Анализ текстов

Ульянкин Филипп

21 сентября 2019 г.

Немного про языковые модели

Языковые модели

Языковые модели

- Пытаются оценить вероятность конкретной последовательности токенов, при каких-то предпосылках на процесс порождения данных
- Пример предпосылки (наивный Байес):

$$P(text) = P(w_1, \dots, w_n) = P(w_1) \cdot \dots \cdot P(w_n)$$

- Другой пример (LDA):

$$P(text) = P(w_1, \dots, w_n) = \prod_{w \in V} \sum_{t \in T} P(w \mid t) \cdot P(t \mid d)$$

- Новый пример:

$$P(text) = P(w_1, \dots, w_n) = P(w_1) \cdot P(w_2 \mid w_1) \cdot \dots \cdot P(w_n \mid w_1, \dots w_n)$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{house}|\text{this is the}) =$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{house}|\text{this is the}) =$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{house}|\text{this is the}) = 0.25$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{house}|\text{this is the}) = 0.25$$

Модель N-грамм

- В примере с домом Джека мы оценивали вероятность на основе 4-грамм, но можно использовать любые n-граммы
- Такие модели называются счётными языковыми моделями
- Давайте посмотрим на пример с биграммой!

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{Jack}|\text{that}) =$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{Jack}|\text{that}) =$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{Jack}|\text{that}) =$$

Дом, который построил Джек

This is the house that Jack built.
This is the malt
That lay in the house that Jack built.
This is the rat,
That ate the malt
That lay in the house that Jack built.
This is the cat,
That killed the rat,
That ate the malt
That lay in the house that Jack built.

$$P(\text{Jack}|\text{that}) = 0.4$$

Модель N-грамм

- В примере с домом Джека мы оценивали вероятность на основе 4-грамм, но можно использовать любые n-граммы
- Мы умеем подсчитывать вероятности и по большому корпусу текстов понимать какое слово пойдёт следующим
- Однако есть несколько нюансов и тонкостей…
- Например, какова вероятность не отдельного текста, а всей последовательности?

И снова о предпосылках

- Наивная Байесовская предпосылка:

$$P(text) = P(w_1, ..., w_n) = P(w_1) \cdot ... \cdot P(w_n)$$

И снова о предпосылках

- Наивная Байесовская предпосылка:

$$P(text) = P(w_1, \dots, w_n) = P(w_1) \cdot \dots \cdot P(w_n)$$

- Цепное правило:

$$P(text) = P(w_1, \dots, w_n) = P(w_1) \cdot P(w_2 \mid w_1) \cdot \dots \cdot P(w_n \mid w_1, \dots, w_{n-1})$$

И снова о предпосылках

- Наивная Байесовская предпосылка:

$$P(text) = P(w_1, \dots, w_n) = P(w_1) \cdot \dots \cdot P(w_n)$$

- Цепное правило:

$$P(text) = P(w_1, \dots, w_n) = P(w_1) \cdot P(w_2 \mid w_1) \cdot \dots \cdot P(w_n \mid w_1, \dots, w_{n-1})$$

- Правило Маркова (смотрим только на k запаздываний):

$$P(w_i \mid w_1, \dots, w_{i-1}) = P(w_i \mid w_{i-k+1}, \dots, w_{i-1})$$

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

$P(\text{this is the house}) =$

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

$$P(\text{this is the house}) = P(\text{this}) \cdot P(\text{is} \mid \text{this}) \cdot P(\text{the} \mid \text{ is}) \cdot \cdot P(\text{house} \mid \text{the})$$

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

$$P(\text{this is the house}) = P(\text{this}) \cdot P(\text{is} \mid \text{this}) \cdot P(\text{the} \mid \text{ is}) \cdot \cdot P(\text{house} \mid \text{the})$$

$$P(\text{this is the house}) = \frac{1}{12} \cdot 1 \cdot 1 \cdot \frac{1}{2}$$

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

$$P(\text{this is the house}) = P(\text{this}) \cdot P(\text{is} \mid \text{this}) \cdot P(\text{the} \mid \text{ is}) \cdot \cdot P(\text{house} \mid \text{the})$$

$$P(\text{this is the house}) = \frac{1}{12} \cdot 1 \cdot 1 \cdot \frac{1}{2}$$

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

$$P(\text{this is the house}) = P(\text{this}) \cdot P(\text{is} \mid \text{this}) \cdot P(\text{the} \mid \text{ is}) \cdot \cdot P(\text{house} \mid \text{the})$$

$$P(\text{this is the house}) = \frac{1}{12} \cdot 1 \cdot 1 \cdot \frac{1}{2}$$

Проблема: первое слово в корпусе That или this, но не другое. Мы можем это учесть и ввести фиктивный токен для старта предложения

Пример: Марковское свойство 2-го порядка

$$P(text) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot ... \cdot P(w_n \mid w_{n-1})$$

$$P(\text{this is the house}) = P(\text{this} \mid \text{start}) \cdot P(\text{is} \mid \text{this}) \cdot P(\text{the} \mid \text{ is}) \cdot \cdot P(\text{house} \mid \text{the})$$

$$P(\text{this is the house}) = \frac{1}{2} \cdot 1 \cdot 1 \cdot \frac{1}{2}$$

Проблема: первое слово в корпусе That или this, но не другое. Мы можем это учесть и ввести фиктивный токен для старта предложения

Пример: Марковское свойство 2-го порядка

- Другая проблема: последовательности бывают разной длины и вероятности в сумме дают 1 только в каждом классе.

Пример: Марковское свойство 2-го порядка

- Другая проблема: последовательности бывают разной длины и вероятности в сумме дают 1 только в каждом классе.
- Лечение: нужен ещё один фейковый токен в самом конце

$$P(text) = P(w_1 \mid start) \cdot P(w_2 \mid w_1) \cdot ... \cdot P(w_n \mid w_{n-1}) \cdot P(end \mid w_n)$$

Оценивание

- Метод максимального правдоподобия:

$$P(text) = \prod_{i=1}^{n} P(w_i \mid w_{i-1});\ w_0 = \text{start}, w_n = \text{end}$$

Оценивание

- Метод максимального правдоподобия:

$$P(text) = \prod_{i=1}^{n} P(w_i \mid w_{i-1});\ w_0 = \text{start}, w_n = \text{end}$$

- Максимизируем и получаем, что

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}w_i)}{c(w_{i-1})}$$

- Новая проблема: нулевые вероятности, заниженные вероятности из-за ограниченности корпуса

Сглаживание

Просто добавляем некоторый коэффициент α к встречаемости каждой N-граммы. Например, при $\alpha = 1$,

$$P(w_i|w_{i-1}) = \frac{\text{count}(w_{i-1}w_i) + \alpha}{\text{count}(w_{i-1}) + \alpha V}$$

V — объем словаря

Сглаживание

- Сглаживание Лапласа (add-one)
- Сглаживание Кнесера-Нея (Kneser-Ney)
- Сглаживание Виттена-Белла (Witten-Bell)
- Сглаживание Гуда-Тьюринга (Good-Turing)
- Откат (backoff)

Цепи Маркова со скрытым состоянием

Постановка задачи

- Есть последовательность токенов
- Хотим сгенерировать последовательность тэгов для неё
- Примеры: тэгирование частей речи, распознавание именованных сущностей и тп

PoS tagging

Open class words	
ADJ	adjective
ADV	adverb
INTJ	interjection
NOUN	noun
PROPN	proper noun
VERB	verb

Other	
PUNCT	punctuation
SYM	symbol
X	other

Closed class words	
ADP	adposition
AUX	auxiliary verb
CCONJ	coordinating conjunction
DET	determiner
NUM	numeral
PART	particle
PRON	pronoun
SCONJ	subordinating conjunction

http://universaldependencies.org

Способы решения задачи

- Rule-based models
- Классификаторы меток для каждого токена
- Sequence models (HMM, MEMM, CRF)
- Нейросетки

PoS tagging with HMMs

- Последовательность видимых состояний (слов):

$$x = x_1, \dots x_n$$

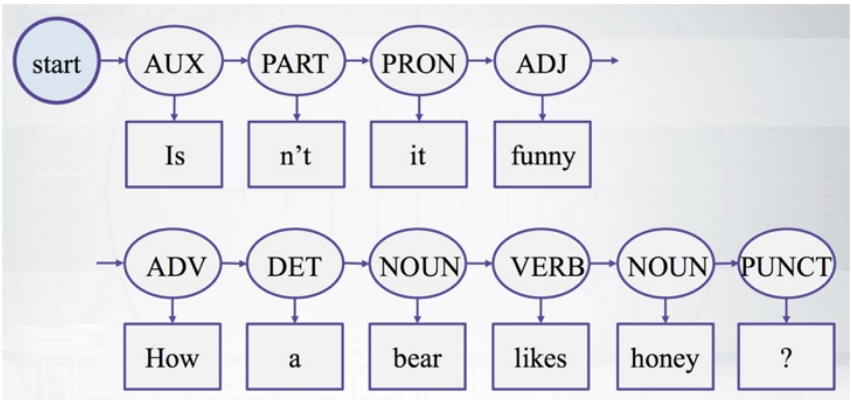
- Последовательность скрытых состояний (тэгов):

$$y = y_1, \dots, y_n$$

- Хотим по x воспроизводить самую вероятную последовательность y

$$y = \arg\max_y P(y \mid x)$$

Процесс порождения данных



PoS tagging with HMMs

- Последовательность видимых состояний (слов):

$$x = x_1, \dots x_n$$

- Последовательность скрытых состояний (тэгов):

$$y = y_1, \dots, y_n$$

- Хотим по x воспроизводить самую вероятную последовательность y

$$y = \arg\max_y P(y \mid x) = \arg\max_y \frac{P(x \mid y) \cdot P(y)}{P(x)} = \arg\max_y P(y, x)$$

Hidden Markov Model

$$P(x, y) = P(x \mid y) \cdot P(y) \approx \prod_{i=1}^{n} P(x_i \mid y_i) \cdot P(y_i \mid y_{i-1})$$

Марковское предположение:

$$P(y) = \prod_{i=1}^{n} P(y_i \mid y_{i-1})$$

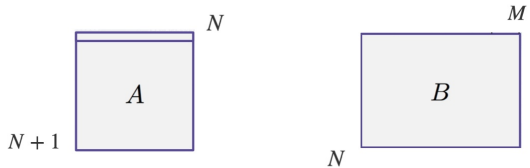
Независимость выходов:

$$P(x \mid y) \approx \prod_{i=1}^{n} P(x_i \mid y_i)$$

Параметры модели

- HMM определяется множеством скрытых состояний $S = \{s_0, \dots, s_N\}$, где s_0 стартовое скрытое состояние
- Матрица перехода между скрытыми состояниями A
- Множество видимых состояний $O = \{o_1, \dots, o_M\}$
- Матрица перехода из скрытых в видимые состояниями B

Параметры модели



Если бы мы знали все метки, мы могли бы получить оценки максимального правдоподобия:

$$a_{ij} = P(s_j \mid s_i) = \frac{c(s_i \to s_j)}{c(s_i)}$$

$$b_{ik} = P(o_k \mid s_i) = \frac{c(s_i \to o_k)}{c(s_i)}$$

Параметры модели

- В реальности мы обычно не знаем меток :(
- В формуле фигурируют вероятность перехода и тэга

$$\arg\max_y P(y, x) = \arg\max_y \prod_{i=1}^{n} P(x_i \mid y_i) \cdot P(y_i \mid y_{i-1})$$

- В предположении, что тэги могут быть любыми можно попробовать найти оптимальную последовательность с помощью алгоритма Витерби

Алгоритм Витерби

- Алгоритм Витерби - это просто динамическое программирование, как в задаче про рюкзак

$$q_{ts} = \max_{s'} q_{t-1,s'} \cdot P(s \mid s') \cdot P(o_t \mid s)$$