Машинное обучение

Классификация последовательностей

Екатерина Черняк
urlechernyak@hse.ru

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

December 8, 2017

1. Классификация посследовательности
   - Скрытые цепи Маркова
   - Марковская модель максимальной энтропии
   - Условные случайные поля

Задача классификации последовательности

	Британская	актриса	и	крестница	принца	Чарльза	Тара	Томкинсон
POS	Прил.	Сущ.	Союз	Сущ.	Сущ.	Им.Собств.	Им. Собств.	Им. Собств
IOB (NE)	O	O	O	O	O	B-Per	B-Per	I-Per
IOBES (NE)	O	O	O	O	O	S-Per	B-Per	E-Per
IOBES (R)	O	O	O	O	O	B-Per-1	B-Per-2	I-Per-2

	была	найдена	мертвой	в	ее	квартире	в	Лондоне
POS	Глаг.	Кр. Прич.	Прил.	Пред.	Мест.	Сущ.	Пред.	Им. Собств
IOB (NE)	O	O	O	O	O	O	O	B-Loc
IOBES (NE)	O	O	O	O	O	O	O	S-Loc
IOBES (R)	O	O	O	O	O	O	O	O

	,	сообщает	BBC	.
POS	Пункт.	Глаг.	Им. Собств	Пункт.
IOB (NE)	O	O	B-Org	O
IOBES (NE)	O	O	S-Org	O
IOBES (R)	O	O	O	O

Определение

Обучающие данные:

- $\boldsymbol{x} = x_1, x_2, \ldots, x_n$, $x_i \in V$, V – словарь
- $\boldsymbol{y} = y_1, y_2, \ldots, y_n$, $y_i \in \{1, \ldots, L\}$ – метки
- $\{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), \ldots, (\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})\}$ – обучающие данные
- экспоненциальная сложность: если длина входной последовательности $= n$, всего возможно L^n решний

Требуется обучить классификатор: $\boldsymbol{x} \rightarrow \boldsymbol{y}$

- y – последовательность
- y – дерево (парсинг)

Методы классификации последовательности

- Sequence labelling
  - Скрытые цепи Маркова [Hidden Markov Model, HMM]
  - Марковские модели максимальной энтропии [Maximum-entropy Markov model, MEMM]
  - Условные случайные поля [Conditional random fields, CRF]
  - Реккурентные нейронные сети (biLSTM)
  - (CNN-)biLSTM-CRF [Ma and Hovy, 2016 End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF]
- Structured prediction
  - SVM^{struct}
  - Structured perceptron
- Slot filing
  - biLSTM-CNN-CRF with attention [Liu and Lane, 2016 Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling]

Меры качества классификации последовательностей

1. token-based

   tp – число истинно-положительных токенов, *fp* – число ложно-положительных токенов, *fn* – число ложно-отрицательных токенов

2. chunk-based

   чанк – именованная сущность (синтаксическая группа, и др.) целиком

   tp – число истинно-положительных чанков, *fp* – число ложно-положительных чанков, *fn* – число ложно-отрицательных чанков

1. Классификация посследовательности
   - Скрытые цепи Маркова
   - Марковская модель максимальной энтропии
   - Условные случайные поля

Скрытая цепь Маркова [Hidden Markov Model, HMM]

$$\hat{T} = \arg\max_T P(T|W)$$

$$\arg\max_T P(W|T)P(T)$$

$$\arg\max_T \prod_i P(w_i|t_i) \prod_i (t_i|t_{i-1})$$

T – конечное множество частеречных тегов

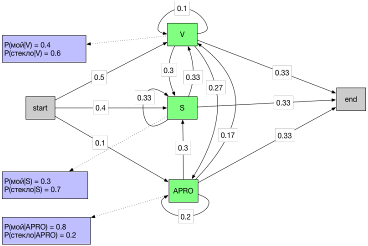
W – конечное множество слов

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

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

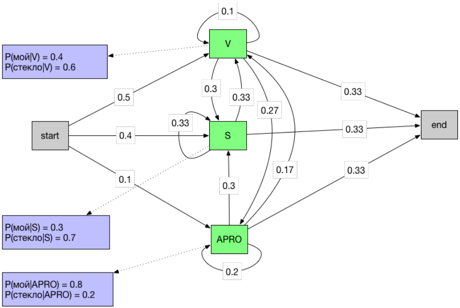
2. Шаг рекурсии

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

3. Завершение

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

Вычисление вероятности последовательности наблюдений “мой стекло”



	start	мой	стекло	end
V	0.5	0.25	0.1219	0.0402
S	0.4	0.12	0.0970	0.0320
APRO	0.1	0.08	0.0167	0.0055

$P(\text{“мой стекло”}) = 0.07775$

Обратный проход

Идея: используем динамическое программирование для вычисления $n \times |Q|$ значений $\beta ij = p(o_{i+1}) \dots o_n, q_i)$:

1. Инициализация

$$\beta_{nj} = a_{jF}, 1 \leq j \leq |Q|$$

2. Шаг рекурсии

$$\beta_{ij} = \sum_{k=1}^{|Q|} \beta_{i+1k} a_{jk} b_k(o_{i+1}), 1 \leq i \leq n, 1 \leq j \leq |Q|$$

3. Завершение

$$p(o) = \sum_{k=1}^{|Q|} a_{0k} b_k(o_1) \beta_{1k}$$

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

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

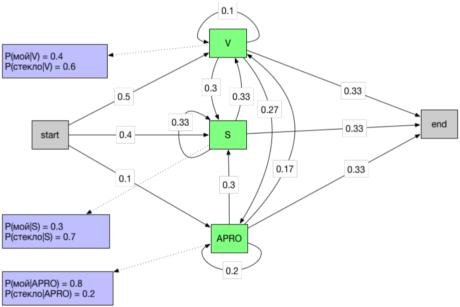
2. Шаг рекурсии

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

3. Завершение

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

Декодирование последовательности наблюдений “мой стекло”



	start	мой	стекло	end
V	0.5	**0.25**, start	0.015, V	0.0046, S
S	0.4	0.12, start	**0.0525**, V	**0.0177**, S
APRO	0.1	0.08, start	0.0135 V	0.0045, S

наиболее вероятная последовательность скрытых состояний: V S

$p(\text{“мой стекло”}, \text{V S}) = 0.0177$

TnT POS-tagger [Brants, 2000]

TnT использует скрытую Марковскую цепь второго порядка для того, чтобы найти частеречные тэги:

$$\arg\max[\prod_j [p(o_i|t_{o-1}, t_{o-2})p(q_i|o_i)]P(o_{T+1}|o_T)$$

Вероятность тэга для данного слова определяется как линейная интерполяция вероятностей, полученных из трех Марковских цепей::

$$P(o_i|o_{i-1}, o_{i-2}) = l_1 * P(o_i) + l_2 * P(o_i|o_{i-1}) + l_3 * P(o_i|o_{i-1}, o_{i-2})$$

nltk.tag.tnt

```
In[1]: from nltk.tag import tnt
In[2]: tnt_pos_tagger = tnt.TnT()
In[3]: tnt_pos_tagger.train(train_data)
In[4]: tnt_pos_tagger.evaluate(test_data)
```

Марковская модель максимальной энтропии [McCallum, 2000], [Toutanova,2003]

Марковская модель максимальной энтропии [Maximum-entropy Markov model, MEMM]

$$\hat{T} = \arg\max_T P(T|W)$$

$$\arg\max_T \prod_i P(t_i) \prod_i (t_i|w_i, t_{i-1})$$

T – конечное множество частеречных тегов

W – конечное множество слов

Метод максимальной энтропии, MaxEnt

Индикаторные признаки:

У/PR страха/S глаза/S велики/(S или A) ./PUNCT

$$f_{11}(c,x) = \begin{cases} 1, & if \;\; t_{-1} = S, c = S \\ 0, & otherwise \end{cases} \qquad f_{12}(c,x) = \begin{cases} 1, & if \;\; t_{-1} = S, c = A \\ 0, & otherwise \end{cases}$$

$$f_{21}(c,x) = \begin{cases} 1, & if \;\; w_{-1}[:-1] = a, c = S \\ 0, & otherwise \end{cases} \qquad f_{22}(c,x) = \begin{cases} 1, & if \;\; w_{-1}[:-1] = a, c = A \\ 0, & otherwise \end{cases}$$

$$f_{31}(c,x) = \begin{cases} 1, & if \;\; w_{+1} = ``.'', c = S \\ 0, & otherwise \end{cases} \qquad f_{32}(c,x) = \begin{cases} 1, & if \;\; w_{+1} = ``.'', c = A \\ 0, & otherwise \end{cases}$$

$$f_{41}(c,x) = \begin{cases} 1, & if \;\; w_{+1} = ``?'', c = S \\ 0, & otherwise \end{cases} \qquad f_{34}(c,x) = \begin{cases} 1, & if \;\; w_{+1} = ``?'', c = A \\ 0, & otherwise \end{cases}$$

$\lambda_{11} = 0.3, \lambda_{21} = 0.4, \lambda_{31} = 0.1, \lambda_{41} = 0.2$

$\lambda_{12} = 0, \lambda_{22} = 0.2, \lambda_{32} = 0, \lambda_{42} = 0.1.$

$P(S|\text{велики}) = \frac{e^{0.3+0.1+0.4}}{e^{0.3+0.1+0.4}+e^{0.2}}$

$P(A|\text{велики}) = \frac{e^{0.2}}{e^{0.3+0.1+0.4}+e^{0.2}}$

$P(S|\text{велики}) > P(A|\text{велики})$

Марковская модель максимальной энтропии

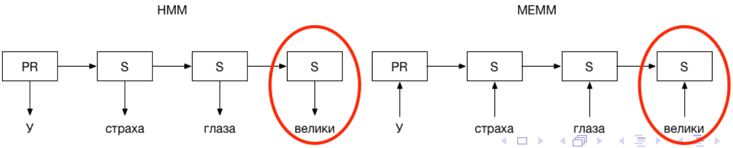
По аналогии с HMM и MaxEnt:

$$P(Q|O) = \prod_{i=1}^{n} P(q|q_{i-1}, o_i)$$

$$P(q|q', o) = \frac{e^{\sum_i w_i f_i(o,q)}}{Z(o, q')}$$

Сравнение HMM и MEMM

1. HMM и MEMM моделируют последовательности: существуют скрытые состояния (частеречные теги), порождающие наблюдения (слова). По последовательности наблюдений требуется определить, какие скрытые состояния их породили;
2. Для декодирования HMM и MEMM используется алгоритмы Витерби, для обучения – EM алгоритм;
3. MEMM позволяет ввести дополнительные индикаторные признаки, поэтому может считаться расширением HMM;
4. HMM – генеративная модель и моделирует $P(O, Q)$, MEMM – дискриминативная и моделирует $P(Q|O)$, что и требуется для декодирования;
5. В MEMM используется локальная нормировка на Z и преимущество получают состояния с меньшей энтропией – меньшим числом переходов, т.н. "label bias problem".



Условные случайные поля [Lafferty, 2001]

Условные случайные поля [Conditional random fields]

$$\hat{T} = \arg\max_T P(T|W) = \phi(t_i, t_{i-1})\phi(t_i, w_i)$$

Условные случайные поля

Вероятность последовательности меток классов для входной последовательности определяется по признакам, которые называются потенциальными функциями. Эти признаки помогают связать класс текущего наблюдения x_i с классами других наблюдений. Для формализации признаков чаще всего используются индикаторные функции. Таким образом, задача обучения сводится к определению весов индикаторных функций.

$$t(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{``}June'' \ and \ y_{i-1} = IN \ and \ y_i = NNP \\ 0, & otherwise \end{cases}$$

$$s(y_{i-1}, x, i) = \begin{cases} 1, & if \ x_i = \text{``}to'' \ and \ y_i \\ 0, & otherwise \end{cases}$$

Условные случайные поля

Для того, чтобы найти вероятность последовательности классов для входной последовательности:

1. извлекаем признаки
2. находим их веса и линейную комбинацию их признаков с найденными весами
3. используем softmax для определения искомых вероятностей.

Обозначим все признаки: $f(y_{i-1}, y_i, x, i)$. Признаки для последовательностей: $F(y, x) = \sum_{i=1}^{n} f(y_{i-1}, y_i, x, i)$. Обозначим веса признаков через λ. Искомая вероятность:

$$p(y|x) = \frac{e^{\sum_{i=1}^{k} \lambda_i F_i(y,x)}}{\sum_{y' \in C^n} e^{\sum_{i=1}^{k} \lambda_i F_i(y',x)}}$$

Пример. POS-тэггинг

Для 4 тэгов (Det, N, Adv, V) задано признаковое пространство:

$$f_1(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{“}chief\text{''} \ and \ y_{i-1} = Det \ and \ y_i = Adj \\ 0, & otherwise \end{cases}$$

$$f_2(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{“}chief\text{''} \ and \ y_{i-1} = N \\ 0, & otherwise \end{cases}$$

$$f_3(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{“}talks\text{''} \ and \ y_{i-1} = Det \ and \ y_i = N \\ 0, & otherwise \end{cases}$$

$$f_4(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{“}talks\text{''} \ and \ y_{i-1} = Adj \ and \ y_i = N \\ 0, & otherwise \end{cases}$$

$$f_5(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{“}talks\text{''} \ and \ y_{i-1} = N \ and \ y_i = V \\ 0, & otherwise \end{cases}$$

$$f_6(y_{i-1}, y_i, x, i) = \begin{cases} 1, & if \ x_i = \text{“}the\text{''} \ and \ y_{i-1} = Det \\ 0, & otherwise \end{cases}$$

Веса: $\lambda_1 = 2, \lambda_2 = 5, \lambda_3 = 9, \lambda_4 = 8, \lambda_5 = 7, \lambda_6 = 20$.

Пример. POS-тэггинг

Сравним вероятности $p(\text{Det N V} \mid \text{the chief talks})$, $p(\text{Det Adj V} \mid \text{the chief talks})$.

1. Det N V: $20 + 5 + 7 = 32$
2. Det Adj V : $20 + 2 + 8 = 30$

Сравненение MEMM и CRF

MEMM

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{wf(\mathbf{x},i,y_{i-1},y_i)}}{Z(\mathbf{x}, y, y_{i-1}; w)}$$

CRF

$$P(\mathbf{y}|\mathbf{x}) = \frac{\sum_{i=1}^{n} e^{wf(\mathbf{x},i,y_{i-1},y_i)}}{Z(\mathbf{x})}$$

- Одинаковые признаки $f(\mathbf{x}, i, y_{i-1}, y_i)$
- MEMM локально нормализованы, CRF – глобально
- Label bias: MEMM поощряет разборы с меньшим количеством переходов