Математические методы анализа текстов

К. В. Воронцов, А. А. Потапенко, А. С. Попов, М. А. Апишев,
Р. Ю. Дербаносов, Н. А. Шаталов

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Математические методы анализа текстов
(курс лекций, К.В.Воронцов, А.А.Потапенко)»

5 сентября 2018

Содержание

Цели курса

- Научиться строить *математические* модели в прикладных задачах анализа текстов
- Знать основные *технологии* анализа текстов и смело применять их в прикладных задачах
- Узнать кучу клёвых красивых алгоритмов, не вошедших в курс машинного обучения

Условная классификация задач анализа текстов

1. По структуре входов–выходов «чёрного ящика»
   - Классификация и предсказательные модели
     вход — текст, **выход** — число
   - Информационный поиск
     вход — текст, **выход** — список документов
   - Преобразование и синтез текста
     вход — текст, **выход** — текст
2. По критерию качества или положению в pipe-line
   - бизнес-задачи
   - вспомогательные задачи компьютерной лингвистики
3. По уровням анализа текста (пирамида NLP)

Задача классификации спама

Дано:

- текстовый документ (e-mail, web-страница)

Найти:

- один из двух классов: спам / не-спам

Критерий:

- AUC, чувствительность и специфичность

Модель классификации строится по обучающей выборке
Основная подзадача: преобразовать текст в векторное признаковое описание фиксированной размерности.

Задача классификации отзывов/обращений

Дано:

- текст отзыва или обращения клиента

Найти:

- класс: куда маршрутизировать запрос / о какой проблеме сообщает клиент или сотрудник

Критерий:

- AUC, чувствительность и специфичность (для многоклассовой классификации)

Модель классификации строится по обучающей выборке.
Основная подзадача: преобразовать текст в векторное признаковое описание фиксированной размерности.

Задача анализа тональности текста (Sentiment Analysis)

Дано:

- текст отзыва или обращения клиента

Найти:

- оценку тональности отзыва в целом, от негативной (-1) через нейтральную (0) до позитивной $(+1)$

Критерий:

- точность определения тональности на размеченных данных

Модель классификации строится по обучающей выборке.
Используются словари тональных слов.

Задача анализа тональности сущностей (Sentiment Analysis)

Дано:

- коллекция текстовых документов
- объект, именованная сущность (named entity)

Найти:

- оценку средней тональности упоминаний объекта в коллекции, от негативной (-1) через нейтральную (0) до позитивной $(+1)$

Критерий:

- точность определения тональности на размеченных данных

Модель классификации строится по обучающей выборке. Используются словари тональных слов и синтаксический анализ для связывания объекта с тональными словами.

Задача категоризации текстовых документов

Дано:

- коллекция текстовых документов
- иерархия категорий (возможно, неполная)
- категории документов (возможно, не все и не всех)

Найти:

- категории неразмеченных документов (многоклассовая классификация с пересечением классов)
- недостающие категории, если таковые имеются (задача кластеризации)

Критерий:

- качество категоризации размеченных документов

Примеры задач регрессии на текстах

MovieReview [Pang, Lee, 2005]
дано: текст отзыва на фильм
найти: рейтинг фильма (1..5), поставленный автором отзыва

Salary (kaggle.com: *Adzuna Job Salary Prediction*)
дано: описание вакансии, предлагаемой работодателем
найти: годовая зарплата

Yelp (kaggle.com: *Yelp Recruiting Competition*)
дано: отзыв (на ресторан, отель, сервис и т.п.)
найти: число голосов «useful», которые получит отзыв

Прогнозирование скачков цен на финансовых рынках
дано: текст новости
найти: изменение цены в последующие 10–60 минут

B. Pang, L. Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales // ACL, 2005.

Конкурс kaggle.com: Avito Context Ad Clicks Prediction

Дано: тройка $\langle$пользователь, запрос, объявление$\rangle$

Найти: вероятность клика по контекстной рекламе, показанной в ответ на запрос пользователя на avito.ru

Критерий:

- внутренний — правдоподобие модели вероятности клика
- внешний бизнес-критерий — доход рекламной площадки

Особенности задачи:

- надо придумывать признаки
- данных много — сотни миллионов показов
- много дополнительных критериев и ограничений

Задача ранжирования поисковой выдачи

Дано: пара $\langle$короткий запрос, документ$\rangle$.

Найти: оценка релевантности документа запросу

Критерий:

- точность и полнота поиска по асессорским данным
- качество ранжирования по асессорским данным

Особенности задачи:

- надо придумывать признаки
- надо распознавать и исправлять опечатки
- надо учитывать словоформы, синонимы, парафразы

Разведочный информационный поиск (exploratory search)

Дано: запрос — один или несколько документов

Найти:

- из каких тем состоит запрос
- семантически близкие документы («о том же самом»)
- релевантные документы по каждой теме
- документы, связанные междисциплинарными связями

Критерий:

- точность и полнота поиска по асессорским данным
- экономия времени пользователей
- эффективность приобретения новых знаний пользователями

Специфические виды поиска

- *Мультиязычный поиск:* найти на всех языках
- *Кроссязычный поиск:* по запросу на одном языке найти документы на заданном другом языке
- *Поиск фрагментов:* не только найти документы, но и указать конкретные места в них
- *Поиск заимствований:* найти фрагменты документа, скопированные из других документов
- *Поиск дубликатов и версий* документа-запроса
- *Поиск ответа на вопрос*

Машинный перевод (Machine Translation)

Дано: текст на одном языке

Найти: его перевод на другой язык

Критерий:

- близость к переводам профессиональных переводчиков
- число исправлений, сделанных переводчиком
- средняя асессорская оценка качества перевода в баллах

Обучающие данные: двуязычные словари и большой корпус параллельных выровненных текстов.

Суммаризация и аннотирование (Summarization)

Дано:

- документ или подборка документов

Найти:

- краткое содержание (реферат)

Критерий:

- точность соответствия (как правило, нескольким) рефератам, написанным людьми (метрики ROUGE, BLUE)

Особенности задачи:

надо учитывать словоформы, синонимы, парафразы

надо выбирать самое важное, но без повторов

Ответы на вопросы (Question Answering)

Дано:

- текст вопроса

Найти:

- текст ответа на поставленный вопрос

Критерий:

- точность выделения фразы ответа на размеченной выборке пар «вопрос – текст-с-ответом»
- средняя асессорская оценка качества ответов в баллах

Обучающие данные: коллекция текстов, возможно, с размеченными ответами на вопросы

Разговорный интеллект (Conversational Intelligence, chatbots)

Дано: текст диалога бота с человеком

Найти: следующую реплику бота

Критерий:

- тест Тьюринга: человек-судья не может отличить собеседника-человека от собеседника-бота
- в бизнес-приложениях: оценка степени удовлетворённости клиента ответом бота
- доля случаев, когда потребность клиента была удовлетворена
- доля случаев, когда оператор принял подсказку бота

Обучающие данные:

коллекция диалогов оператора с клиентом

Пирамида NLP (Natural Language Processing)

Задачи морфологического и синтаксического анализа

- Морфологический анализ, выделение морфем
- Исправление опечаток
- Лемматизация (lemmatization)
- Синтаксический анализ (syntax analysis)
- Автоматическое выделение терминов (automatic term extraction)
- Распознавание именованных сущностей (named entity recognition)
- Разрешение *анафоры* и *кореференции*
- Распознавание *эллипсиса* (намеренного пропуска слов)
- Выделение фактов «объект-субъект-действие» (fact extraction)
- Выделение значений полей (slot filling)

Задачи семантического анализа

- Оценивание семантической близости (semantic similarity)
- Семантические векторные представления (word embedding)
- Тематическое моделирование (topic modeling)
- Сегментация текста (text segmentation)
- Отслеживание событий (event tracking)
- Обнаружение и отслеживание тем (topic detection & tracking)
- Выявление связей между объектами (relation learning)
- Обучение онтологий (ontology learning)

Подходы и методы

- На основе правил (rule-based, regular expression)
- На основе лингвистических ресурсов
- Машинное обучение
- Преобразование последовательностей (sequence-to-sequence)

Лингвистические ресурсы

- Неразмеченные корпуса текстов
- Размеченные корпуса текстов
- Тезаурусы
- Онтологии

Оценивание качества моделей

Анализ модели

- Качественный
  — демонстрируются примеры успешной и неуспешной работы модели
- Количественный
  — используются численные критерии

Количественные оценки модели

- Внутренние (intrinsic)
  — модель оценивается по тому критерию, по которому происходила оптимизация её параметров
- Внешние (extrinsic)
  — модель оценивается по бизнес-критерию или с помощью краудсорсинга (асессоров, кодировщиков)