Введение в анализ данных: Классификация текста

Юля Киселёва
juliakiseleva@yandex-team.ru
Школа анализа данных

План на сегодня

- Постановка задачи, подходы и применения
- Построение и обучение классификатора
- Оценка качества классификации

Виды классификации

Вид ответа:

- точная классификация $F : C \times D \to \{0,1\}$
- ранжирование ответа классификатора: $F : C \times D \to [0,1]$

Необходимые входные данные:

- Корпус, коллекция документов
- Таксономия (категории)

Соотношение категорий:

- Категории не пересекаются
- Категории могут пересекаться
- Бинарная классификация: две непересекающиеся категории

Постановка задачи

Дано:

Категории: $C = \{c_1, \ldots, c_n\}$

Документы: $D = \{d_1, \ldots, d_m\}$

Неизвестная целевая функция: *F : C x D ->{0,1}*

Цель:

Построить классификатор *F'* максимально близкий к *F*

Уточнение:

Построение списка категорий для документа

Построение списка документов для данной категории

Напоминание

Коллекция документов:

- Обучающая коллекция документов
- Дополнение: иногда требуется проверочная коллекция документов для тьюнинга
- Тестовая коллекция документов

Приложения

- Фильтрация документов: распознавание спама
- Автоматическая система управления
- Составление каталогов для веб-страниц
- Классификация новостей
- Интернет - реклама
- Выявление плагиата

Этапы классификации

1. Представление документов в едином формате
2. Обучение классификатора
   - Общая форма классифицирующего правила
   - Настройка параметров
3. Оценка качества
   - Оценка абсолютного качества
   - Сравнение с другими классификаторами

Базовый подход

Исходный документ:

Документ = коллекция слов (термов)

Терм имеет вес по отношению к документу

Множество всех термов ***T***

Каждый терм имеет вес ***wij*** по отношению к документу

Вес терма:

Известный подход: ***wij = TFij*IDFij***

Новые подходы

- По-другому выбирать термы
  - Есть ли варианты?
- По-другому выбирать вес терма в документе
- Использовать дополнительные термы

Уменьшение размерности документов

- Виды уменьшения размерностей (Единый метод для коллекции /Свой для каждой категории)
  - Выбор термов
    - «Средне встречающиеся» термы
    - Использование различных коэффициентов полезности
  - Искусственные термы
    - Кластеризация термов

План на сегодня

- Постановка задачи, подходы и применения
- Построение и обучение классификатора
- Оценка качества классификации

Ранжирование и четкая классификация

- Строим функцию *CSV: D -> [0, 1]*
- Выбираем пороговое значение t*i*

Классификация:

- Пропорциональный метод
- Каждому документу выбрать k- ближайших категорий

Потоковый линейный классификатор

- Документ: $d = (d_1, \ldots, d_n)$
- Вектор полезности каждого терма для категории: $c_i = (c_1, \ldots, c_n)$

$$CSV_i(d) = \vec{d} \bullet \vec{c}_i = \prod c_{ij} d_j$$

$$CSV_i(d) = \frac{\vec{d} \bullet \vec{c}_i}{|\vec{d}\,||\,\vec{c}_i\,|}$$

Как подобрать характеризующий вектор $c_i = (c_1, \ldots, c_n)$?

Потоковый линейный классификатор (2)

Схема обучения:

1. Начинаем: $\vec{c}_i = (1,\ldots,1)$
2. Для каждого документа из обучающей выборки применяем текущее правило
3. При неудаче вносим поправки $+\alpha, -\beta$ в координаты, соответствующие словам неудавшегося документа

Потоковый линейный классификатор (2)

- **Вариации**
  - Поправки при удачной классификации
  - Поправки в неактивные слова
- проверочное множество – это индикатор остановки обучения
- **Преимущества:**
  - Есть обратная связь – обучение можно продолжать и за пределами обучающей коллекции

Метод регрессии

- Обучающая коллекция в матричном виде:
  - Каждый документ – это вектор из весов термов
  - Документы образуют матрицу I размера |Tr| x |T|
  - Степень принадлежности документа к категориям – вектор => для всех документов – матрица O размера |C| x |TR|
- Найти:

MI - O =0 = > min |||MI – O||

$$M = \arg\min_{M} \| MI - O \|$$

Пример для метода регрессии

I -матрица

	T1	T2
D1	0	1
D2	1	1

* M -

O -матрица

	C1	C2
D1	0	1
D2	1	1

= 0

M =

	T1	T2
C1	1	0
C2	0	1

Метод Rocchio

Профайл категории (прототип документа): список взвешенных термов, присутствие или отсутствие которых наиболее хорошо отличает категорию ***Ci*** от других.

Профайл для категории ***Ci*** : $\vec{C}_i =< w_{1i},\ldots,w_{|T|i} >$

$$w_{ik} = \beta \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \alpha \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|}$$

$$POS_i = \{d_j \in T \mid F(d_j, c_i) = 1\} \qquad NEG_i = \{d_j \in T \mid F(d_j, c_i) = 0\}$$

Другие методы

- Вероятностные классификаторы
- Нейронные сети
- Support Vector Machines

План на сегодня

- Постановка задачи, подходы и применения
- Построение и обучение классификатора
- Оценка качества классификации

Как выбрать результат

1. Выбор большинства – результат, который дает большинство
2. Взвешенная линейная комбинация – степень доверия каждому классификатору:

$$\sum_i n_i F'_i(d,c)$$

3. Динамический выбор классификатора – в зависимости от категории

Как выбрать результат(2)

4. Динамическая комбинация классификаторов – объединение «взвешенной линейной комбинации» и «динамического выбора классификатора»

Метрики из информационного поиска

- **Полнота:** отношение количества найденных документов из категории к общему числу документов из категории:

$$\mathrm{Re}\,call = \frac{|D_{rel} \bigcap D_{retr}|}{|D_{rel}|}$$

- **Точность:** доля документов действительно из категории а общем числе документов

$$\mathrm{Pr}\,ecision = \frac{|D_{rel} \bigcap D_{retr}|}{|D_{retr}|}$$

Метрики из информационного поиска (2)

F-мера: среднегармоническое между точностью и полнотой

$$F - measure = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}, \alpha \in [0,1]$$

$$F - measure = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \beta = \frac{(1-\alpha)}{\alpha}$$

$$F_1 - measure = \frac{2PR}{P+R}, \alpha = 1/2, \beta = 1$$

Сравнение двух классификаторов

Явный метод:

- Одинаковая тестовая коллекция (например, новости Reuters)
  - для русского языка есть дорожки РОМИП
- Одинаковый обучающий набор

Неявный метод:

- Сравнить с базовым методом

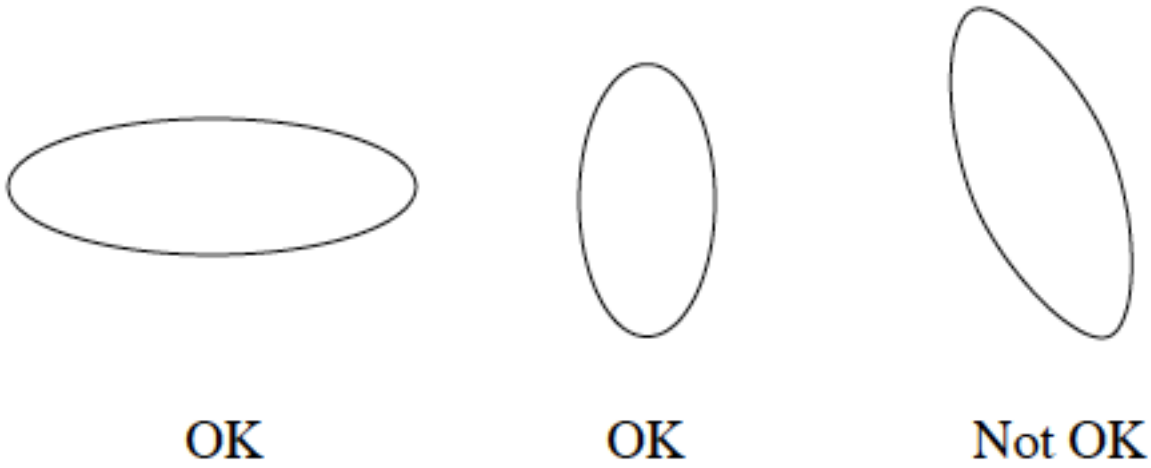
Резюме

Узнали:

- Постановка задачи, подходы и применения
- Построение и обучение классификатора
- Оценка качества классификации

BFR Алгоритм

- BFR[Bradley-Fayyad-Reina] – это вариант алгоритма k-means, который был спроектирован для работы с большими объемами данных
- Предполагается, что кластеры распределены относительно центройда и имеют определенную форму:



План на сегодня

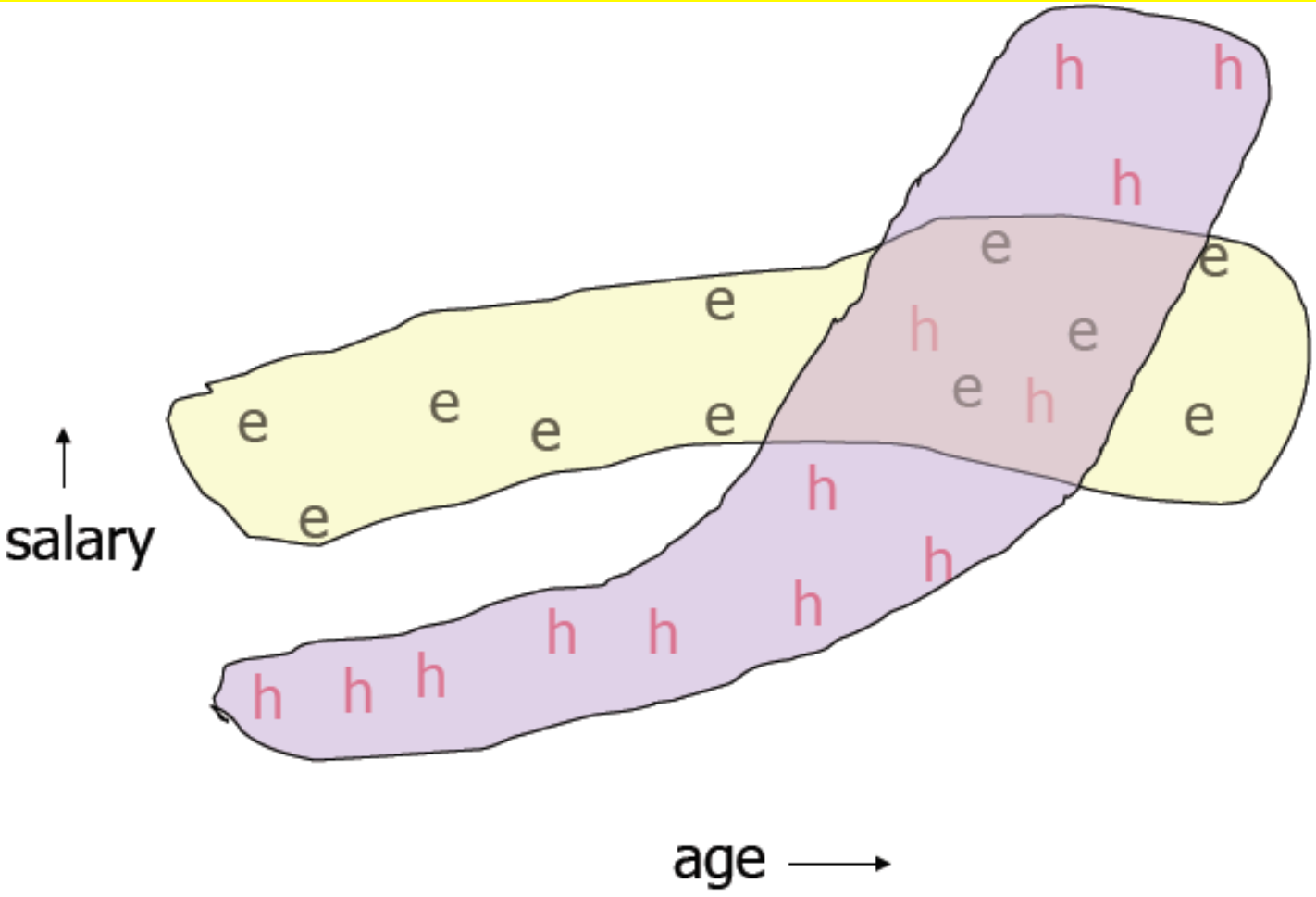
- Задача кластеризации
- Методы кластеризации
- Алгоритм k-means
- **Алгоритм CURE**

CURE Алгоритм

CURE = **C**lustering **U**sing **RE**presentatives

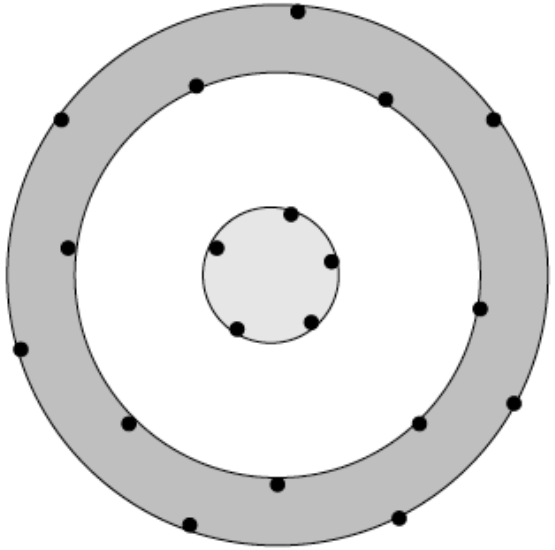
- Евклидово пространство
- Не заботится о форме кластеров
- Кластер представляется коллекцией репрезентативных точек

Пример: зарплата в Стэндфордском Университете



CURE Алгоритм

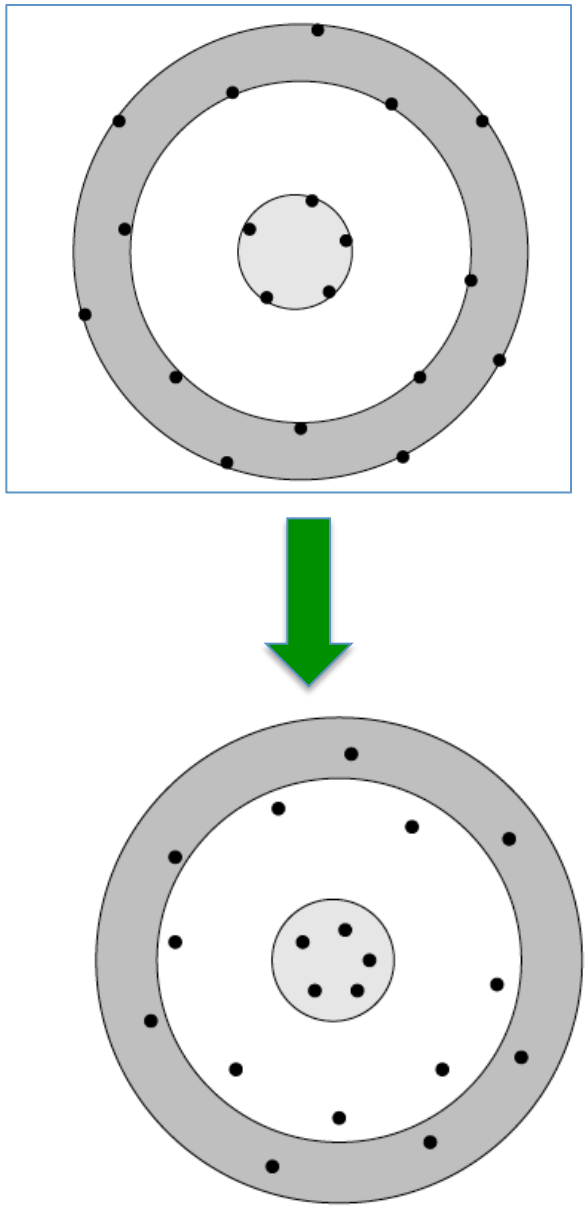
1. Случайным образом выбираем набор точек, которые помещаются в память
2. Кластеризуем этот набор помощью иерархического метода – группируем наиболее близкие точки
3. Для каждого кластера выбираем набор точек (представителей), которые удалены друг от друга насколько это возможно



CURE Алгоритм (2)

4. Из набора нужно выбрать представителей, перемещая их (скажем) 20% в сторону центра тяжести кластера
5. Затем обходим каждую точку ***p*** и перемещаем ее в ближайший кластер.

Определение: «Ближайшим» к ***p*** называется кластер, который содержит большее число ближайших к ***p*** точек



Метрики для оценки

- ***C-index*** (Dalrymple-Alford, 1970)
- ***Gamma*** (Baker & Hubert, 1975)
- ***Adjusted ratio of clustering*** (Roenker et al., 1971)
- ***D-index*** (Dalrymple-Alford, 1970)
- ***Modified ratio of repetition*** *(*Bower, Lesgold, and Tieman, 1969)
- ***Dunn's index*** (Dunn, 1973)
- **Variations of Dunn's index** (Bezdek and Pal, 1998)
- **Jagota index** (Arun Jagota 2003)
- ***Strict separation*** (based on Balacan, Blum, and Vempala, 2008)
- And many more...

Оценка (1)

- Jagota предложил метрику, которая отражает однородность кластера:

$$Q = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} d(\boldsymbol{x}, \mu_i)$$

- где $|\boldsymbol{C_i}|$ - это число элементов в кластере *l*
- *Q* будет маленьким, если (в среднем) точки в кластере близки друг к другу

Gamma

- За d(+) обозначим число раз, когда две точки, которые были кластеризованы вместе в кластер С имели расстояние большее, чем другие две точки не помещенные в один кластер
- За d(-) обозначим противоположный результат

$$\gamma = \frac{d(+) - d(-)}{d(+) + d(-)}$$

Резюме

- Познакомились с задачей кластеризации
- Ввели несколько определений
- k-means
- CURE
- Ввели методы оценки