

# Интеллектуальные информационные системы

Тема 7

## Тематическая категоризация

Павел Исаакович Браславский  
[pb@imach.uran.ru](mailto:pb@imach.uran.ru)  
весенний семестр 2006

### План

- Зачем? Области применения
- Постановка задачи
- Представление документов (индексирование)
  - Снижение размерности пространства признаков
    - выбор признаков
    - репараметризация (LSI)
  - Мера важности терминов
- Методы классификации
  - Наивный Байесовский подход
  - kNN
  - Метод Роккио (центроиды классов)
  - SVM

# Тематическая категоризация

- Тематическая классификация, text classification, topic spotting...
- Распределить документы по предварительно определенным категориям

2006

П.И. Браславский - Интеллектуальные ИС

3

# Почтовый спам

## А Р Е Н Д А квартир в Москве: Любые квартиры в любом районе

Позвоните нам - (495) IOI-14-17  
и мы Вам поможем

### к Вашим услугам:

- наш 10-летний опыт работы
- квартиры любых типов и категорий
- помесная оплата, без залога и предоплаты
- бесплатный подбор и просмотр всех вариантов
- гарантии от выселения и поднятия цены на весь срок сдачи
- база из 120 000 постоянно создаваемых вариантов во всех районах

### Наши «горячие» предложения за сегодня:

- 1к.кв. в Новопеределкино за 430\$, м.Юго-Западная, 19/22эт. в наличии телефон, наборная мебель, холодильник, ТВ
- 1к.кв. на Изумрудной ул за 450\$, м.Бабушкинская, 4/5эт. в наличии телефон, мет'дв, набор мебели, холодильник, ТВ
- 1к.кв. на ул. Речников за 500\$, м.Коптевская, 5/5эт. в наличии телефон, мет'дв, набор мебели, холодильник, ТВ
- 1к.кв. на ул.Гришина за 600\$, м.Кунцевская, 3/5эт. в наличии телефон, косм, ремонт, гарнитурная мебель, холодильник, ТВ
- 1к.кв. на Беловорской ул за 500\$, м.Речной вокзал, 7/9эт. в наличии телефон, мет'дв, набор мебели, холодильник, ТВ
- 1к.кв. на Проспекте Буденного за 600\$, м.Семеновская, 6/9эт. в наличии телефон, мет'дв, косм, ремонт, гарнитуру мебели, вся техника
- 1к.кв. на Протопоповском пер за 1000\$, м.Проспект Мира, 8/14эт. в наличии ремонт под евро, сейфовая дверь, мягкая мебель, горка, встроенная кухня, вся бытовая техника
- 2к.кв. на Винникой ул за 600\$, м.Университет, 4/9эт. в наличии телефон, вся необходимая мебель, холодильник, ТВ
- 2к.кв. на Живописной ул за 620\$, м.Полтежская, 2/5эт. в наличии телефон, мет'дв, гарнитурная мебель, холодильник, ТВ
- 2к.кв. на Гравеиновской ул за 640\$, м.Техстальники, 6/9эт. в наличии только кухонный гарнитур, холодильник, ТВ
- 2к.кв. на ул.Василисы Кожиной за 700\$, м.Братинковская, 3/12эт. в наличии паркетный пол, после ремонта, вся необходимая мебель, холодильник, ТВ
- 2к.кв. на Ельнинской ул за 700\$, м.Молодцова, 5/12эт. в наличии диван, телефон, гарнитурная мебель, вся техника
- 2к.кв. на ул.Молосовых за 700\$, м.Новосигарево, 3/9эт. в наличии телефон, набор мебели в комнатах, кухонный гарнитур, стир/машина
- 2к.кв. на Ореховом б-ре за 700\$, м.Красногвардейская, 2/9эт. в наличии телефон, мет'дв, гарнитурная мебель, техника
- 2к.кв. на Рублевском ш. за 750\$, м.Крылатское, 10/17эт. в наличии только телефон, мет'дв
- 2к.кв. на Кропоткинском пер за 1000\$, м.Кропотинская, 2/5эт. в наличии только телефон.
- 2к.кв. на ул.Удальцова за 3000\$, м.Проспект Вернадского, 22/25эт. евродом, охраня территория, подземный паркинг, евроремонт, столовая - гостиная / 35, спальня - 20, холл - 18, теплые полы на балконе и на кухне, джакузи, встроенная кухня, вся бытовая техника
- 3к.кв. на Свободном пр-те за 750\$, м.Новосигарево, 8/9эт. в наличии только телефон, мет'дв

Все готово к въезду, заселение в течение 2-х часов

Телефон IOI-14-17 Почта [ofis1602@hotpop.com](mailto:ofis1602@hotpop.com)

2006

П.И. Браславский - Интеллектуальные ИС

4

# Каталоги

Arts in the Yahoo! Directory - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное

Адрес: http://dir.yahoo.com/Arts/

Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]

YAHOO! SEARCH Directory Search: the Web | the Directory | this category Search

Arts [Email this page](#)

Directory > Arts

CATEGORIES [What's This?](#)

Top Categories

- [By Region](#) (56513) **NEW!**

Additional Categories

- [Art History](#) (1772)
- [Art Weblogs@](#)
- [Artists](#) (2027) **NEW!**
- [Arts Therapy@](#)
- [Awards](#) (17)
- [Booksellers@](#)
- [Censorship](#) (10)
- [Chats and Forums](#) (21)
- [Crafts](#) (955) **NEW!**
- [Education](#) (628) **NEW!**
- [Events](#) (297)
- [Humanities](#) (56138) **NEW!**
- [Institutes](#) (33)
- [Job and Employment Resources](#) (34)
- [Museums, Galleries, and Centers](#) (999)
- [News and Media](#) (234)
- [Organizations](#) (306)
- [Performing Arts](#) (8035) **NEW!**

2006

5

# Новостные рубрики

**Yandex**  
Найдётся всё

[Везде](#) [Каталог](#) [Новости](#) [Маркет](#) [Адреса](#) [Словари](#) [Картинки](#) [Все службы...](#)

автоматически обработан 1151 источник, обновлено 21.03.2006 11:17 мск

## Главные новости

выпуск: Россия | [Украина](#)

### [Сегодня наступает астрономическая весна \(60\)](#)

21 марта день сравняется с ночью. В этот день солнце переходит из Южного полушария в Северное. С астрономической точки зрения, 21 марта наступает весна и продолжается до 22 июня - до дня летнего ...

### [Митинг оппозиции на Октябрьской площади Минска продолжался всю ночь \(102\)](#)

Принимающие участие в несанкционированной акции протеста против победы Александра Лукашенко на президентских выборах в Белоруссии не покидали Октябрьской площади Минска ...

В понедельник вечером в митинге в столице Белоруссии принимали участие несколько тысяч человек.

### [Россия ждет рост цен на телефон \(46\)](#)

Согласно информации этой службы, с 1 апреля 2006 года вводятся новые тарифы на услуги телефонной связи, предоставляемые компаниями, которые входят в холдинг ОАО ...

В МГТС рассказали, что будет введено три тарифных плана - с абонентской платой, с поперменной системой оплаты и с комбинированной системой оплаты.

### [Белорусы стартовали на чемпионате мира по фигурному катанию \(44\)](#)

09.15 - мужчины, квалификация, 14.35 - пары, короткая программа, 18.54 - мужчины,

## Главные новости

- [Политика](#)
- [В мире](#)
- [Общество](#)
- [Экономика](#)
- [Спорт](#)
- [Присшествия](#)
- [Культура](#)
- [Наука](#)
- [Здоровье](#)
- [Hi-Tech](#)
- [Интернет](#)
- [Авто](#)
- [Туризм](#)

2006

П.И. Браславский - Интеллектуальные ИС

6

## Ручные классификаторы

- Категория → набор ключевых слов
- + тщательная работа позволяет добиться хороших результатов
- много ручного труда, concept drift: необходимость поддержки

## Машинное обучение

- Обучающее множество
- Автоматический механизм построения классификатора по обучающему множеству
- + меньше ручного труда (или ниже требования к квалификации)
  - возможен overfitting

## Классификация с учителем

Классы:  $C = \{c_1, \dots, c_n\}$

Объекты:  $x \in X$

Найти классификационную функцию

$f: X \rightarrow C$

или

$\forall c \in C f_c: X \rightarrow \{0, 1\}$  или  $f_c: X \rightarrow [0, 1]$

Однозначная/многозначная классификация.

Особый случай: бинарная классификация

Обучающая выборка

$T = \{(x^*, c^*) | x^* \in X, c^* \in C\}$

## Постановка задачи – 2

- В общем случае категории описываются *только* набором позитивных примеров (а не набором ключевых слов, например)
- Тексты предполагаются неструктурированными, не снабженными метаданными (дата, автор и т.д.)

## Представление документов

- Похоже на индексирование в задачах ИП
- термины (terms) = признаки (features) (терминология распознавания образов)
- Отличие от ИП: снижение размерности пространства признаков (из-за вычислительной сложности; overfitting)

## Построение классификатора

- Обучающее множество (*training set*): построение классификатора
- Множество для оценки (*validation set*): настройка параметров
- тестовое множество (*test set*): тестирование, оценка качества классификации



## Снижение размерности

- Выбор признаков (*feature selection*)
- Извлечение признаков (*feature extraction, reparameterization*)
- Лингвистические методы (например, выделение устойчивых словосочетаний)
- Локальное/глобальное снижение размерности



## Способы отбора признаков

- Document frequency (или порог встречаемости): удаление очень редких терминов
- Information gain (expected mutual information)

# Меры для отбора признаков

Function	Denoted by	Mathematical form
<i>Document frequency</i>	$\#(t_k, c_i)$	$P(t_k, c_i)$
<i>Information gain</i>	$IG(t_k, c_i)$	$P(t_k, c_i) \cdot \log \frac{P(t_k, c_i)}{P(c_i) \cdot P(t_k)} + P(\bar{t}_k, c_i) \cdot \log \frac{P(\bar{t}_k, c_i)}{P(c_i) \cdot P(\bar{t}_k)}$
<i>Chi-square</i>	$\chi^2(t_k, c_i)$	$\frac{g \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]^2}{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}$
<i>Correlation coefficient</i>	$CC(t_k, c_i)$	$\frac{\sqrt{g} \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}}$
<i>Relevancy score</i>	$RS(t_k, c_i)$	$\log \frac{P(t_k c_i) + d}{P(\bar{t}_k \bar{c}_i) + d}$

Sebastiani, 1999

# Извлечение признаков

- Кластеризация терминов (*term clustering*) на основе совместной встречаемости (см. предыдущую тему)
- Латентно-семантическое индексирование (*latent semantic indexing, LSI*)



# Latent Semantic Indexing

- Модель ИП, направленная на устранение проблемы синонимии и полисемии слов
- Основана на сингулярном разложении матрицы термин-документ
- В задаче классификации сингулярное разложение строится на основе обучающего множества
- «Идея»: позволяет учитывать не только «сильные термины», но и группы «слабых».

## LSI

- $A_{mn}$  – матрица термин-документ;  
 $a_{ij}$  – вес термина  $i$  в документе  $j$
- $U_{mm}$  состоит из ортонормальных собственных векторов матрицы  $AA^T$
- $V_{nn}$  состоит из ортонормальных собственных векторов матрицы  $A^T A$
- $S_{nn}$  – диагональная матрица, состоящая из неотрицательных квадратных корней собственных чисел матрицы  $A^T A$
- Сингулярное разложение  $A = USV^T$
- Оптимальная малоранговая аппроксимация  $A$ :  
 $A_k = U_k S_k V_k^T$
- $d' = U_k^T S_k^{-1} d$

см. Добрынин, 2002

# Методы классификации

- Параметрические
  - Наивный Байесовский классификатор
  - Support Vector Machine, SVM
- Непараметрические
  - Профиль класса («Идеальный представитель»)
  - Группа примеров

# Байесовский подход

Использует теорему Байеса: вычисляются апостериорные вероятности

$$p(c | d) = \frac{p(c)p(d | c)}{p(d)}$$

$p(c|d)$  – вероятность того, что документ  $d$  принадлежит классу  $c$ ;

$p(c)$  – вероятность того, что наугад взятый документ принадлежит классу  $c$ ;

$p(d|c)$  – вероятность «встретить» документ  $d$  в классе  $c$ ;

$p(d)$  – вероятность «встретить» документ  $d$  (не зависит от класса)

# Байесовский подход

Документ – это набор признаков (соответствуют словам):

$$p(d|c) \rightarrow p(x_1, \dots, x_n|c)$$

Наивность подхода – в предположении независимости признаков:

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i, c)$$

Оценка вероятностей:

$$p^*(x_i|c) = \# \text{ объектов со свойством } i / \text{ всего объектов}$$

# НБ: Обучение

- Составить словарь (*Vocabulary*) по обучающей выборке (ОВ)
- Посчитать  $P(c_j)$  и  $P(x_k | c_j)$ 
  - Для каждого класса  $c_j$ 
    - $docs_j \leftarrow$  подмножество ОВ, соответствующее  $c_j$
    - $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{размер ОВ}|}$$
  - $Text_j \leftarrow$  объединение всех  $docs_j$  в один документ
  - Для каждого слова  $x_k$  из *Vocabulary*
    - $n_k \leftarrow$  количество слов  $x_k$  в  $Text_j$
    - $$P(x_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$
 ← сглаживание

## НБ: классификация

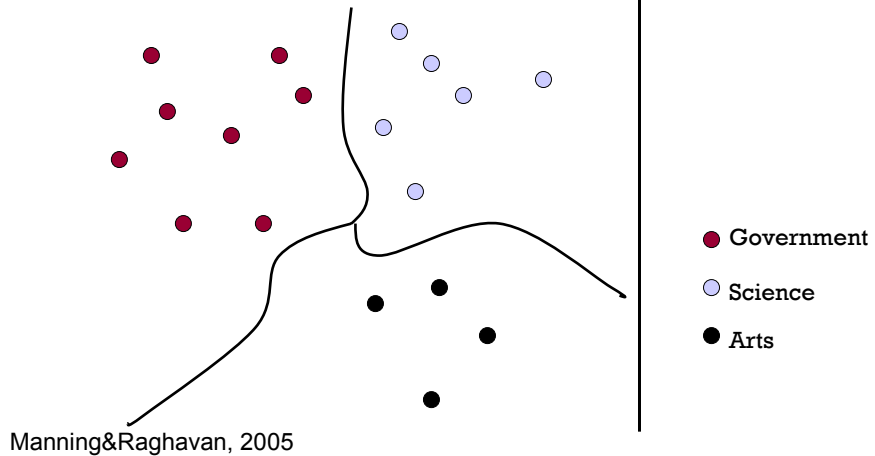
- *positions* ← все позиции слов в документе, подлежащем классификации, которые содержат слова из словаря (*Vocabulary*)
- Документ относится к классу  $c_{NB}$ , такому что:

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

## НБ: преимущества

- Простота
- Вычислительная эффективность:  
линейное время обучения и  
классификации

# Векторное пространство

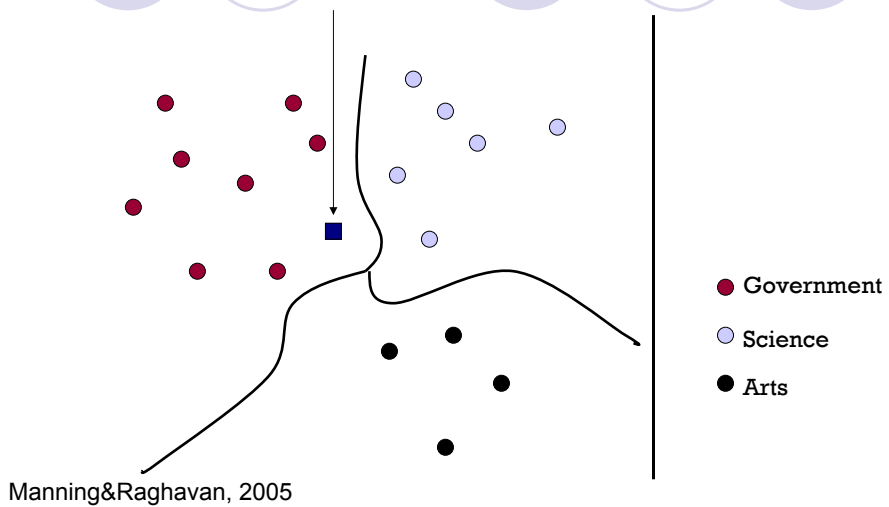


2006

П.И. Браславский - Интеллектуальные ИС

25

# Тестирование



2006

П.И. Браславский - Интеллектуальные ИС

26

## k-Nearest Neighbors (k-NN)

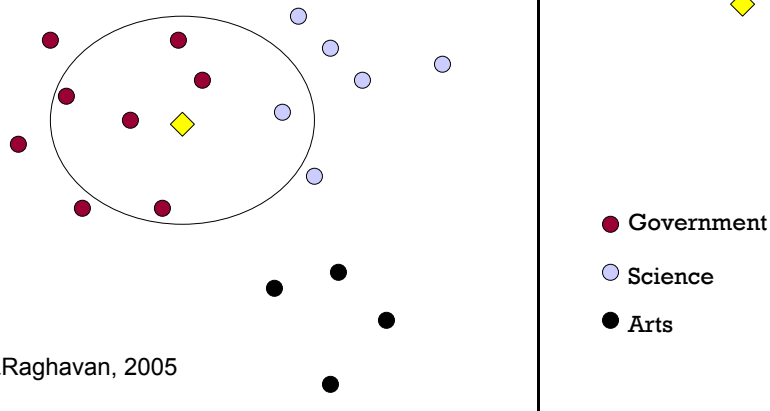
- Найти  $k$  ближайших соседей из тестовой выборки
- Посмотреть, как они распределились по классам
- Выбрать класс с максимальной долей в  $k$
- Для классификации текстов в качестве метрики обычно используется  $\cos$
- Представление документов –  $tf \cdot idf$

2006

П.И. Браславский - Интеллектуальные ИС

27

## Пример: $k=6$ (6NN)



Manning&Raghavan, 2005

2006

П.И. Браславский - Интеллектуальные ИС

28

## Особенности kNN

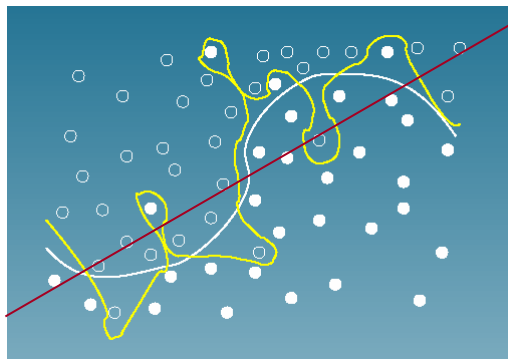
- Нет собственно фазы обучения
- Сильно чувствителен к метрике
- Чувствителен к  $k$
- Вычислительно сложнее, чем NB или линейные классификаторы
- Достаточно высокое качество классификации
- Гибкость

2006

П.И. Браславский - Интеллектуальные ИС

29

## «Гибкость/прямолинейность»



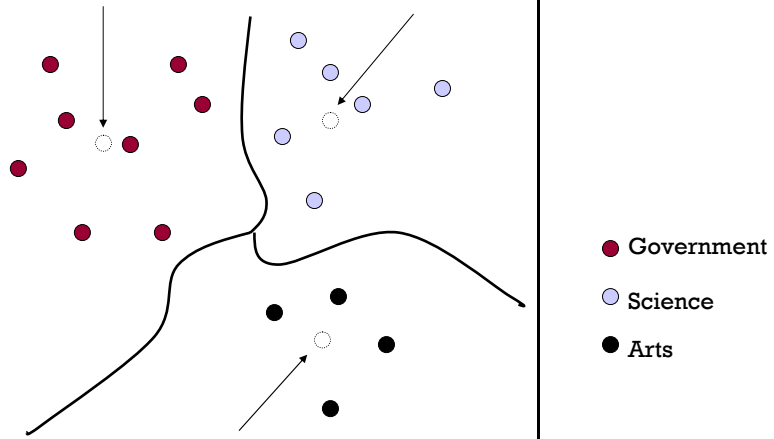
Manning&Raghavan, 2005

2006

П.И. Браславский - Интеллектуальные ИС

30

# Центриды (или метод Роккио)



2006

П.И. Браславский - Интеллектуальные ИС

31

# Метод Роккио

$$w_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|}$$

Sebastiani, 2002

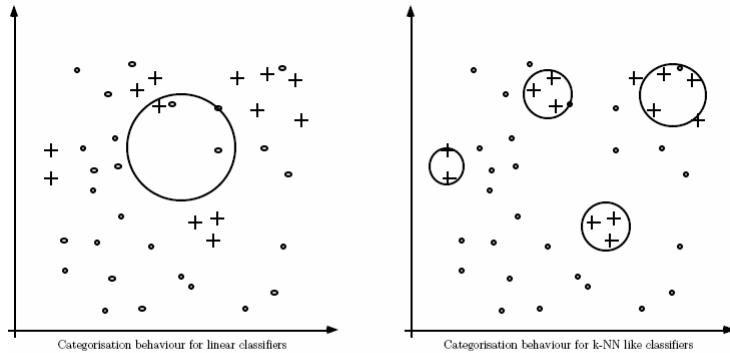
2006

П.И. Браславский - Интеллектуальные ИС

32



## Центриды vs kNN



Sebastiani, 1999

o Negative training examples  
+ Positive training examples  
○ Influence area of the classifier

2006

П.И. Браславский - Интеллектуальные ИС

33

## Support Vector Machines (SVM)

- Наиболее эффективный современный метод
- Математика: Владимир Вапник, применение к текстам: Торстен Иоахимс
- Идея: разделить позитивные и негативные примеры максимально широкой границей

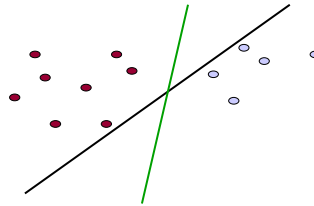
2006

П.И. Браславский - Интеллектуальные ИС

34

# Вводные замечания

- Бинарная классификация
- Разделяющая поверхность – гиперплоскость (в т.ч. NB, Rocchio)
- Для двумерного случая:  $ax+bx=c$
- Общем случае – бесконечно много решений



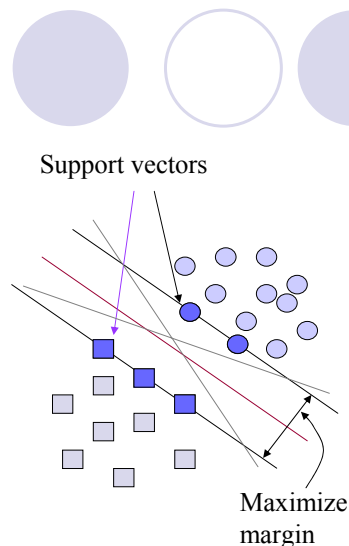
2006

П.И. Браславский - Интеллектуальные ИС

35

# SVM – идея

- Максимизация границы около разделяющей гиперплоскости
- Разделяющая гиперплоскость определяется опорными векторами («проблемные примеры»)
- Задача квадратичного программирования



Manning&Raghavan, 2005

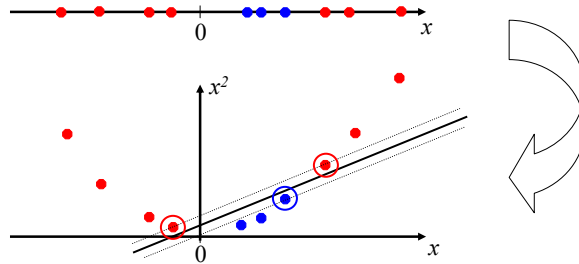
2006

П.И. Браславский - Интеллектуальные ИС

36

# Нелинейные SVM

Отображение в пространство большей размерности, где обучающее множество линейно разделяемо



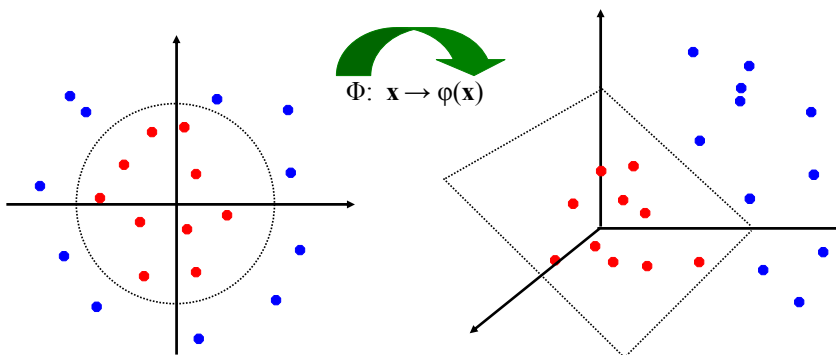
Manning&Raghavan, 2005

2006

П.И. Браславский - Интеллектуальные ИС

37

# Нелинейные SVM - 2



Manning&Raghavan, 2005

2006

П.И. Браславский - Интеллектуальные ИС

38