

## Автоматическая классификация текстов

Лекция N 6 курса  
"Алгоритмы для Интернета"

Юрий Лифшиц

поми РАН - СПбГУ ИТМО

Осень 2006

1 / 30

"... классификация осуществляется на добровольной основе"

Владимир Стржалковский // из сообщений REGNUM

Библия классификатора: Fabrizio Sebastiani  
"Machine Learning in Automated Text Categorization"



2 / 30

### План лекции

- 1 Постановка задачи, подходы и применения
  - Постановка задачи
  - Основные шаги
- 2 Индексация документов
- 3 Построение и обучение классификатора
- 4 Оценка качества классификации

3 / 30

### Часть I

Как строго поставить задачу классификации текстов?

Области применения классификации текстов?

Три основных этапа классификации

4 / 30

### Акценты лекции

**Автоматическая** классификация

**Не:** подбор правил вручную

Автоматическая **классификация**

**Не:** автоматическая кластеризация

**Используем методы:**

Информационного поиска (Information Retrieval)

Машинного обучения (Machine Learning)

5 / 30

### Постановка задачи

**Данные задачи**

Категории  $\mathcal{C} = \{c_1, \dots, c_{|\mathcal{C}|}\}$

Документы  $\mathcal{D} = \{d_1, \dots, d_{|\mathcal{D}|}\}$

**Неизвестная** целевая функция  $\Phi : \mathcal{C} \times \mathcal{D} \rightarrow \{0, 1\}$

**Классификатор**

Наша задача построить классификатор  $\Phi'$   
максимально близкий к  $\Phi$

**Что мы знаем?**

Значение  $\Phi$  на начальной коллекции документов  
Коллекцию разделяют на "учебную", "проверочную"  
и "тестовую"

6 / 30

### Виды классификации

**Вид ответа:**

Точная классификация  $\Phi' : \mathcal{C} \times \mathcal{D} \rightarrow \{0, 1\}$

Ранжирование:  $\Phi' : \mathcal{C} \times \mathcal{D} \rightarrow [0, 1]$

**Порядок обработки данных**

Построение списка категорий для данного документа

Построение списка документов для данной категории

**Соотношение категорий**

Категории не пересекаются

Категории могут пересекаться

Бинарная классификация: две непересекающиеся  
категории

7 / 30

### Применения классификации текстов

**Где используются методы классификации текстов:**

- Фильтрация документов, распознавание спама
- Автоматическая аннотирование
- Снятие неоднозначности (автоматические переводчики)
- Составление интернет-каталогов
- Классификация новостей
- Распределение рекламы
- Персональные новости

8 / 30

**Индексация документов**

Переводим документы в единый экономный формат

**Обучение классификатора**

Общая форма классифицирующего правила  
Настройка параметров

**Оценка качества классификации**

Оценка абсолютного качества  
Сравнение классификаторов между собой

9 / 30

**Часть II**

В каком виде хранить документ?

Как уменьшить количество характеристик?

10 / 30

**Базовый подход****Исходное представление документа:**

Документ = коллекция слов (термов)  
Каждый терм имеет **вес** по отношению к документу

**Вес термина**

Стандартный подход:  $w_{ij} = TF_{ij} \cdot IDF_i$   
Проводится **нормализация** по документу

**Новые подходы:**

По-другому выбирать термы  
По-разному определять вес термина в документе  
Индексировать "фразы"  
Использовать дополнительные термы (не связанные со словами)

11 / 30

**Уменьшение размерности****Виды уменьшения размерности:**

Единый метод / свой для каждой категории  
Создание искусственных термов / выбор термов

**Выбор термов**

Оставлять "средне-встречающиеся" термы  
Использование различных "коэффициентов полезности"

**Искусственные термы**

Кластеризация термов  
Сингулярное разложение

12 / 30

**Часть III**

В какой форме строить классифицирующее правило?

Как подобрать параметры классификатора?

13 / 30

**Ранжирование и четкая классификация****Два этапа:**

Строим функцию  $CSV_i : \mathcal{D} \rightarrow [0, 1]$   
Выбираем пороговое значение  $\tau_i$

**Переход к точной классификации**

Пропорциональный метод  
Каждому документу выбрать  $k$  ближайших категорий

14 / 30

**Линейный on-line классификатор (1/3)**

**Документ:**  $d = (d_1, \dots, d_n)$

**Правило классификации:** скалярное произведение

$$CSV_i(d) = \bar{d} \cdot \bar{c}_i = \prod c_{ji} d_j$$

После нормализации получается косинус между векторами:

$$CSV_i(d) = \frac{\bar{d} \cdot \bar{c}_i}{|\bar{d}| |\bar{c}_i|}$$

Как подобрать  $c_{1i}, \dots, c_{ni}$ ?

15 / 30

**Линейный on-line классификатор (2/3)****On-line обучение**

Начинаем с  $\bar{c}_i = (1, \dots, 1)$   
Для каждого учебного документа применяем текущее правило  
При неудаче вносим поправки  $+\alpha, -\beta$  в координаты, соответствующие словам "проваленного" документа

**Вариации:**

Мультипликативные поправки  
Поправки при удачной классификации  
Поправки в "не активные" слова

16 / 30

**Преимущества**

Если будет обратная связь, обучение можно продолжать и за пределами учебной коллекции  
Можно уменьшать пространство термов on-line

Как применять линейный классификатор для случаев документо-центрированной классификации и категория-центрированной классификации?

**Учебная коллекция в матричном виде:**

Каждый документ — это вектор из весов термов  
Все вместе документы образуют матрицу  $I$  размера  $|Tr| \times |T|$   
Степень принадлежности документа категориям — вектор  
Для всех документов вместе — матрица  $O$  размера  $|C| \times |Tr|$

**Цель:**

Найти матрицу линейных правил  $M$ , минимизирующую

$$\|MI - O\|$$

**Цель:**

Найти матрицу линейных правил  $M$ , минимизирующую  $\|MI - O\|_F$

**Матричная норма Фробениуса:**

Корень из суммы квадратов всех элементов

**Интерпретация:**

Хотим минимизировать корень из суммы квадратов всех ошибок

**Алгоритм минимизации:**

Отдельно для каждой категории  
Как найти  $\bar{c}_i$  минимизирующее  $l\bar{c}_i - \bar{d}_i$ ?  
Нужно взять проекцию  $\bar{d}_i$  на линейную оболочку строк  $l$

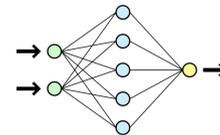
**Вид классификатора:** принадлежность категории определяется ДНФ-формулой:

Если (ворота&вратарь)  $\vee$   
(лук& $\neg$ жареный)  $\vee$   
(хоккей), то  $d \in$  "Спорт"

**Обучение:**

- 1. Начинаем с огромной формулы описывающей все документы категории (и отрицающей все внешние документы из учебной коллекции)
- 2. Проводим серию упрощений и слияний скобок
- 3. Проводим вторую серию упрощений, жертвуя всеобщей точностью на тренировочной коллекции

Иллюстрация из Wikipedia



**Вычисления с помощью нейронных сетей:**

- Входные, промежуточные, выходные элементы
- Коэффициенты на ребрах
- Пороги активации во внутренних вершинах
- Цель: подобрать коэффициенты для наилучшего вычисления желаемой функции

**Классификация с помощью нейронных сетей:**

- Входной уровень — веса термов в документе
- Ноль, один, несколько промежуточных уровней
- Уровень ответов состоит из клеток принадлежности категориям

Какого вида правило мы получим при отсутствии промежуточных уровней?

**Обучение нейронных сетей:**

- 1. Провести вычисления на учебном документе
- 2. Для каждой категории с существенной ошибкой внести поправки в коэффициенты на ребрах, которые ведут в нее
- 3. Пройти по этим ребрам назад
- 4. Провести корректировку для внутренних вершин и ребер, ведущих в них

## Часть IV

Как оценить качество классификатора?

25 / 30

## Метрики из информационного поиска

- **Полнота:** отношение количества найденных документов из категории к общему количеству документов категории
- **Точность:** доля документов действительно из категории в общем количестве найденных документов
- **Аккуратность:** доля верно соотнесенных документов во всех документах

Чем плоха аккуратность?

26 / 30

## Сравнение двух методов

### Явный метод (benchmarks):

- Одинаковая коллекция (например, новости Reuters)
- Одинаковая индексация
- Одинаковый обучающий набор

### Неявный метод

- Сравнивать каждый метод с неким "эталонным" примитивным методом

27 / 30

## Задача

Пусть мы узнали, что вероятности принадлежности документов к категории равны  $p_1 \geq \dots \geq p_n$ . По какому порогу надо принять решение о принадлежности, чтобы ожидание функции эффективности

$$u_{TP} \cdot \#TP + u_{TN} \cdot \#TN + u_{FP} \cdot \#FP + u_{FN} \cdot \#FN$$

было максимально (мы считаем, что  $u_{TP}, u_{TN} > u_{FP}, u_{FN}$ )?

28 / 30

## Главные моменты

### Сегодня мы узнали:

- Классификация текстов использует методы информационного поиска и машинного обучения
- Три этапа: индексация, построение классификатора, оценка качества
- Классификаторы: линейный, ДНФ-правило, метод регрессий, нейронные сети

Вопросы?

29 / 30

## Источники

Страница курса <http://logic.pdmi.ras.ru/~yura/internet.html>

Использованные материалы:

-  Fabrizio Sebastiani  
Machine Learning in Automated Text Categorization  
<http://nmls.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf>
-  Юрий Лифшиц  
Лекция по классификации текстов (конспект)  
<http://logic.pdmi.ras.ru/~yura/modern/06modernnote.pdf>

30 / 30