Интеллектуальные информационные системы

Тема 7

Тематическая категоризация

Павел Исаакович Браславский
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

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
- помесячная 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 сдаваемых 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$, м.Коломенская, 5\5эт, в наличии телефон, мет*дв, набор мебели, холодильник, ТВ
1к.кв, на ул.Гришина за 500$, м.Кунцевская, 3\5эт, в наличии телефон, косм. ремонт, гарнитурная мебель, холодильник, ТВ
1к.кв, на Беломорской 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\9! эт, в наличии телефон, мет*дв, косм. ремонт, гарнитур 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, строенная кухня, вся бытовая техника

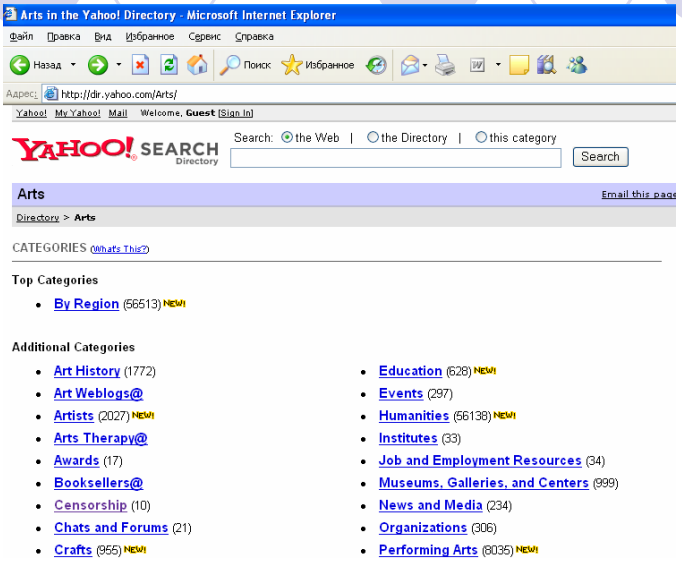
2к.кв, на Винницкой ул за 600$, м.Университет, 4\9эт, в наличии телефон, вся необходимая мебель, холодильник, ТВ
2к.кв, на Живописной ул за 620$, м.Полежаевская, 2\5эт, в наличии телефон, мет*дв, гарнитурная мебель, холодильник, ТВ
2к.кв, на Грайвороновской ул за 640$, м.Текстильщики, 6\9эт, в наличии только кухонный гарнитур, холодильник, ТВ
2к.кв, на ул.Василисы Кожиной за 700$, м.Багратионовская, 3\12эт, в наличии паркетный пол, после ремонта, вся необходимая мебель, холодильник, ТВ
2к.кв, на Ельнинской ул за 700$, м.Молодежная, 5\12эт, в наличии домофон, телефон, гарнитурная мебель, вся техника
2к.кв, на ул.Молостовых за 700$, м.Новогиреево, 3\9эт,1 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 , 16\17эт, в наличии только телефон, мет*дв
2к.кв, на Кропоткинском пер за 1000$, м.Кропоткинская, 2\5эт, в наличии только телефон,
2к.кв, на ул.Удальцова за 3500$, м.Проспект Вернадского, 22\25эт, вародом, охрана территории, 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.Новогиреево, 8\9эт, в наличии только телефон, мет*дв

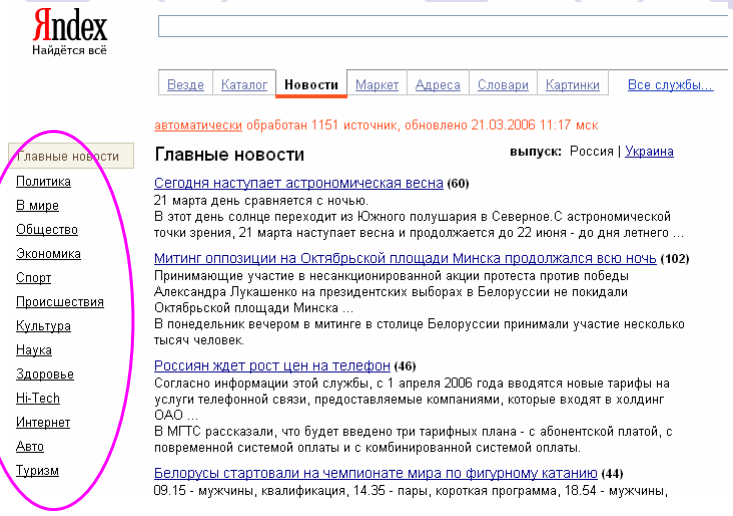
Все готово к въезду, заселение в течение 2-х часов

Телефон IOI–14–17 **Почта** ofis1602@hotpop.com

Каталоги



Новостные рубрики



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: $C=\{c_1, \ldots, c_n\}$
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
Document frequency	$\#(t_k, c_i)$	$P(t_k, c_i)$
Information gain	$IG(t_k, c_i)$	$P(t_k, c_i) \cdot log\frac{P(t_k, c_i)}{P(c_i) \cdot P(t_k)} + P(\overline{t}_k, c_i) \cdot log\frac{P(\overline{t}_k, c_i)}{P(c_i) \cdot P(\overline{t}_k)}$
Chi-square	$\chi^2(t_k, c_i)$	$\frac{g \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]^2}{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}$
Correlation coefficient	$CC(t_k, c_i)$	$\frac{\sqrt{g} \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}}$
Relevancy score	$RS(t_k, c_i)$	$log\frac{P(t_k \mid c_i) + d}{P(\overline{t}_k \mid \overline{c}_i) + d}$

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^TA
- S_{nn} – диагональная матрица, состоящая из неотрицательных квадратных корней собственных чисел матрицы A^TA
- Сингулярное разложение $A = USV^T$
- Оптимальная малоранговая аппроксимация A:
  $A_k = U_kS_kV_k^T$
- $d' = U_k^TS_k^{-1}d$

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

$$p(c \mid d) = \frac{p(c)\,p(d \mid c)}{p(d)}$$

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

$p(c)$ – вероятность того, что наугад взятый документ принадлежит класса 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):

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

Наивность подхода – в предположении *независимости* признаков:

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

Оценка вероятностей:

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

НБ: Обучение

- Составить словарь (*Vocabulary*) по обучающей выборке (ОВ)
- Посчитать $P(c_j)$ и $P(x_k \mid c_j)$
  - 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 $Text_j$
    - $$P(x_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$ ← сглаживание

НБ: классификация

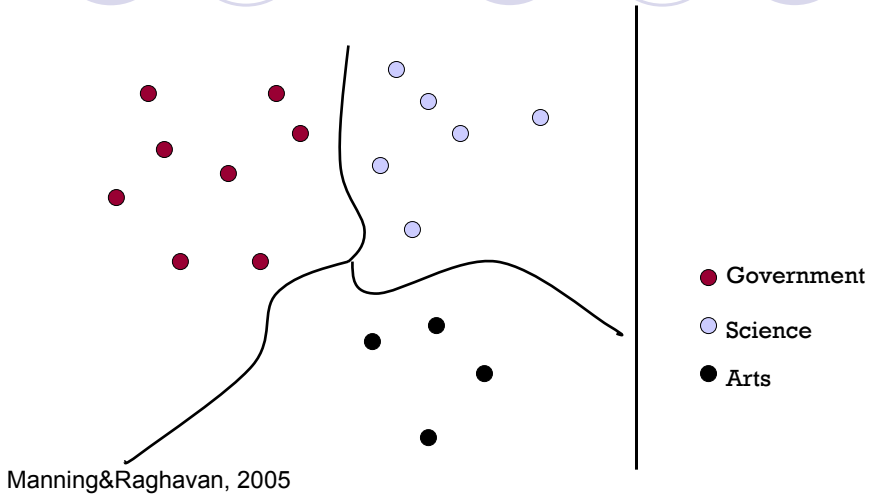
- *positions* ← все позиции слов в документе, подлежащем классификации, которые содержат слова из словаря (*Vocabulary)*
- Документ относится к классу c_{NB}, такому что:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

НБ: преимущества

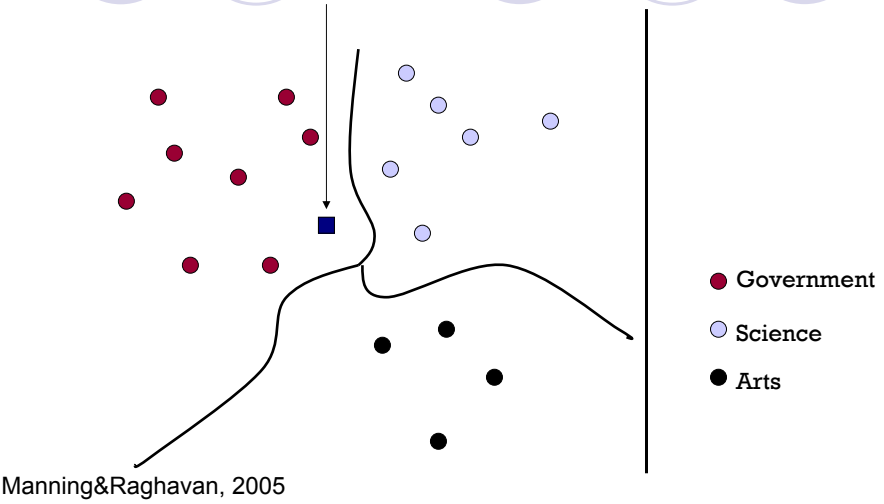
- Простота
- Вычислительная эффективность: линейное время обучения и классификации

Векторное пространство



Manning&Raghavan, 2005

Тестирование

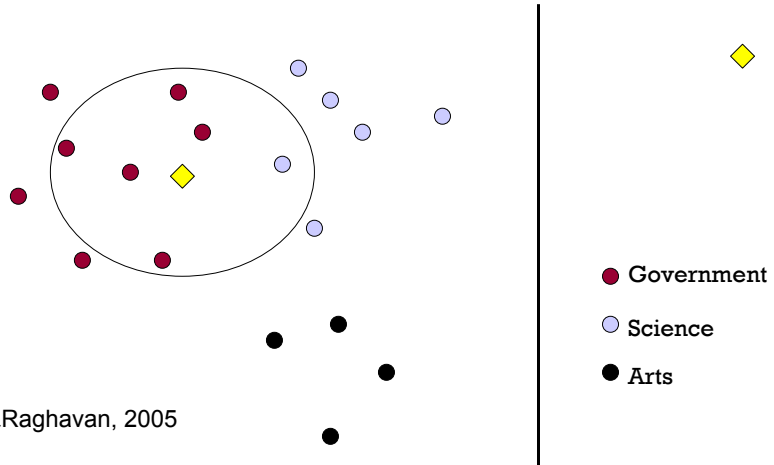


Manning&Raghavan, 2005

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*idf

Пример: k=6 (6NN)

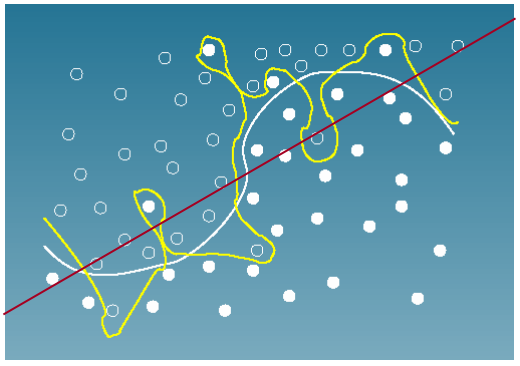


Manning&Raghavan, 2005

Особенности kNN

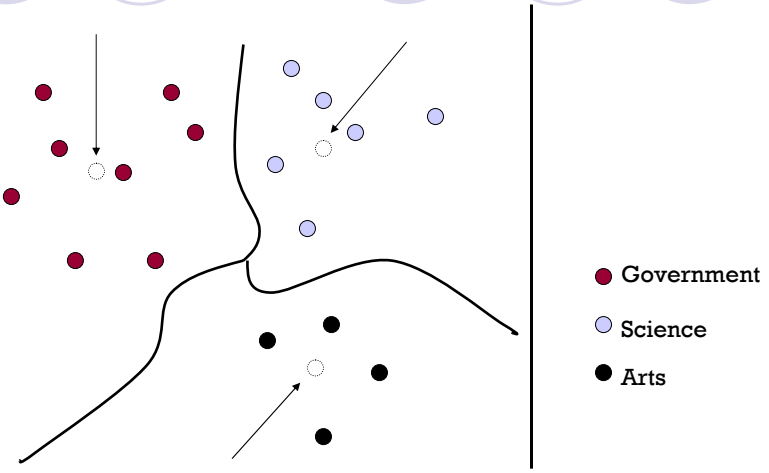
- Нет собственно фазы обучения
- Сильно чувствителен к метрике
- Чувствителен к k
- Вычислительно сложнее, чем NB или линейные классификаторы
- Достаточно высокое качество классификации
- Гибкость

«Гибкость/прямолинейность»



Manning&Raghavan, 2005

Центроиды (или метод Роккио)

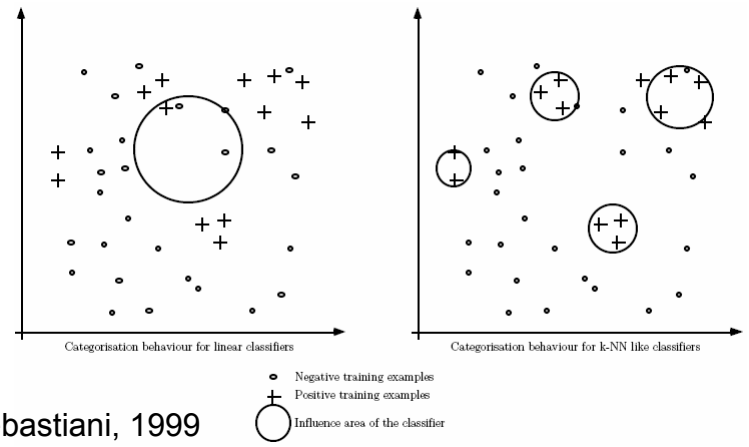


Метод Роккио

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

Sebastiani, 2002

Центроиды vs kNN



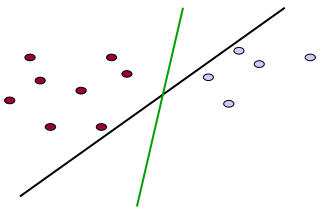
Sebastiani, 1999

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

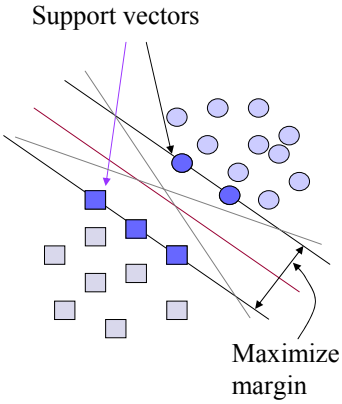
Вводные замечания

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



SVM – идея

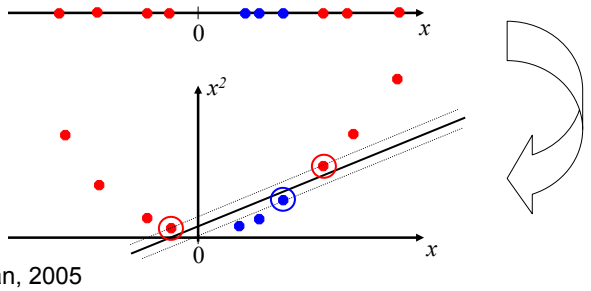
- Максимизация границы около разделяющей гиперплоскости
- Разделяющая гиперплоскость определяется опорными векторами («проблемные примеры»)
- Задача квадратичного программирования



Manning&Raghavan, 2005

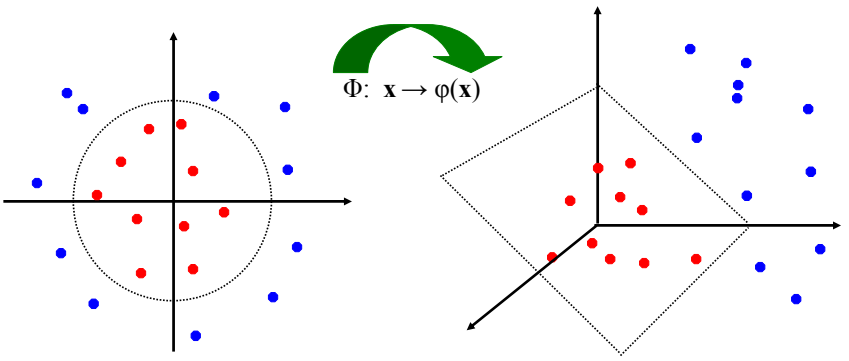
Нелинейные SVM

Отображение в пространство большей размерности, где обучающее множество линейно разделяемо



Manning&Raghavan, 2005

Нелинейные SVM - 2



Manning&Raghavan, 2005