Text mining

3. Классификация в АОТ

Дмитрий Ильвовский, Екатерина Черняк

dilvovsky@hse.ru, echernyak@hse.ru

Национальный Исследовательский Университет – Высшая Школа Экономики
НУЛ Интеллектуальных систем и структурного анализа

February 10, 2017

1. Задача классификации текстов
   - Наивный Байесовский классификатор
   - Логистическая регрессия (метод максимальной энтропии)

2. Задача классификации последовательности
   - Языковые модели
   - Скрытые цепи Маркова
   - Условные случайные поля

1. Задача классификации текстов
   - Наивный Байесовский классификатор
   - Логистическая регрессия (метод максимальной энтропии)

2. Задача классификации последовательности
   - Языковые модели
   - Скрытые цепи Маркова
   - Условные случайные поля

Наивный Байесовский классификатор

Требуется оценить вероятность принадлежности документа $d \in D$ классу $c \in C$: $p(c|d)$. Каждый документ – мешок слов, всего слов $|V|$.
$p(c|d)$ – апостериорная вероятность класса c
$p(c)$ – априорная вероятность класса c

$$p(c|d) = \frac{p(d|c)p(c)}{p(d)}$$

$$c_{MAP} = \texttt{argmax}_{c \in C} p(c|d) = \texttt{argmax}_{c \in C} \frac{p(d|c)p(c)}{p(d)} \propto$$

$$\propto \texttt{argmax}_{c \in C} p(d|c)p(c) = \texttt{argmax}_{c \in C} p(x_1, x_2, \ldots, x_{|V|}|c)p(c)$$

Предположение о независимости

1. **Мешок слов**: порядок слов не имеет значения
2. **Условная независимость**: вероятности признаков $p(x_i|c_j)$ внутри класса c_j независимы

$$p(x_1, x_2, \ldots, x_{|V|}|c) \times p(c) = p(x_1|c) \times p(x_2|c) \times \cdots \times p(x_{|V|}|c) \times p(c) =$$

$$= p(c) \times \prod_{x_i \in X} p(x_i|c)$$

Обучение наивного Байесовского классификатора

От признаков x_i переходим к словам w_i. ММП оценки:

$$\hat{p}(c_j) = \frac{|\{d|d \in c_j\}|}{|D|}$$

$$\hat{p}(w_i|c_j) = \frac{\texttt{count}(w_i, c_j)}{\sum_{w \in V} \texttt{count}(w, c_j)}$$

Создаем $|C|$ мегадокументов: каждый документ = все документы в одном классе, склеенные в один мегадокумент и вычисляем частоты w в мегадокументах

Проблема нулевых вероятностей: $\texttt{count}(w_i, c_j)$ может быть равно нулю. Допустим, что каждое слово встречается как минимум α раз в мешке слов.

Преобразование Лапласа: $\frac{+\alpha}{+\alpha|V|}$

$$\hat{p}(w_i|c_j) = \frac{\texttt{count}(w_i, c_j) + \alpha}{(\sum_{w \in V} \texttt{count}(w, c_j)) + \alpha|V|}$$

Пример. Тематическая классификация

			class
train	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
test	5	Chinese Chinese Chinese Tokyo Japan	?

$p(c) = \frac{3}{4}, p(j) = \frac{1}{4}$

$p(\texttt{Chinese}|\texttt{c}) = (5+1)/(8+6) = 6/14 = 3/7$

$p(\texttt{Chinese}|\texttt{j}) = (1+1)/(3+6) = 2/9$

$p(\texttt{Tokyo}|\texttt{c}) = (0+1)/(8+6) = 1/14$

$p(\texttt{Tokyo}|\texttt{j}) = (1+1)/(3+6) = 2/9$

$p(\texttt{Japan}|\texttt{c}) = (0+1)/(8+6) = 1/14$

$p(\texttt{Japan}|\texttt{j}) = (1+1)/(3+6) = 2/9$

$p(c|d_5) = 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$

$p(j|d_5) = 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \approx 0.0001$

1. Задача классификации текстов
   - Наивный Байесовский классификатор
   - Логистическая регрессия (метод максимальной энтропии)

2. Задача классификации последовательности
   - Языковые модели
   - Скрытые цепи Маркова
   - Условные случайные поля

Логистическая регрессия (метод максимальной энтропии)

Требуется оценить вероятность принадлежности документа $d \in D$ классу $c \in C$: $p(c|d)$. Пусть заданы признаки $f_i \in F$ – множество признаков и w_i – их веса. Признаки могут зависеть от классов: $f_i(c, d)$
Линейная комбинация этих признаков $\sum_{i=1}^{k} w_i f_i(c, d)$.
Как связана $\sum_{i=1}^{k} w_i f_i(c, x)$ и $p(c|d)$?

$$p(c|d) = \frac{1}{Z} e^{\sum_{i=1}^{k} w_i f_i(c,d)},$$

где

$$\frac{1}{Z} = \frac{1}{\sum_{c' \in C} e^{\sum_{i=1}^{k} w_i f_i(c',d)}}.$$

Логистическая регрессия (метод максимальной энтропии)

$$\hat{c} = \mathtt{argmax}_{c \in C} p(c|d) = \mathtt{argmax}_{c \in C} \frac{e^{\sum_{i=1}^{k} w_i f_i(c,d)}}{\sum_{c' \in C} e^{\sum_{i=1}^{k} w_i f_i(c',d)}} \propto$$

$$\propto \mathtt{argmax}_{c \in C} e^{\sum_{i=1}^{k} w_i f_i(c,d)} \propto$$

$$\propto \mathtt{argmax}_{c \in C} \sum_{i=1}^{k} w_i f_i(c,d).$$

Пример. Классификация по тональности на $C = +, -$

Используем индикаторные признаки.

"... there are virtually no surprises, and the writing is second-rate. So why did I enjoy it so much? For one thing, the cast is great ..."

$$f_1(c_1, d) = \begin{cases} 1, & if \text{ "great"} \in d \text{ and } c = +, \\ 0, & otherwise \end{cases}$$

$$f_2(c_1, d) = \begin{cases} 1, & if \text{ "second} - \text{rate"} \in d \text{ and } c = -, \\ 0, & otherwise \end{cases}$$

$$f_3(c_1, d) = \begin{cases} 1, & if \text{ "no"} \in d \text{ and } c = -, \\ 0, & otherwise \end{cases}$$

$$f_4(c_1, d) = \begin{cases} 1, & if \text{ "enjoy"} \in d \text{ and } c = -, \\ 0, & otherwise \end{cases}$$

$$f_5(c_1, d) = \begin{cases} 1, & if \text{ "great"} \in d \text{ and } c = -, \\ 0, & otherwise \end{cases}$$

$w_1 = 1.9, w_2 = 0.9, w_3 = 0.7, w_4 = -0.8, w_5 = -0.6$

класс $+$:
$1.9 + 0 + 0 + 0 + 0 = 1.9$

класс $-$:
$0 + 0.9 + 0.7 - 0.8 - 0.6 =$
$= 0.2$

$p(+|d) = \frac{e^{1.9}}{e^{1.9} + e^{0.2}}$

$p(-|d) = \frac{e^{1.9}}{e^{1.9} + e^{0.2}}$

Поиск весов логистической регрессии

- Для каждой пары (c, d):

$$\hat{w} = \texttt{argmax}_w \log p(c|d)$$

- Максимизация логарифмического правдоподобия:

$$L(w) = \sum_j \log p(c_j|d)$$

- При использовании индикаторных признаков, методы выпуклой оптимизации позволяют выбрать модель с максимальной энтропией.

Генеративные и дискриминативные классификаторы

- Генеративный классификатор строит модель порождения документа d классом c

$$\texttt{argmax}_{c \in C} p(c|x) = \texttt{argmax}_{c \in C} \frac{p(d|c)p(c)}{p(d)}$$

- Дискриминативный классификатор умеет напрямую различать разные классы c

$$\texttt{argmax}_{c \in C} p(c|x)$$

1. Задача классификации текстов
   - Наивный Байесовский классификатор
   - Логистическая регрессия (метод максимальной энтропии)

2. Задача классификации последовательности
   - Языковые модели
   - Скрытые цепи Маркова
   - Условные случайные поля

Задача классификации последовательности

Британская	актриса	и	крестница	принца	Чарльза	Тара	Томкинсон
Прил.	Сущ.	Союз	Сущ.	Сущ.	Им.Собств.	Им. Собств.	Им. Собств.
O	O	O	O	O	B-Per	B-Per	I-Per
O	O	O	O	O	S-Per	B-Per	E-Per

была	найдена	мертвой	в	ее	квартире	в	Лондоне
Глаг.	Кр. Прич.	Прил.	Пред.	Мест.	Сущ.	Пред.	Им. Собств
O	O	O	O	O	O	O	B-Loc
O	O	O	O	O	O	O	S-Loc

,	сообщает	BBC	.
Пункт.	Глаг.	Им. Собств	Пункт.
O	O	B-Org	O
O	O	S-Org	O

1. Задача классификации текстов
   - Наивный Байесовский классификатор
   - Логистическая регрессия (метод максимальной энтропии)

2. Задача классификации последовательности
   - Языковые модели
   - Скрытые цепи Маркова
   - Условные случайные поля

Языковые модели

- Языковые модели отвечают на вопрос:
  Насколько вероятно, что случайная последовательность слов на естественном языке согласована и соответствует правилам этого языка? $\rightarrow p_{LM}$
- Языковые модели помогают с упорядочиванием слов:
  $p_{LM}($ the house is small $) > p_{LM}($ small the is house $)$
  $p_{LM}($ он поступил в хороший вуз $) > p_{LM}($хороший вуз в он поступил $)$
- Языковые модели помогают с выбором правильного слова:
  $p_{LM}($ I am going home $) > p_{LM}($ I am going house $)$
  $p_{LM}($ я иду на работу $) > p_{LM}($я иду в работу $)$

Языковые модели N-gram

- Пусть дана последовательность слов $W = w_1, w_2, w_3, \ldots, w_n$
- Какова $p(W)$?
- Нельзя оценить по даже по очень большим корпусам – нельзя перечислить все предложения!
- Правило цепи (цепное правило):

$$p(w_1, w_2, w_3, \ldots, w_n) = p(w_1) \times p(w_2|w_1) \times$$

$$\times p(w_3|w_1, w_2) \times \cdots \times p(w_n|w_1, w_2, w_3, \ldots, w_{n-1})$$

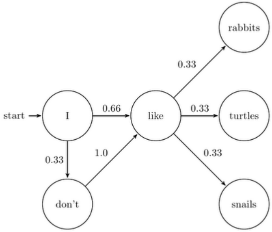
Легче не становится: $p(w_n|w_1, w_2, w_3, \ldots, w_{n-1})$

Цепи Маркова

- Цепь Маркова порядка k – только k предыдущих предысторий (состояний) важны
- Модель биграм:

$$p(w_1, w_2, w_3, \ldots, w_n) = p(w_1) \times p(w_2|w_1) \times$$

$$\times p(w_3|w_2) \times \cdots \times p(w_n|w_{n-1})$$



- ММП оценка вероятностей:

$$p(w_2|w_1) = \frac{\texttt{count}(w_1, w_2)}{\texttt{count}(w_1)}$$

Проблема нулевых вероятностей

Если в каком-то корпусе биграмма "вчерашний снег" не встречалась и $p(\text{снег} \mid \text{вчерашний}) = 0$, это не означает, что эта биграмма неправильная. Но тогда вероятность порождения предложения $p(\text{``растаял вчерашний снег''}) = 0$ тоже нулевая!
Преобразование Лапласа:

$$p(w^n) = \frac{\texttt{count}(w^n) + \alpha}{\texttt{count}(w^{n-1}) + \alpha|V|}$$

Слишком сильное преобразование: число биграмм в корпусе в разы меньше числа возможных биграмм
Преобразование Гуд-Тьюринга:

$$\texttt{count}(w^n)^* = \frac{(\texttt{count}(w^n) + 1) * N_{c+1}}{N_c}$$

N_c – количество биграмм, которые встречаются $\texttt{count}(w^n)$ раз

Использование языковой модели для поиска

1. Построить языковую модель M_d для каждого документа $d \in D$
2. Оценить вероятность порождения запроса q каждой моделью M_d

$$\mathrm{p}(q|M_d) = K_q \prod_{t \in q} P(t|M_d)^{tf(t,q)},$$

   где K_q – мультиномиальный коэффициент для запроса q и может быть проигнорирован, поскольку не зависит от документа
3. Упорядочить документы по убыванию $\mathrm{p}(q|M_d)$

1 Задача классификации текстов
- Наивный Байесовский классификатор
- Логистическая регрессия (метод максимальной энтропии)

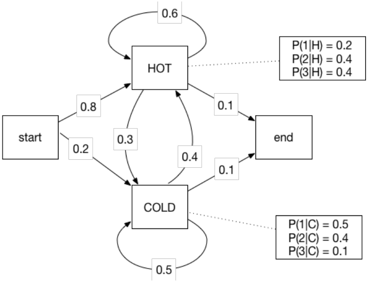
2 Задача классификации последовательности
- Языковые модели
- Скрытые цепи Маркова
- Условные случайные поля

Скрытая цепь Маркова

$< Q, A, O, B, q_0, q_F >$:

- $Q = q_1, \ldots, q_N$ – конечное множество состояний;
- A – матрица вероятностей переходов размером $|Q| \times |Q|$, $0 \le a_{ij} \le 1$;
- O – конечное множество наблюдений;
- B – вероятности наблюдений, $b_i \to \mathbb{R}, \sum_{o \in O} b_i(o) = 1, 1 \le i \le |Q|$;
- q_0, q_F – специальные начальные и конечные символы и соответствующие им вероятности переходов a_{0i}, a_{iF}, $0 \le a_{0i}, a_{iF} \le 1$, $1 \le i \le |Q|$;

$\sum_{j=1}^{|Q|} a_{ij} + a_{iF} = 1, 0 \le i \le |Q|$



Скрытая цепь Маркова

Марковские допущения о независимости:

1. Текущее состояние зависит только от предыдущего состояния:

$$p(q_{i_n}|q_{i_1}\dots q_{i_{n-1}}) = p(q_{i_n}|q_{i_{n-1}})(= a_{i_{n-1}i_n})$$

2. Текущее наблюдение зависит только от текущего состояния:

$$p(o_{i_j}|q_{i_1}\dots q_{i_{n-1}}, o_{i_1}\dots o_{i_{n-1}}) = p(o_{i_j}|q_{i_j})(= b_{i_j}(o_{i_j}))$$

Три задачи скрытых цепей Маркова

1. Оценить вероятность последовательности наблюдений в модели;
2. Найти последовательность состояний, которая с наибольшей вероятностью порождает данную последовательность наблюдений;
3. Оценить параметры модели (обучение по реальным данным).

Первая задача

По последовательности наблюдений $o = o_1 \dots o_n$ оценить вероятность последовательности o. Мы знаем, что:

$$p(o, q) = p(o|q)p(q)$$

Используем допущения о независимости:

$$p(o, q) = \prod_{i=1}^{n} p(o_i|q_i) \prod_{i=1}^{n} p(q_i|q_{i-1})$$

Тогда для всей последовательности наблюдений o:

$$p(o) = \sum_{q \in Q^n} \prod_{i=1}^{n} p(o_i|q_i) \prod_{i=1}^{n} p(q_i|q_{i-1}) p(q_F|q_n)$$

Прямой проход

Идея: используем динамическое программирование для вычисления $n \times |Q|$ значений $\alpha_{ij} = p(o_1 \dots o_i, q_i)$:

1. Инициализация

$$\alpha_{1j} = a_{0j} b(o_1), 1 \leq j \leq |Q|$$

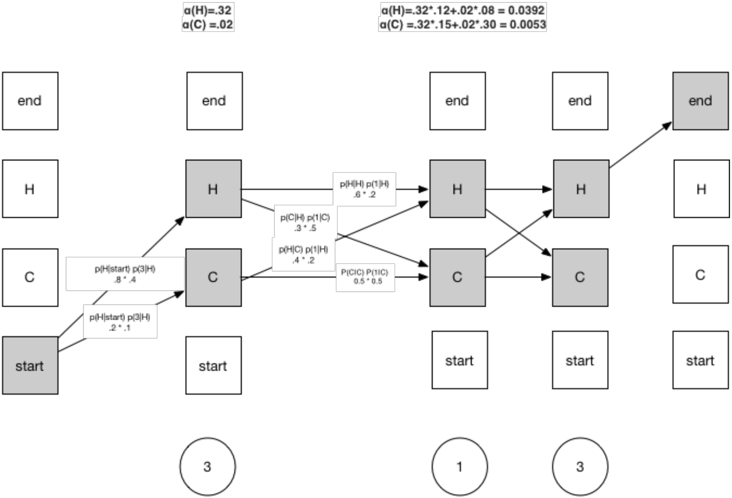
2. Шаг рекурсии

$$\alpha_{ij} = \sum_{k=1}^{|Q|} \alpha_{i-1k} a_{kj} b_j(o_i), 1 \leq i \leq n, 1 \leq j \leq |Q|$$

3. Завершение

$$p(o) = \sum_{k=1}^{|Q|} \alpha_{nk} a_{kF}$$

Вычисление вероятности последовательности наблюдений 313



Декодирование

По последовательности наблюдений $o = o_1 \dots o_n$ определить наиболее вероятную последовательность $q = q_1 \dots q_n \in Q^n$:

$$\mathtt{argmax}_{q \in Q^n} p(o, q) = \mathtt{argmax}_{q \in Q^n} p(o|q) p(q)$$

Используем допущения о независимости:

$$\mathtt{argmax}_{q \in Q^n} p(o, q) = \mathtt{argmax}_{q \in Q^n} \prod_{i=1}^{n} p(o_i|q_i) \prod_{i=1}^{n} p(q_i|q_{i-1})$$

Алгоритм Витерби

Идея: используем динамическое программирование для вычисления $n \times |Q|$ значений $v_{ij} = \max_{q \in Q^{i-1}} p(o_1 \dots o_i, q_1 \dots q_i)$:

1. Инициализация

$$v_{1j} = a_{0j} b(o_1), 1 \le j \le |Q|$$

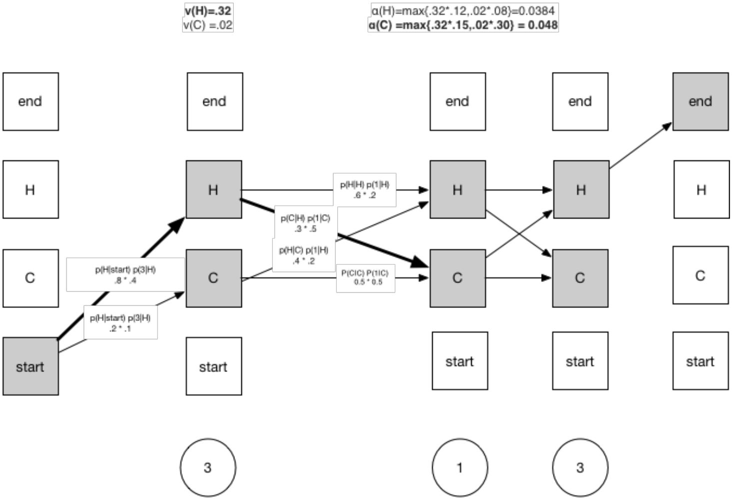
2. Шаг рекурсии

$$v_{ij} = \max v_{i-1k} a_{kj} b_j(o_i), 1 \le i \le n, 1 \le j \le |Q|$$

3. Завершение

$$\max_{q \in Q^n} p(o, q) = \max_{1 \le k \le |Q|} v_{nk} a_{kF}$$

Декодирование последовательности наблюдений 313



1. Задача классификации текстов
   - Наивный Байесовский классификатор
   - Логистическая регрессия (метод максимальной энтропии)

2. Задача классификации последовательности
   - Языковые модели
   - Скрытые цепи Маркова
   - Условные случайные поля

Условные случайные поля

	документ	последовательность
генеративный	Наивный Байес	Скрытые цепи Маркова
дискриминативный	Логистическая регрессия	Условные случайные поля