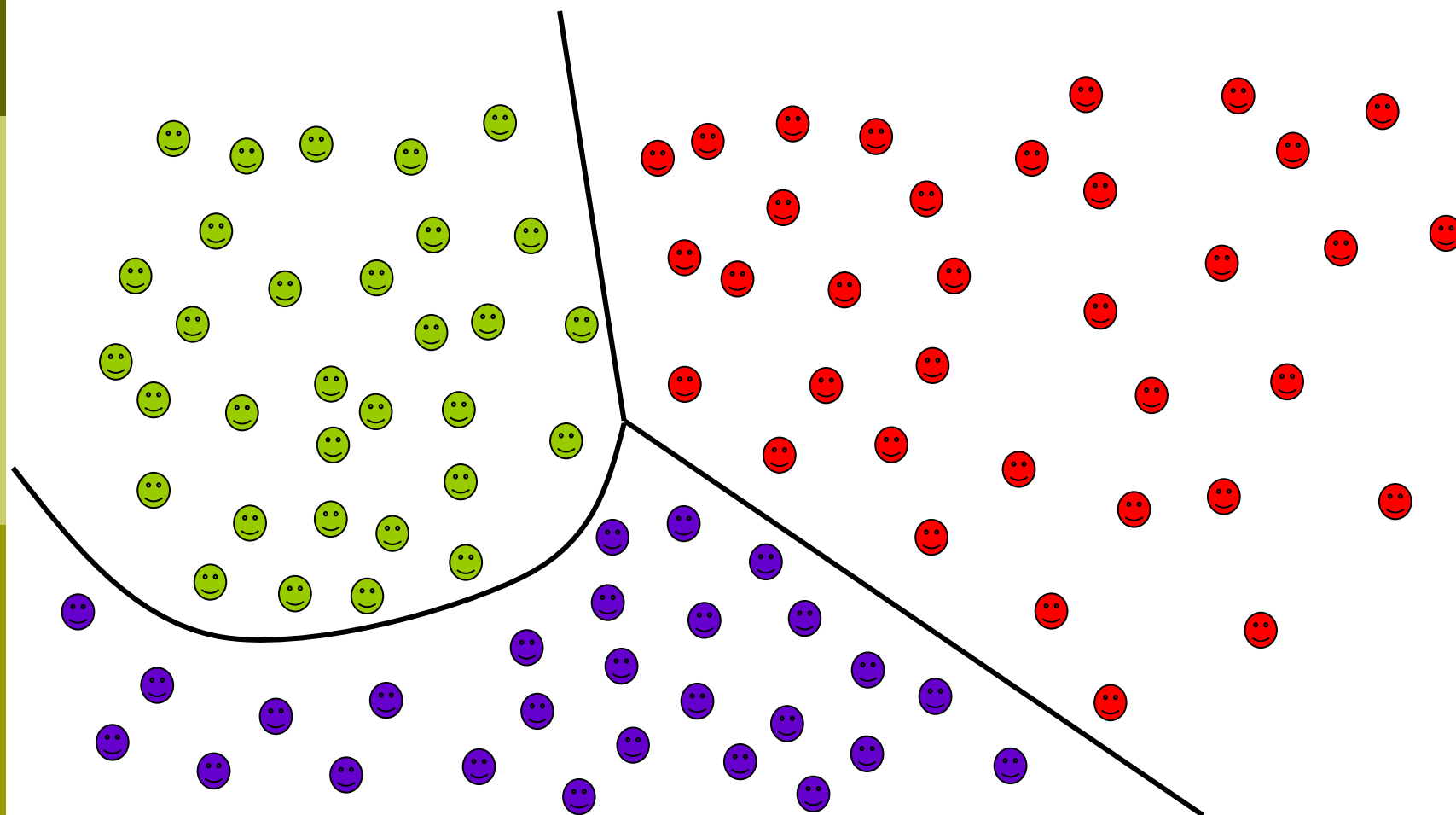
$$f(x) = \text{sign}(w_1 x^2 + w_2 z^2 + w_3 xz + b)$$

Класификатори

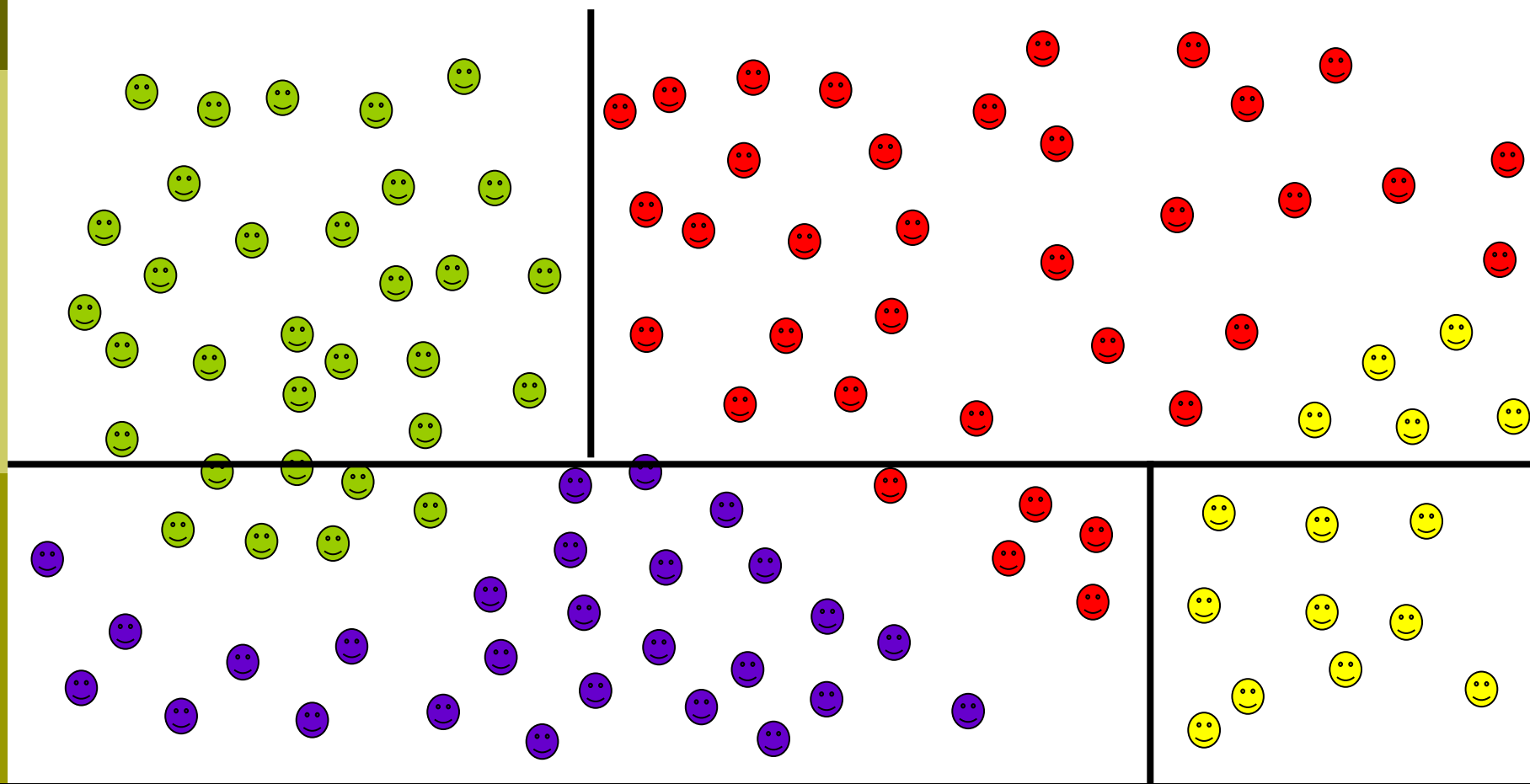


Множествена класификация

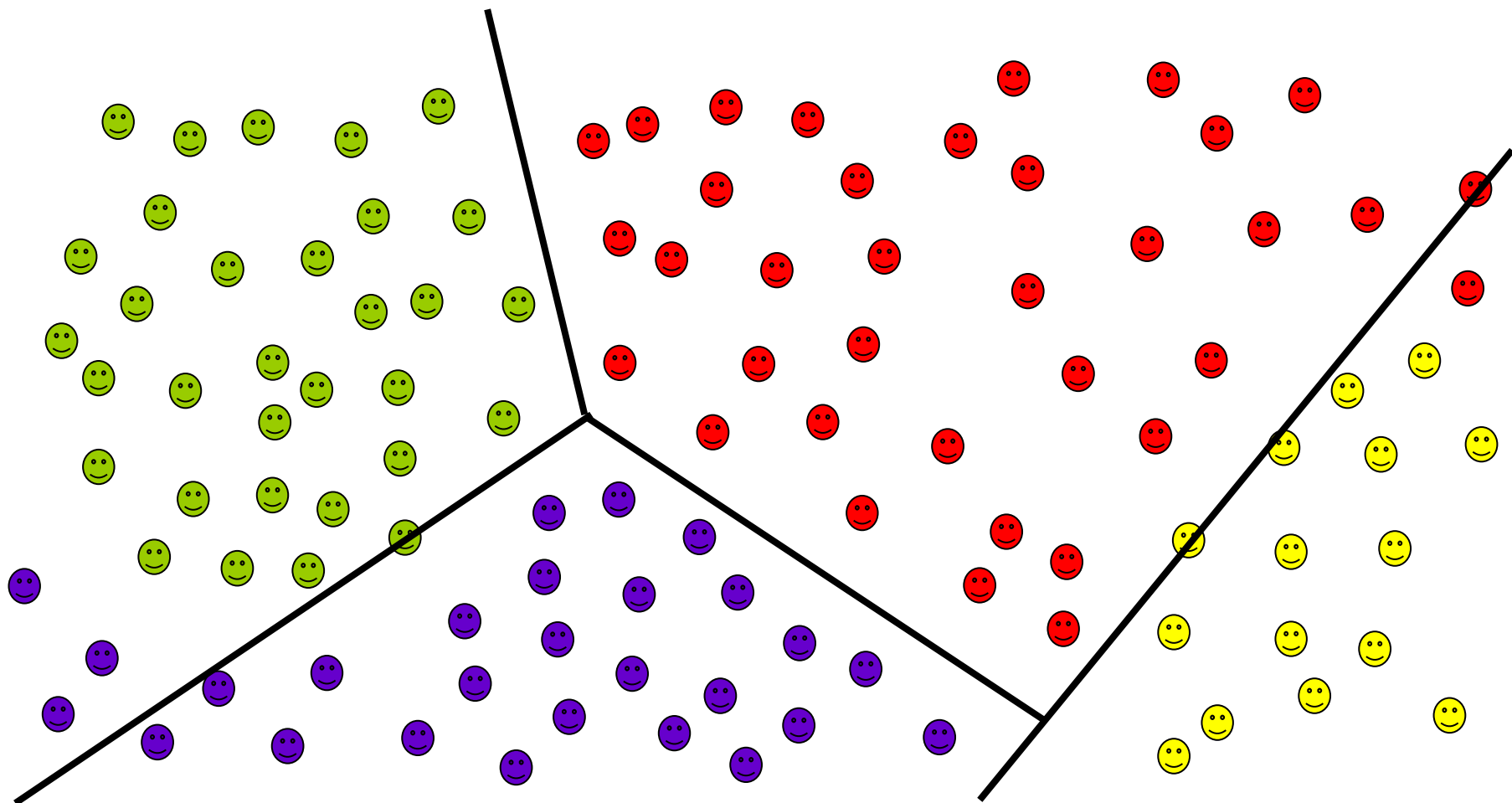
Множествена класификация



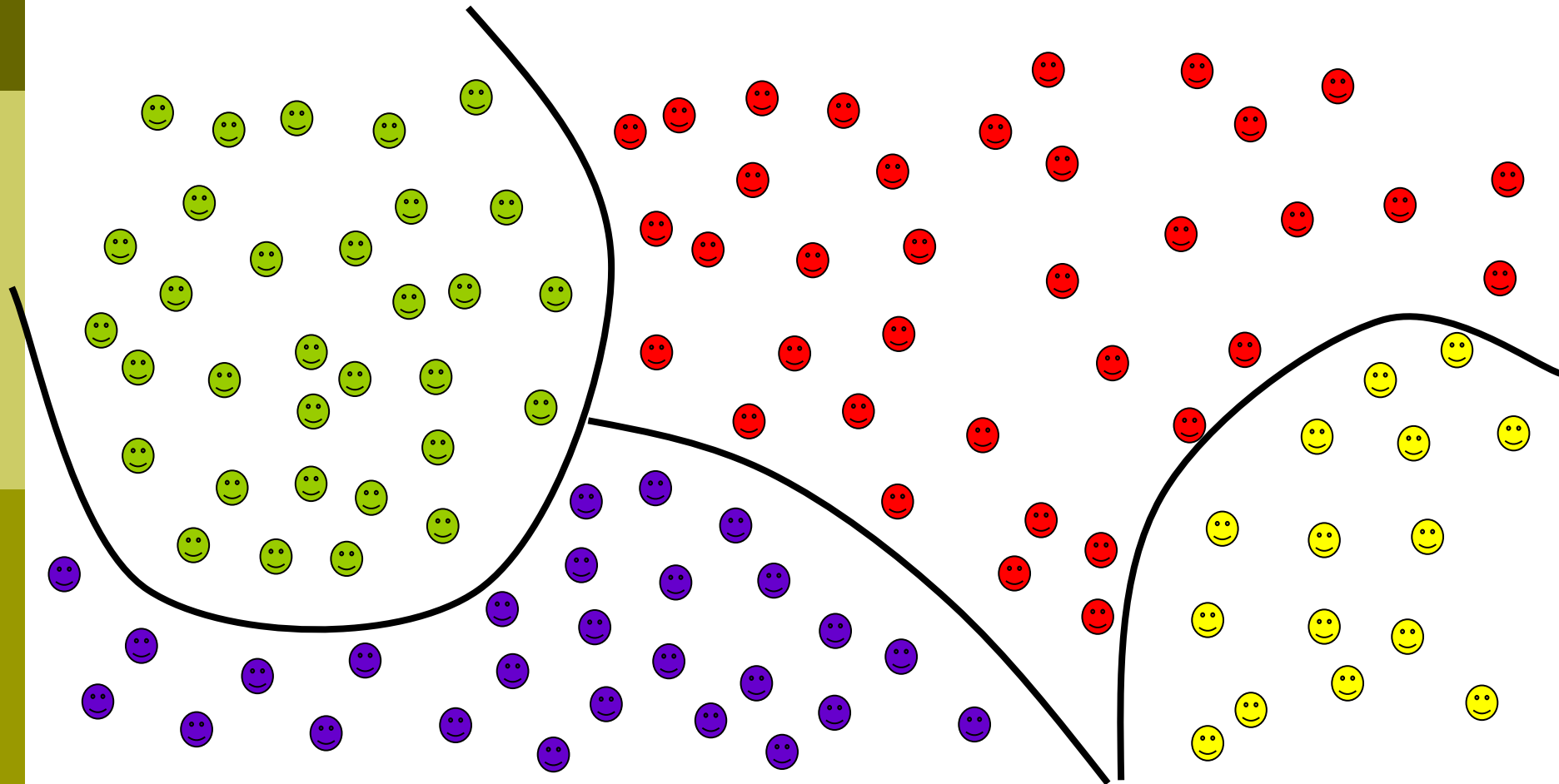
Линеен, паралелен разделител (дърво на решенията)



Линеен непаралелен разделител (напр. Naïve Bayes)



Нелинеен (напр. k -и най-близък съсед)

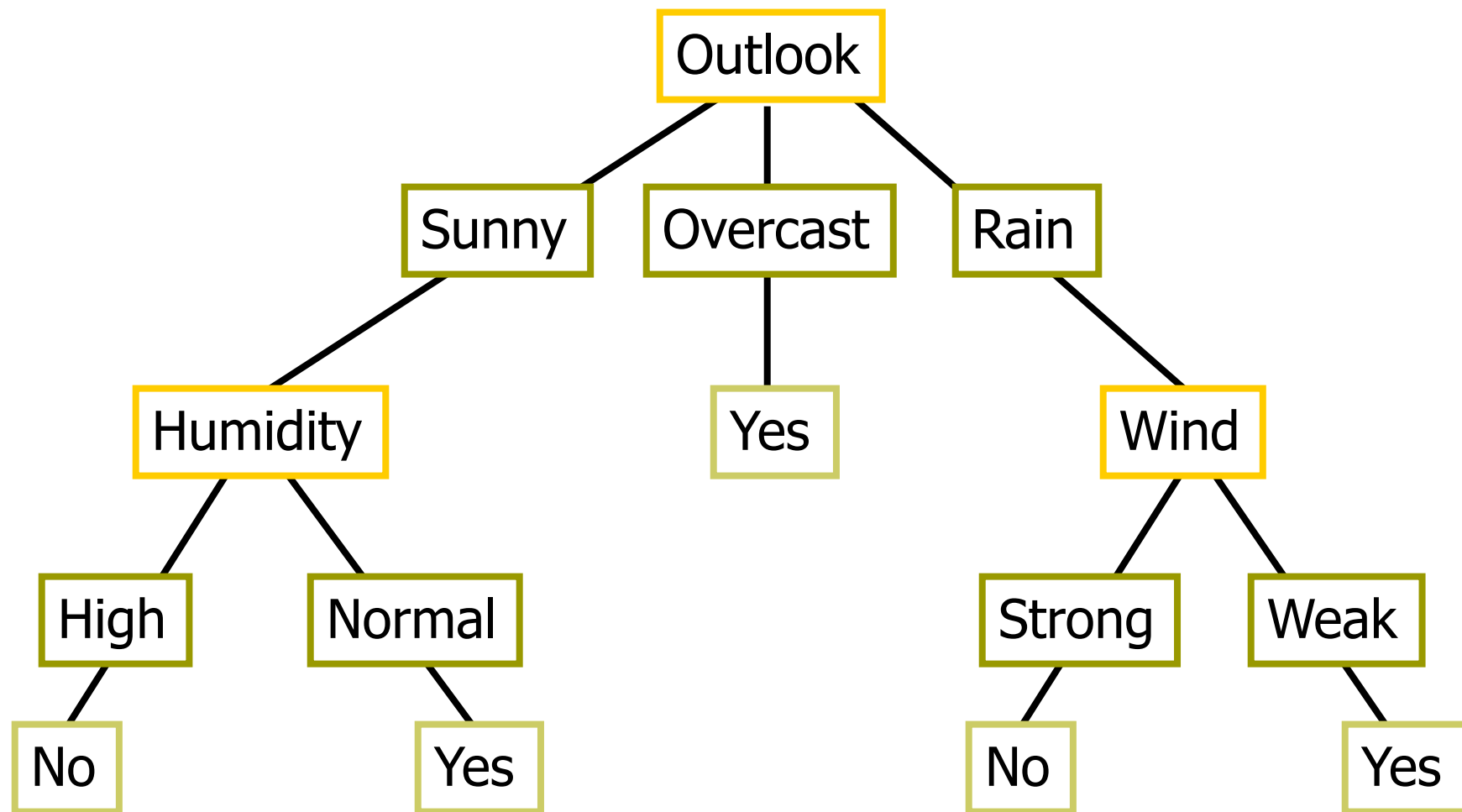


Учебни данни

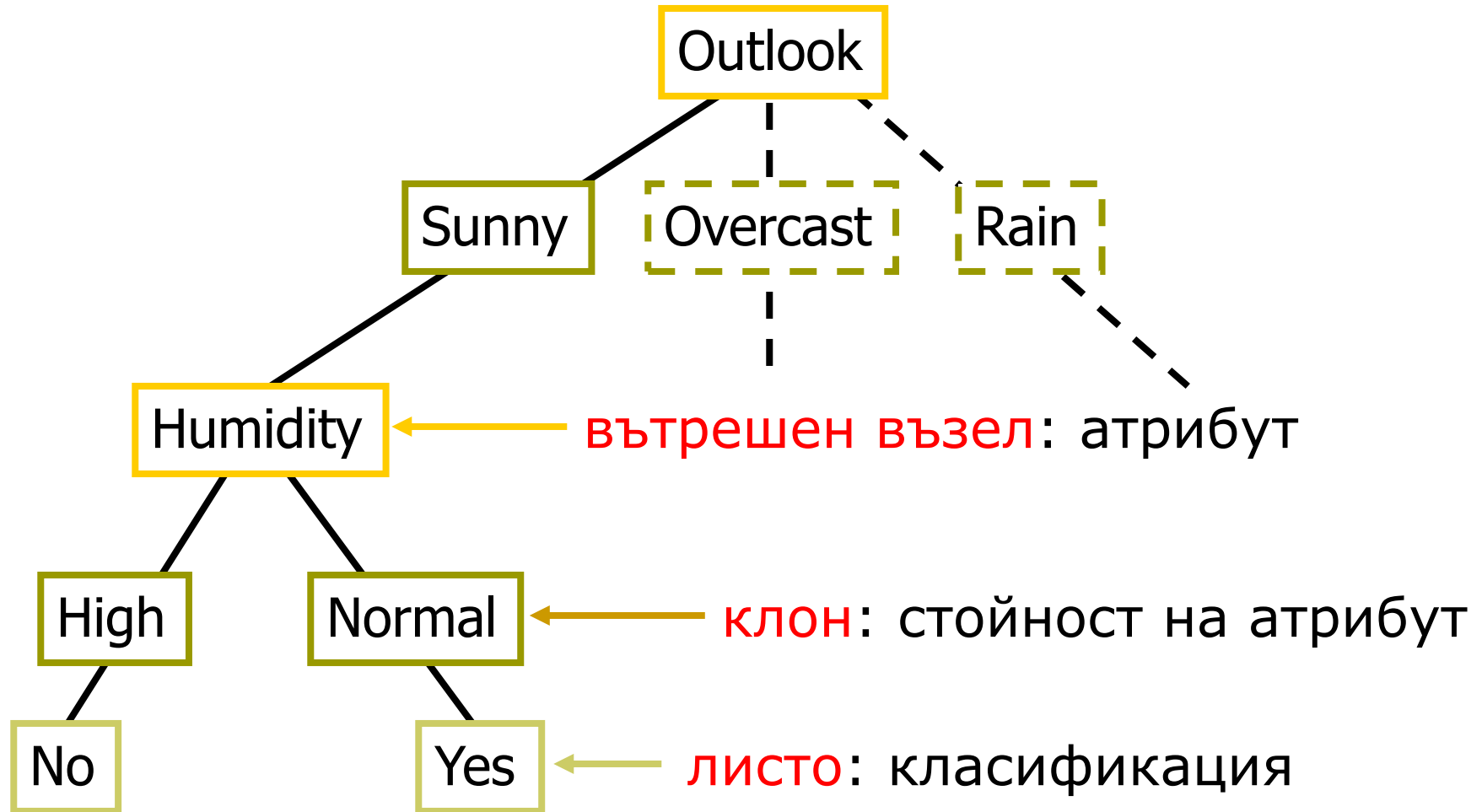
Goal: learn when we can play Tennis and when we cannot

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Класификационно дърво



Класификационно дърво



Пример от Reuters

```
<REUTERS NEWID="11">
<DATE>26-FEB-1987 15:18:59.34</DATE>
<TOPICS><D>earn</D></TOPICS>
<TEXT>
<TITLE>COBANCO INC &lt;CBCO> YEAR NET</TITLE>
<DATELINE> SANTA CRUZ, Calif., Feb 26 - </DATELINE>
<BODY>Shr 34 cts vs 1.19 dlrs
      Net 807,000 vs 2,858,000
      Assets 510.2 mln vs 479.7 mln
      Deposits 472.3 mln vs 440.3 mln
      Loans 299.2 mln vs 327.2 mln
      Note: 4th qtr not available. Year includes 1985
      extraordinary gain from tax carry forward of 132,000 dlrs,
      or five cts per shr.
      Reuter
</BODY></TEXT>
</REUTERS>
```

Figure 16.3 An example of a Reuters news story in the topic category “earnings.” Parts of the original have been omitted for brevity.

Класификационно дърво за Reuters

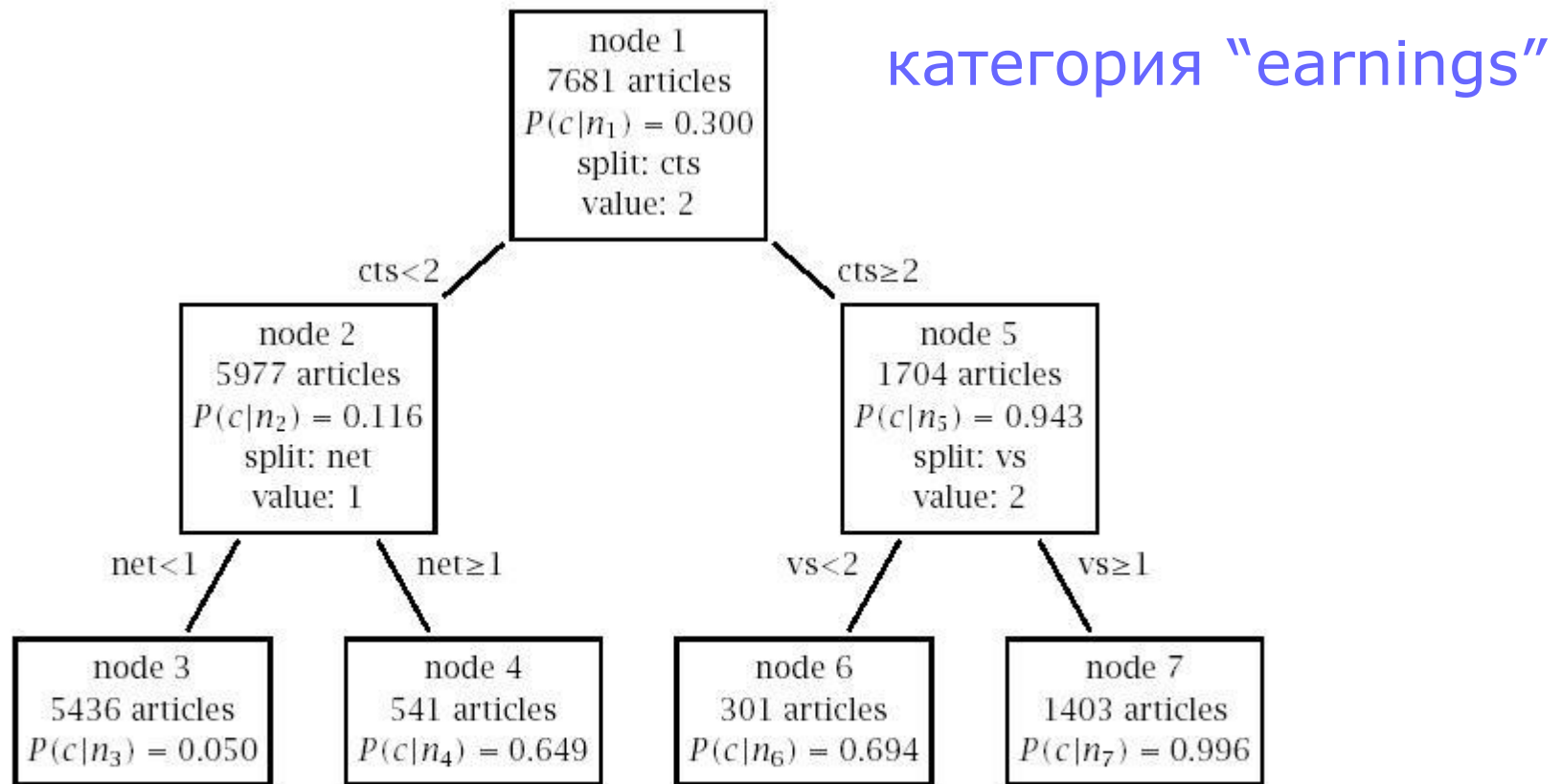
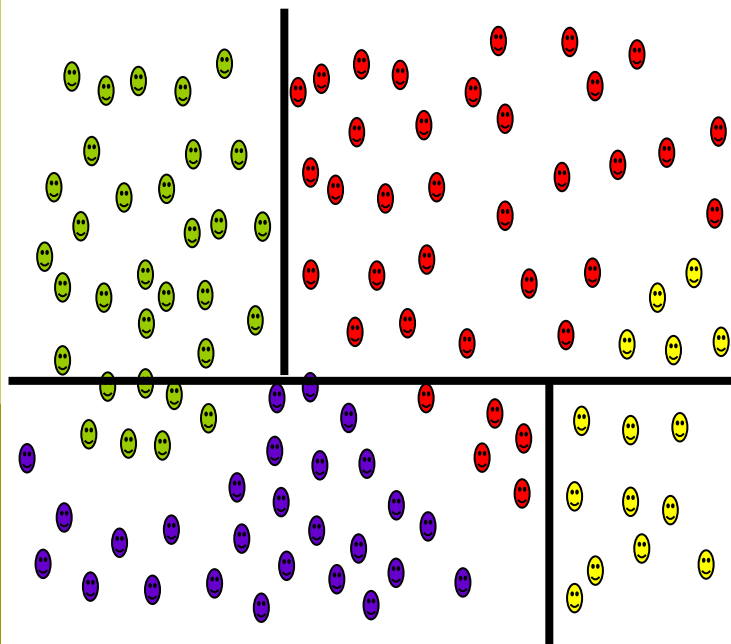


Figure 16.1 A decision tree. This tree determines whether a document is part of the topic category "earnings" or not. $P(c|n_i)$ is the probability of a document at node n_i to belong to the "earnings" category c .

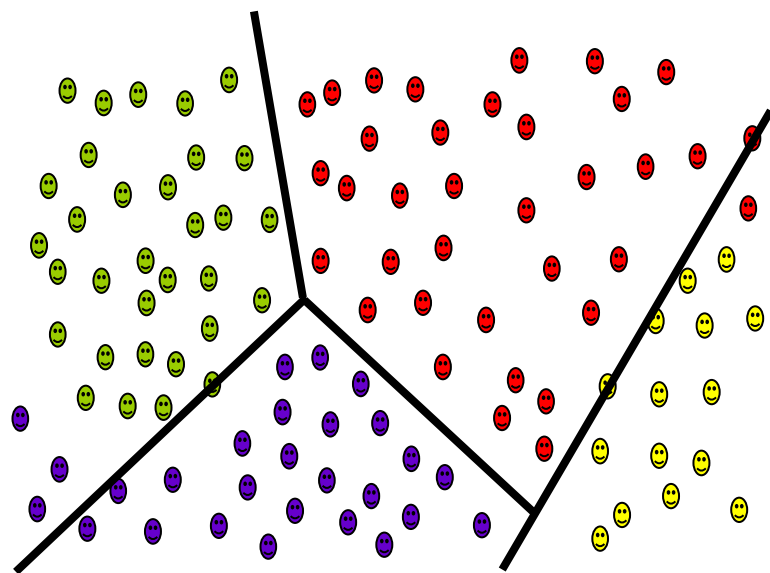
Наивен Бейсов класификатор

По-добър от класификационните дървета

класиф. дървета

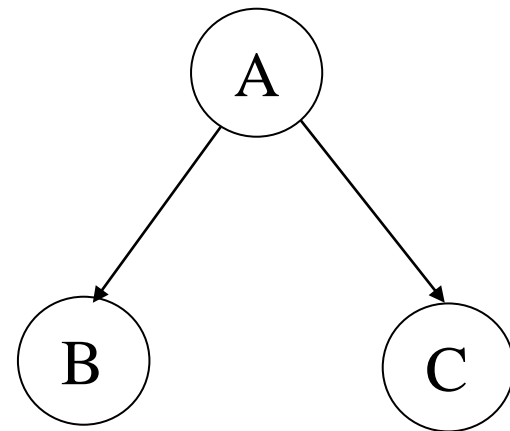


наивен Бейсов клас.



Bayesian Models

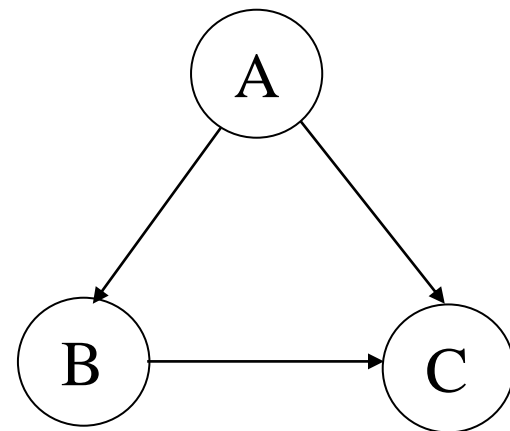
- ▣ Graphical Models:
graph theory plus
probability theory
- ▣ Nodes are variables
- ▣ Edges are conditional
probabilities



$P(A)$
 $P(B|A)$
 $P(C|A)$

Bayesian Models

- Graphical Models:
graph theory plus
probability theory
- Nodes are variables
- Edges are conditional
probabilities
- Absence of an edge
between nodes implies
independence
between the variables
of the nodes



$P(A)$

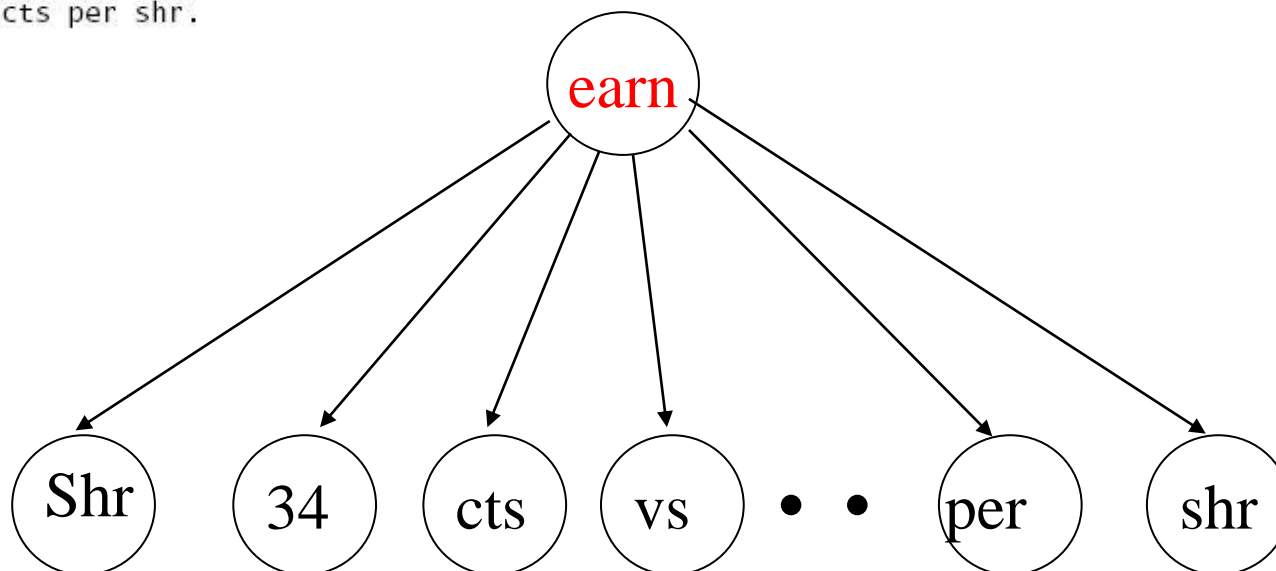
$P(B|A)$

$P(C|A)$

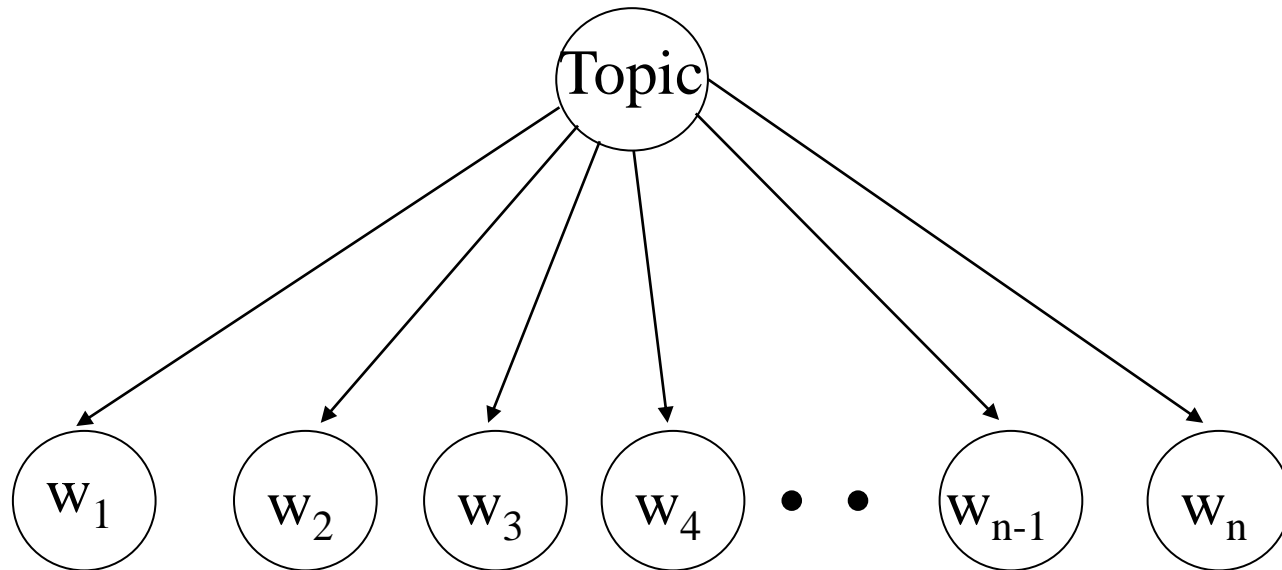
$\rightarrow P(C|A,B)$

Наивен Бейсов текстов класификатор

```
<TOPICS><D>earn</D></TOPICS>  
<TEXT>  
<TITLE>COBANCO INC &lt;CBCO> YEAR NET</TITLE>  
<DATELINE> SANTA CRUZ, Calif., Feb 26 - </DATELINE>  
<BODY>Shr 34 cts vs 1.19 dlrs  
    Net 807,000 vs 2,858,000  
    Assets 510.2 mln vs 479.7 mln  
    Deposits 472.3 mln vs 440.3 mln  
    Loans 299.2 mln vs 327.2 mln  
    Note: 4th qtr not available. Year includes 1985  
extraordinary gain from tax carry forward of 132,000 dlrs,  
or five cts per shr.  
    Reuter
```

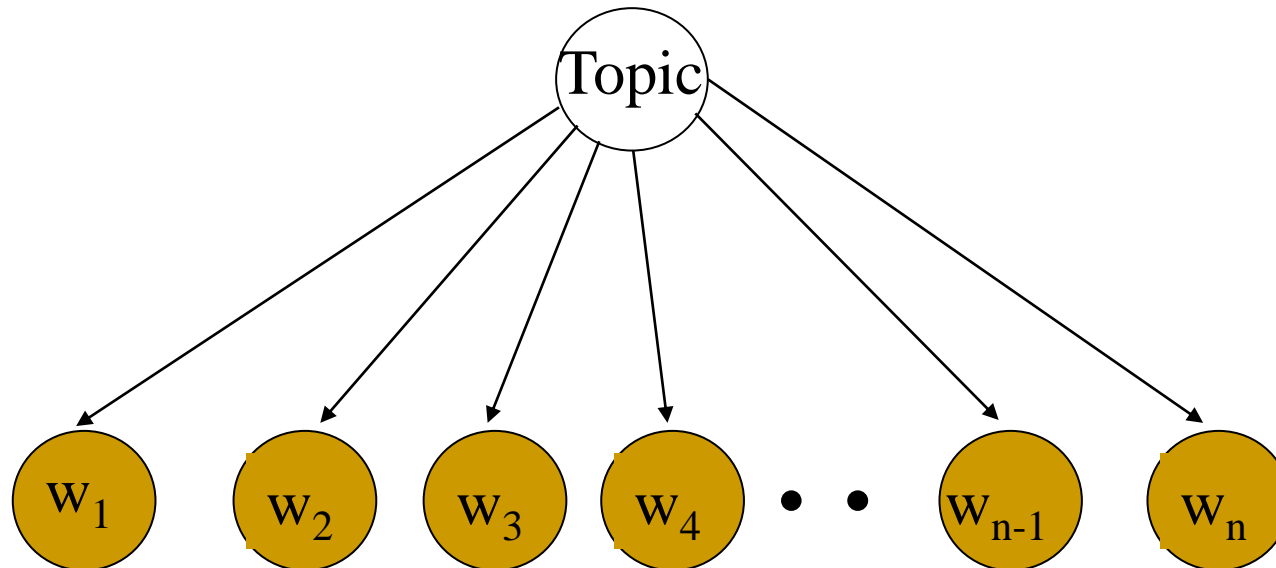


Naïve Bayes for Text Classification



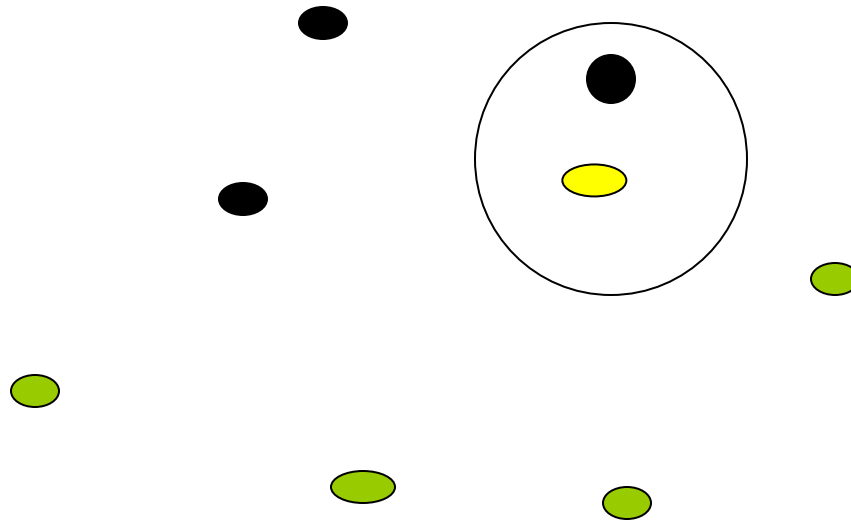
- The words depend on the topic: $P(w_i | \text{Topic})$
 - $P(\text{cts} | \text{earn}) > P(\text{tennis} | \text{earn})$
- Naïve Bayes assumption: all words are independent given the topic
- From training set we learn the probabilities $P(w_i | \text{Topic})$ for each word and for each topic in the training set

Наивен Бейсов текстов класификатор

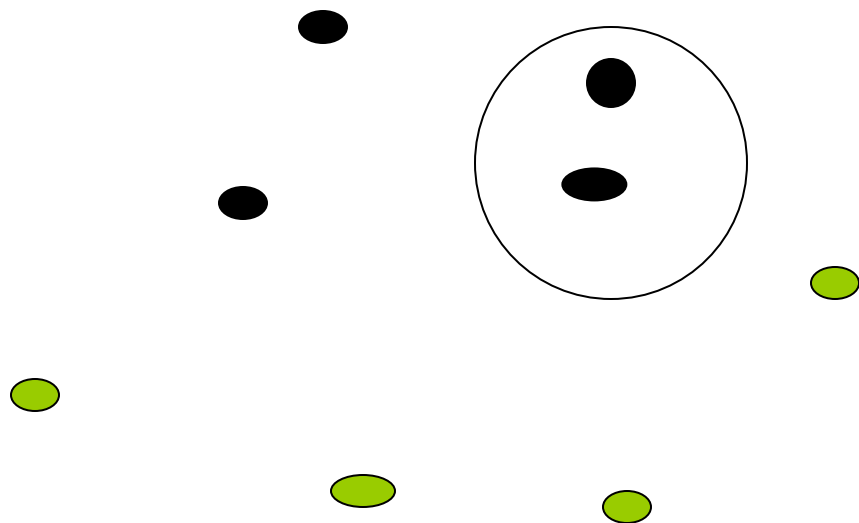


- За класифициране на нов пример
- Пресмятаме $P(\text{Topic} \mid w_1, w_2, \dots, w_n)$ за всяка тема
- Правило на Бейс:
 - Избор на тема T' , за която
 - $P(T' \mid w_1, w_2, \dots, w_n) > P(T \mid w_1, w_2, \dots, w_n)$ за всяко $T \neq T'$

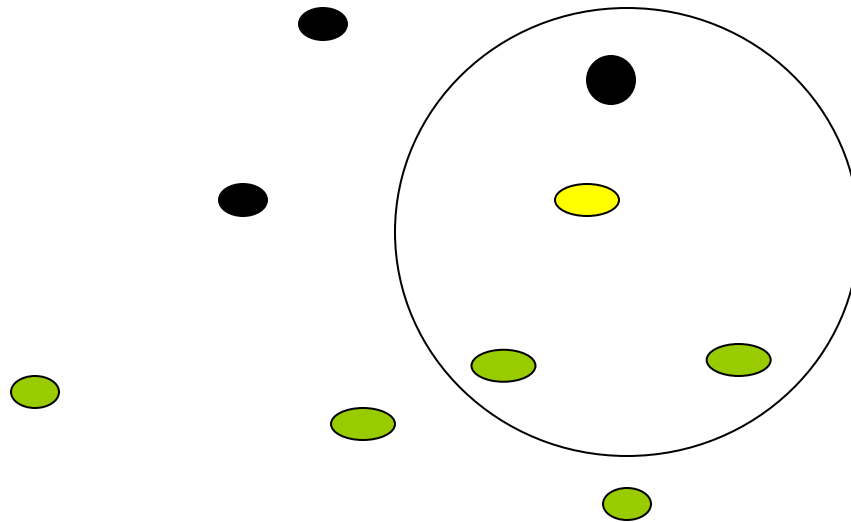
1 най-близък съсед (1-NN)



1 най-близък съсед (1-NN)



3 най-близки съседа (3-NN)



3 най-близки съседа (3-NN)

Но този пример е по-близо... Можем да претеглим съседите по тяхната прилика.



Избираме класа на мнозинството съседни.

Избор на атрибути



Атрибути

- избор на атрибути
- претегляне: TF.IDF
- нормализация на термини
- намаляване на размерността

Атрибути за класификация на текст

□ Лингвистични

■ думи

- малка/главна буква? (да правим ли разлика?)
- нормализирани? (напр. "думите" → "дума")

■ фрази

■ n -грами на ниво дума

■ пунктуация

■ части на речта

□ Нелингвистични

■ форматиране на документа

■ информативни последователности (напр. *<t*)

■ n -грами на ниво символ

Пример от Reuters: кат. “earnings”

```
<REUTERS NEWID="11">  
<DATE>26-FEB-1987 15:18:59.34</DATE>  
<TOPICS><D>earn</D></TOPICS>  
<TEXT>  
<TITLE>COBANCO INC &lt;CBCO> YEAR NET</TITLE>  
<DATELINE> SANTA CRUZ, Calif., Feb 26 - </DATELINE>  
<BODY>Shr 34 cts vs 1.19 dlrs  
      Net 807,000 vs 2,858,000  
      Assets 510.2 mln vs 479.7 mln  
      Deposits 472.3 mln vs 440.3 mln  
      Loans 299.2 mln vs 327.2 mln  
      Note: 4th qtr not available. Year includes 1985  
extraordinary gain from tax carry forward of 132,000 dlrs,  
or five cts per shr.  
      Reuter  
</BODY></TEXT>  
</REUTERS>
```


Пример от Reuters: кат. “earnings”

vs	5
mln	5
cts	3
;	3
&	3
000	4
loss	0
,	0
"	0
3	4
profit	0
dlrs	3
1	2
pct	0
is	0
s	0
that	0
net	3
lt	2
at	0

$\vec{x} =$

Добрите атрибути не
са непременно
лингвистично
смислени...

Документът се
представя като вектор
с координати
атрибутите му и
стойности – техните
честоти

Кога има нужда от избор на атрибути?

□ Когато алгоритъмът не може да работи с всичките

- напр. разпознаване на език за 100 езика с използване на всички думи от тези езици като атрибути
- класификация на текст с използване на n -грами

□ Добрите атрибути могат да доведат до по-добър класификатор

- И все пак – защо да **избираме** някои от тях?
- Защо да не запазим всички?
 - Дори ненадеждните атрибути могат да се окажат полезни.
 - Но трябва да ги **претеглим**:
 - В екстремалния случай, “лошите” атрибути ще имат тегло 0, което е форма на **избор на атрибути**

Защо да избираме атрибути?

- Не всички атрибути са еднакво добри
 - **лоши атрибути:** най-добре да се премахнат
 - **редки**
 - не очакваме да ги срещнем отново
 - съвместното им срещане с даден клас може да е случайно
 - **прекалено чести**
 - главно функционални (стоп) думи
 - **равномерно разпределени** между категориите
 - **добри атрибути:** трябва да ги запазим
 - срещат се с определена категория
 - **не** се срещат с определена категория
 - **останалите:** добре е да ги запазим

Избор на атрибути

- Манипулации с атрибути
 - Обикновено:
 - елиминиране на атрибути
 - претегляне на атрибути
 - нормализиране на атрибути
 - Понякога - **трансформация**
 - напр. латентен семантичен анализ
- Изборът зависи от
 - задачата
 - класификатора: kNN, невронна мрежа и др.
- Може да се използват примерните документи

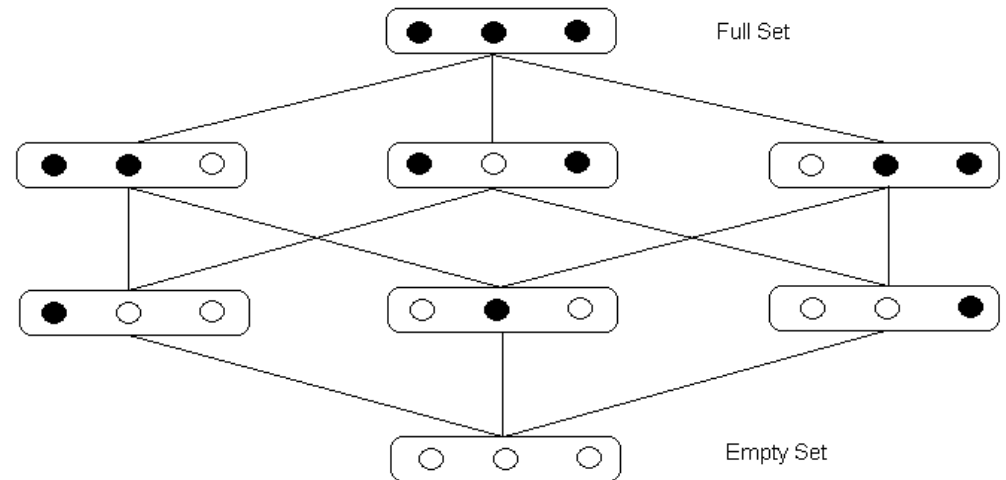
Зависимости между атрибутите

- ▣ Проблеми при изолирано разглеждане
 - ▣ няколко добри, но корелирани атрибута
 - напр. *Пало Алто, Сиера Леоне*
 - един от тях е достатъчен?
 - ▣ Един атрибут **се нуждае** от друг
 - *ябълки* и *круши*: нуждаем се от ширина и височина
- 2^n възможни подмножества

Избор на атрибути

■ Илюстрация на проблема

- Пълно множество
- Празно множество
- Изброяване



■ Търсене

- изчерпващо (изброяване/разклонения&граници)
- евристика (последователно напред/назад)
- стохастично (генериране+тестване)
- индивидуални атрибути или подмножества (генериране+тестване)

Избор на атрибути

- Независими от задачата методи
 - честота на срещане в документа (TF)
 - брой документи, съдържащи атрибута (DF)

- Зависими от задачата методи
 - Information Gain (IG)
 - взаимна информация (MI)
 - χ^2 статистика (CHI)

Емпирично сравнени от Yang & Pedersen (1997):

<http://dl.acm.org/citation.cfm?id=657137>

Брой документи, в които се съдържа (DF)

DF: брой документи, в които се среща атрибутът

- Закон на Зипф: честота * ранг = const
- Премахване на **редките**: (срещани 1-2 пъти)
 - неинформативни
 - ненадеждни – могат да бъдат просто шум
 - не влияят на крайното решение
 - едва ли ще се срещнат в нови документи
- Плюс
 - лесно се смята
 - **не зависи от задачата**: дори не се интересуваме кои са класовете
- Минус
 - евристичен подход
 - редките термини могат да бъдат добри дискриминатори

Ами честите?

Какво е "рядък" атрибут?

Най-чести думи в Brown Corpus

	Word	Instances	% Frequency		Word	Instances	% Frequency
1.	<u>The</u>	69970	6.8872	18.	<u>at</u>	5377	0.5293
2.	<u>of</u>	36410	3.5839	19.	<u>by</u>	5307	0.5224
3.	<u>and</u>	28854	2.8401	20.	<u>I</u>	5180	0.5099
4.	<u>to</u>	26154	2.5744	21.	<u>this</u>	5146	0.5065
5.	<u>a</u>	23363	2.2996	22.	<u>had</u>	5131	0.5050
6.	<u>in</u>	21345	2.1010	23.	<u>not</u>	4610	0.4538
7.	<u>that</u>	10594	1.0428	24.	<u>are</u>	4394	0.4325
8.	<u>is</u>	10102	0.9943	25.	<u>but</u>	4381	0.4312
9.	<u>was</u>	9815	0.9661	26.	<u>from</u>	4370	0.4301
10.	<u>He</u>	9542	0.9392	27.	<u>or</u>	4207	0.4141
11.	<u>for</u>	9489	0.9340	28.	<u>have</u>	3942	0.3880
12.	<u>it</u>	8760	0.8623	29.	<u>an</u>	3748	0.3689
13.	<u>with</u>	7290	0.7176	30.	<u>they</u>	3619	0.3562
14.	<u>as</u>	7251	0.7137	31.	<u>which</u>	3561	0.3505
15.	<u>his</u>	6996	0.6886	32.	<u>one</u>	3297	0.3245
16.	<u>on</u>	6742	0.6636	33.	<u>you</u>	3286	0.3234
17.	<u>be</u>	6376	0.6276	34.	<u>were</u>	3284	0.3232

Премахване на стоп-думите

- Често срещани думи от даден списък
 - главно думи от “затворен клас”
 - едва ли ще бъде добавена нова дума
 - включва: спомагателни глаголи, съюзи, предлози, местоимения, частици, квантификатори и др.
 - някои думи от “отворен клас”
 - напр. числителни
- Лоши дискриминатори
 - равномерно разпределени по класове
 - могат да бъдат премахнати от речника
 - Дали **винаги** е добра идея? (разпознаване на автор)

Статистика χ^2 (СНІ)

Дали "ягуар" предсказва надеждно клас "автомобили" ?

	дума = ягуар	дума \neq ягуар
клас = авто	2	500
клас \neq авто	3	9500

Сравняваме:

- **наблюдаваното** разпределение; и
- **нулева хипотеза**: терминът *ягуар* и класът *авто* са независими

Статистика χ^2 (CHI)

χ^2 гледа $(f_o - f_e)^2 / f_e$ сумирано по всички клетки:

$$\chi^2(j, a) = \sum (O - E)^2 / E = (2 - .25)^2 / .25 + (3 - 4.75)^2 / 4.75 + (500 - 502)^2 / 502 + (9500 - 9498)^2 / 9498 = 12.9 \quad (p < .001)$$

Нулевата хипотеза се отхвърля с ниво на доверие .999, тъй като $12.9 > 10.83$ (стойността за .999).

	дума = ягуар	дума \neq ягуар	
клас = авто	2 (0.25)	500 (502)	очаквано: f_e
клас \neq авто	3 (4.75)	9500 (9498)	наблюд.: f_o

$$P(\text{дума=ягуар}) = (5/10005) * P(\text{клас=авто}) = (502/10005) * 10005 = 0.25$$

Статистика χ^2 (СНІ)

□ Плюс

- нормализирана
- $\chi^2(t, c)$ е 0, когато t и c са независими
- може да се сравни с разпределение χ^2 , с 1 степен на свобода

□ Минус

- ненадеждна при редки атрибути
- "скъпа" за смятане

Информационна печалба

IG: брой битове информация, които печелим, ако знаем, че терминът присъства или отсъства

$$G(t) = - \sum_{i=1}^m P_r(c_i) \log P_r(c_i) \\ + P_r(t) \sum_{i=1}^m P_r(c_i|t) \log P_r(c_i|t) \\ + P_r(\bar{t}) \sum_{i=1}^m P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t})$$

t - термин,

c_i - клас

ентропия: $H(c)$

условна
ентропия
 $H(c|t)$

условна
ентропия
 $H(c|\neg t)$

Взаимна информация (MI)

- Вероятността да видим t и c заедно разделена на вероятността да видим t в текст и вероятността на c

$$I(t, c) = \log \frac{P_r(t \wedge c)}{P_r(t) \times P_r(c)} = \log P_r(t|c) - \log P_r(t)$$

редките
термини са
по-тежки

Приблизително: $I(t, c) \approx \log \frac{A \times N}{(A + C) \times (A + B)}$

$A = \#(t, c)$	$C = \#(\neg t, c)$
$B = \#(t, \neg c)$	$D = \#(\neg t, \neg c)$

Не използва
отсъствие
на термина

Взаимна информация

□ Плюс

- $I(t, c)$ е 0, когато t и c са независими
- връзка с теория на информацията

□ Минус

- малки числа – ненадеждни резултати
- “скъпа” за пресмятане
- **не използва отсъствието на термин**

Множество категории

Как се ползват χ^2 и MI за много категории?

Пресмятаме напр. χ^2 за всяка категория и комбинираме:

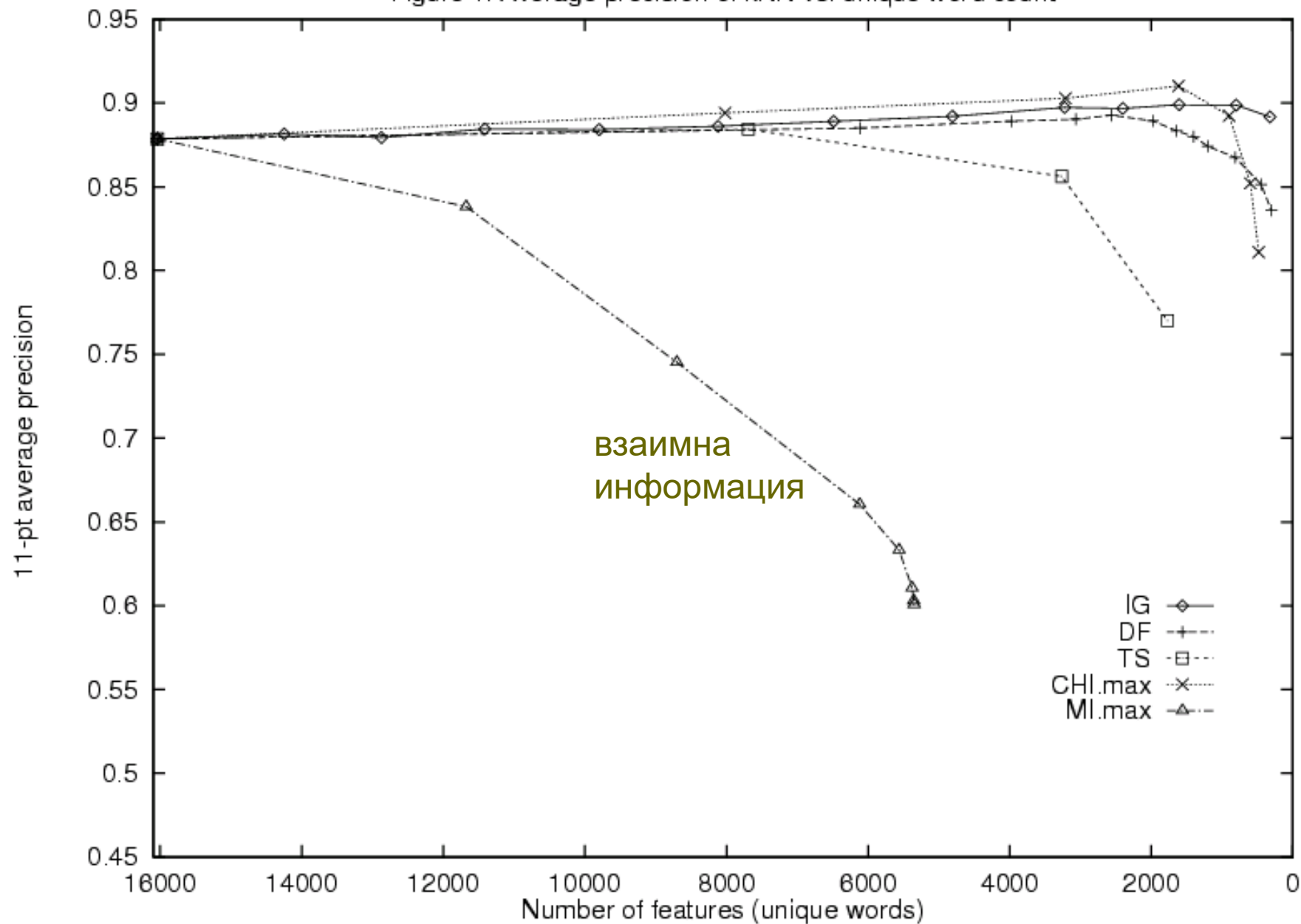
- **всички категории:** математическо очакване

$$\chi_{avg}^2(t) = \sum_{i=1}^m P_r(c_i) \chi^2(t, c_i)$$

- **една категория:** максимум по класовете

$$\chi_{max}^2(t) = \max_{i=1}^m \{\chi^2(t, c_i)\}$$

Figure 1. Average precision of kNN vs. unique word count



взаимна
информация

Сравнение: DF, TS, IG, CHI, MI

Method	DF	IG	CHI	MI	TS
favoring common terms	Y	Y	Y	N	Y/N
using categories	N	Y	Y	Y	N
using term absence	N	Y	Y	N	N
performance in kNN/LLSF	excellent	excellent	excellent	poor	ok

■ DF, IG and CHI за добри и силно корелирани

- DF е добра, евтина и независима от задачата
- DF се предпочита, когато IG и CHI са скъпи

■ MI е лоша

- предпочита редки атрибути (които са лоши)

■ MI vs. IG

Информационна печалба

$$G(t) = \sum_{i=1}^n P_r(t, c_i) I(t, c_i) + \sum_{i=1}^n P_r(\bar{t}, c_i) I(\bar{t}, c_i)$$

взаимна информация

Претегляне на термини

- При избора на атрибут, предполагахме, че е булев
 - 1, ако терминът се съдържа в документа
 - 0, иначе
- Добре е да ги претеглим
- Стандартна техника: TF.IDF

Претегляне на термини с TF.IDF

□ TF: честота на термин

- дефиниция : $TF = t_{ij}$
 - Честота на термин i в документ j
- цел: прави по-тежки думите от *документа*

□ IDF: инвертирана честота на документ

- дефиниция : $IDF = \log(N/n_i)$
 - n_i : брой документи, съдържащи термина i
 - N : общ брой документи
- цел: прави тежки думите, които се съдържат в *малко документи*

□ TF.IDF

- дефиниция: $t_{ij} \times \log(N/n_i)$

Нормализиране на термини

- Комбиниране на различни форми на дума
 - Стеминг/морфологичен анализ
 - bought, buy, buys -> buy
 - Общи категории
 - \$23.45, 5.30 Yen -> MONEY
 - 1984, 10,000 -> DATE, NUM
 - PERSON
 - ORGANIZATION
 - LOCATION
 - Гупиране по лексикална йерархия
 - WordNet, MeSH

Стеминг & лематизация

- Цел: групира морфологичните варианти
 - **Стеминг:** *псевдодума*
 - напр. "more" и "morals" стават "mor" (стемър на Портър)
 - **Лематизация:** *основна форма*
 - напр. "more" и "morals" стават "more" и "moral"
- **Плюс:**
 - намалява размера на речника
 - намалява броя на редките термини
- **Минус:**
 - губят се важни атрибути (често дори `to_lowercase()` е лошо!)
 - съмнителна полза (може би просто "-s", "-ing" и "-ed"?)

Какво се прави на практика?

1. Избор на атрибути

- ▣ премахване на редките
 - редки в цялата колекция (т.е. DF)
 - срещнати в единствен документ
- ▣ премахване на най-честите (стоп-думи)

2. Нормализиране:

- ▣ стеминг (*често*)
- ▣ класове думи (*понякога*)

3. Претегляне на атрибути: TF.IDF или IDF

4. Намаляване на размерността. (*понякога*)

Векторен пространствен модел

Пример от Reuters: кат. "earnings"

```
<REUTERS NEWID="11">  
<DATE>26-FEB-1987 15:18:59.34</DATE>  
<TOPICS><D>earn</D></TOPICS>  
<TEXT>  
<TITLE>COBANCO INC &lt;CBCO> YEAR NET</TITLE>  
<DATELINE> SANTA CRUZ, Calif., Feb 26 - </DATELINE>  
<BODY>Shr 34 cts vs 1.19 dlrs  
      Net 807,000 vs 2,858,000  
      Assets 510.2 mln vs 479.7 mln  
      Deposits 472.3 mln vs 440.3 mln  
      Loans 299.2 mln vs 327.2 mln  
      Note: 4th qtr not available. Year includes 1985  
      extraordinary gain from tax carry forward of 132,000 dlrs,  
      or five cts per shr.  
      Reuter  
</BODY></TEXT>  
</REUTERS>
```

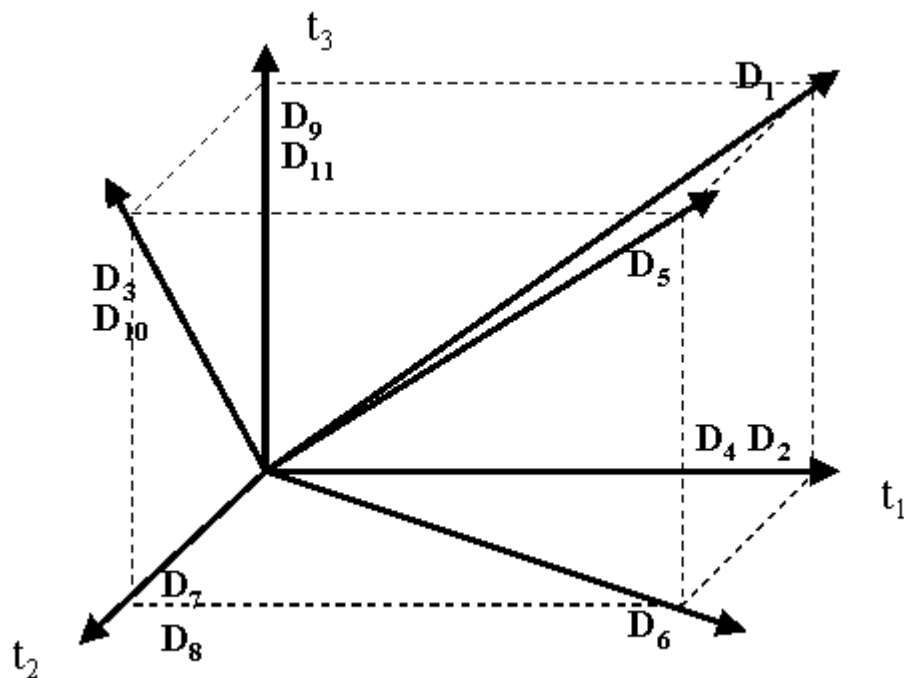
Пример от Reuters: кат. "earnings"

vs	5
mln	5
cts	3
;	3
&	3
000	4
loss	0
,	0
"	0
3	4
profit	0
dlrs	3
1	2
pct	0
is	0
s	0
that	0
net	3
lt	2
at	0

$\vec{x} =$

Документът се представя
като вектор с координати
атрибутите му и стойности
– техните честоти

Векторен пространствен модел



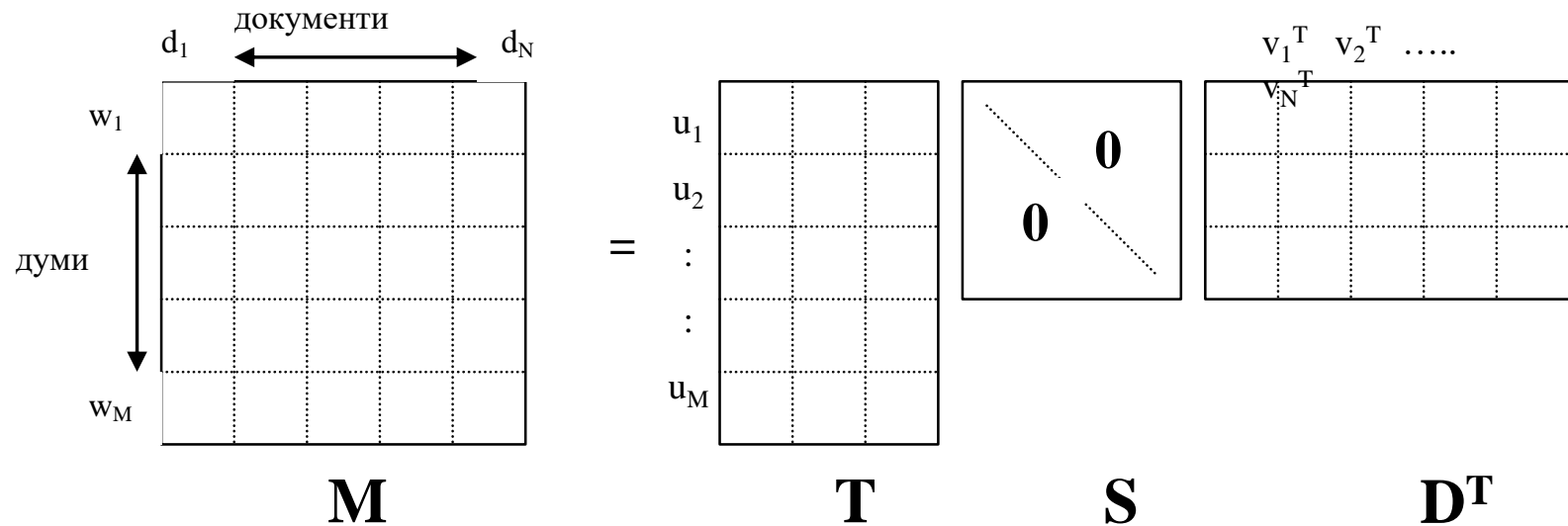
Прилика:

$$\cos \theta = \frac{\mathbf{X} \bullet \mathbf{Y}}{\|\mathbf{X}\| * \|\mathbf{Y}\|}$$

Документът се представя като вектор с координати атрибутите му и стойности – техните честоти

Латентен семантичен анализ

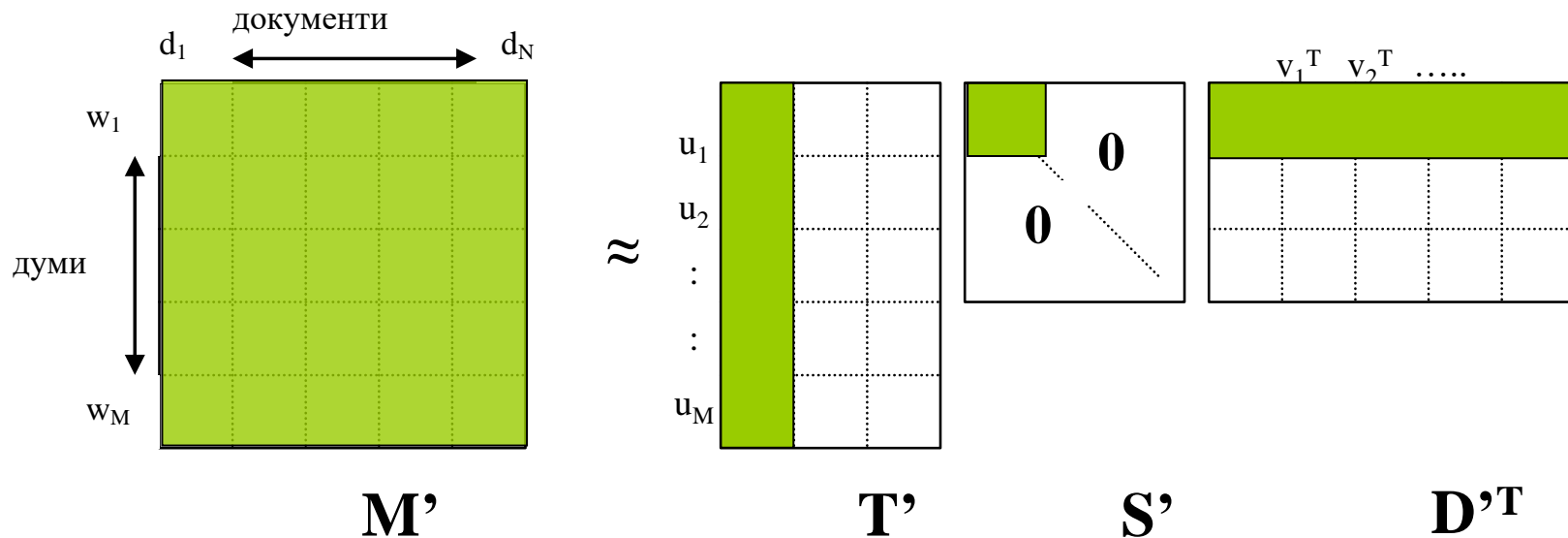
Декомпозиция по сингулярни стойности



T, D: ортонормални

S: диагонална

Отрязване на матриците



T', D' : ортогонални

S' : диагонална

Сравняване на два документа

$$\begin{aligned} M^T M &= (TSD^T)^T (TSD^T) \\ &= DST^T TSD^T \\ &= DSSD^T \\ &= (DS)(DS)^T \end{aligned}$$

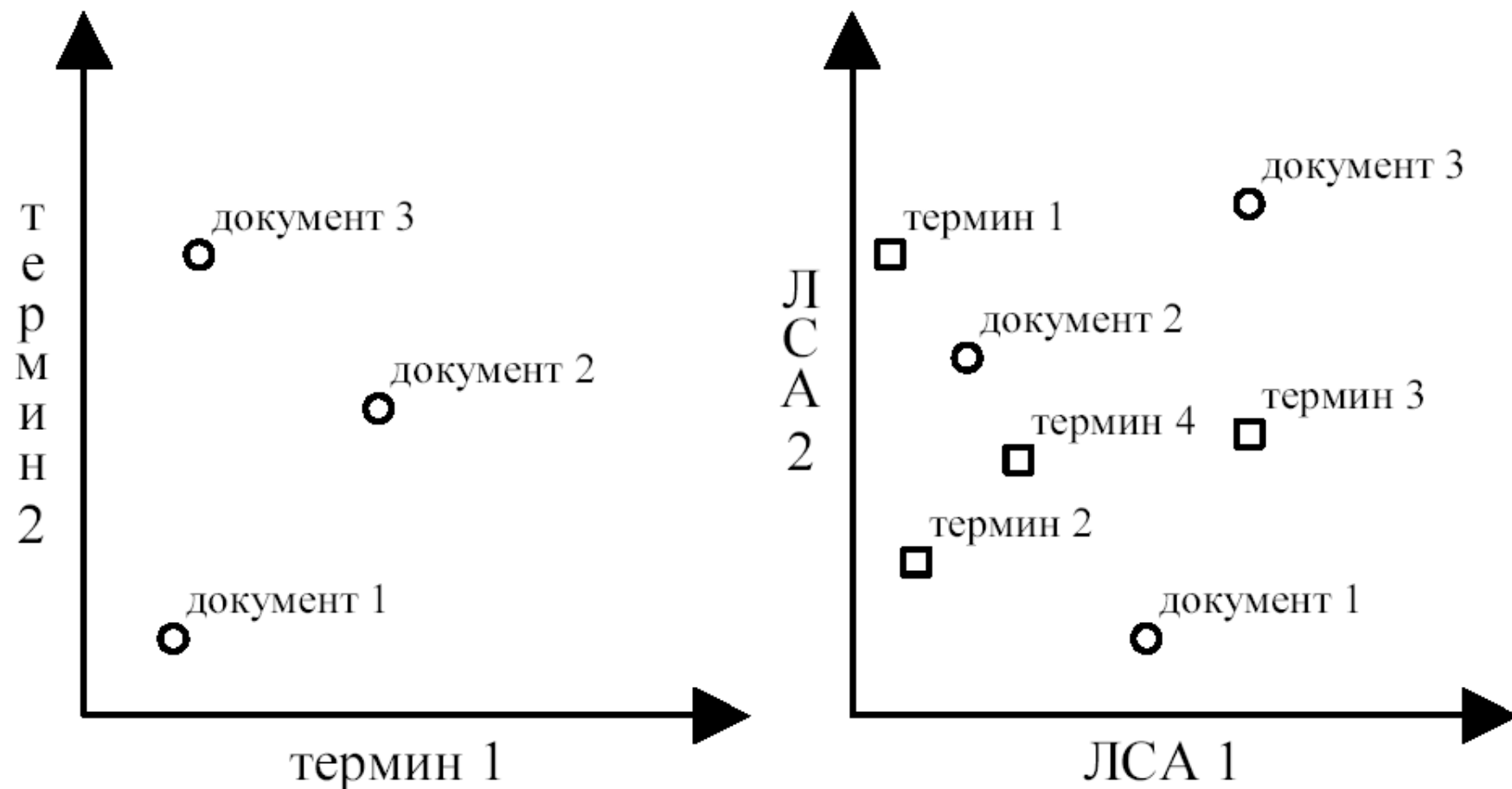
т.е. **DS** е вектор на **документ**

Сравняване на два термина

$$\begin{aligned}MM^T &= (TSD^T)(TSD^T)^T \\&= TSD^T DST^T \\&= TSST^T \\&= (TS)(TS)^T\end{aligned}$$

т.е. **TS** е вектор на **термин**

Векторен модел и ЛСА



Резюме

- Класификация на текст
 - примери
 - класификация и клъстеризация
- По-важни класификатори
- Атрибути
 - типове
 - селекция
 - претегляне
 - трансформация: ЛСА

За четене

□ **Classification, kernels**

- Dan Klein's tutorial

- <http://www.cs.berkeley.edu/~klein/papers/classification-tutorial-naacl2007.pdf>

□ **Text Categorization (TC)**

- Sebastiani survey on machine learning for TC

- <http://dl.acm.org/citation.cfm?id=505283>

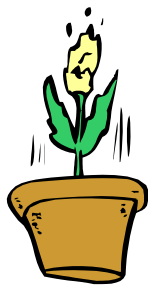
□ **Feature selection**

- Yang & Pedersen

- <http://dl.acm.org/citation.cfm?id=657137>

□ **Latent Semantic Analysis**

- Preslav's file



Въпроси и коментари?