Классификация текстов

Лекция N 6 курса
“Современные задачи
теоретической информатики”

Юрий Лифшиц
yura@logic.pdmi.ras.ru

ИТМО

Осень’2005

План лекции

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

План лекции

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

Акценты лекции

Автоматическая классификация
  Не: подбор правил вручную

Автоматическая **классификация**
  Не: автоматическая кластеризация

Используем методы:
  Информационного поиска (Information Retrieval)
  Машинного обучения (Machine Learning)

Постановка задачи

Данные задачи

Категории $\mathfrak{C} = \{c_1, \ldots, c_{|\mathfrak{C}|}\}$

Документы $\mathfrak{D} = \{d_1, \ldots, d_{|\mathfrak{D}|}\}$

Неизвестная целевая функция $\Phi : \mathfrak{C} \times \mathfrak{D} \to \{0, 1\}$

Классификатор

Наша задача построить классификатор Φ' максимально близкий к Φ

Что мы знаем?

Значение Φ на начальной коллекции документов

Коллекцию разделяют на "учебные" и "тестовые"

Виды классификации

Вид ответа:

Точная классификация $\Phi' : \mathfrak{C} \times \mathfrak{D} \to \{0, 1\}$

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

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

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

Три этапа классификации

Индексация документов

Переводим все документы в единый экономный формат

Обучение классификатора

Общая форма классифицирующего правила

Настройка параметров

Оценка качества классификации

Оценка абсолютного качества

Сравнение классификаторов между собой

План лекции

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

Базовый подход

Исходное представление документа:
- Документ = коллекция слов (термов)
- Каждый терм имеет **вес** по отношению к документу

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

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

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"
- Выбор "зависимых" термов

Искусственные термы
- Кластеризация термов
- Сингулярное разложение (из прошлой лекции)

План лекции

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

Ранжирование и четкая классификация

Два этапа:

Строим функцию $CSV_i : \mathfrak{D} \to [0, 1]$

Выбираем пороговое значение τ_i

Переход к точной классификации

Пропорциональный метод

Каждому документу выбрать k ближайших категорий

Метод Rocchio

Построение профайла категории

Считаем среднее арифметическое векторов-документов

Определяем $CSV_i(d)$

как расстояние от вектора d до профайла

Разрешающие деревья

Строим дерево для обучающего набора

Выбираем терм

Документы, его содержащие, кладем направо, остальные налево

Повторяем, пока не получим однородные группы

Как выбирать разделяющий терм?

По корреляции с категорией

Трудность

Дерево не должно быть слишком глубоким = "проблема переобучения"

Метод k соседей

Идея:

Определять категорию документа через категории соседних учебных документов

Реализация:

$CSV_i(d) = \sum_{d_z \in Tr_k(d)} |d, d_z| \cdot \Phi(c_i, d_z)$

Рекомендации: k порядка 20-50

Другие методы

Не успеваем затронуть:

- Вероятностные классификаторы
- Нейронные сети
- Support Vector Machines

Комитеты классификаторов

Естественная идея:

Объединить несколько разных алгоритмов для принятия коллективного решения

Методы объединения:

- Выбор большинства
- Взвешенная линейная комбинация
- Динамический выбор классификатора
- Динамическая комбинация классификаторов

Последовательное обучение

- Классификаторы строятся по очереди
- Вводится понятие "трудности документа"
- Каждый следующий классификатор учитывает документы, на которых ошибся предыдущий, с большим весом

План лекции

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

Метрики из информационного поиска

Кто помнит метрики из прошлой лекции?

- **Полнота:** отношение количества найденных документов из категории к общему количеству документов категории
- **Точность:** доля документов действительно из катеогории в общем количестве найденных документов
- **Benchmarks:** показатели системы на контрольных запросах и специальных коллекциях документов (Reuters Collection)
- **Аккуратность:** доля верно соотнесенных документов во всех документах

Чем плоха аккуратность?

Сравнение двух методов

Явный метод

Одинаковая коллекция
Одинаковая индексация
Одинаковый обучающий набор

Неявный метод

Сравнивать каждый метод с неким “эталонным” примитивным методом

Последний слайд

Если не запомните ничего другого:

- Классификация текстов использует методы информационного поиска и машинного обучения
- Три этапа: индексация, построение классификатора, оценка качества
- Классификаторы: разрешающие деревья, метод k соседей, метод Rocchio

Вопросы?