Обработка естественного языка, лекция 3: вероятностные алгоритмы

Пожидаев М. С.

24 ноября 2021 г.

Применение СММ

Скрытые марковские модели применяются для решения следующих задач в лингвистике:

1. Определение частей речи.
2. Распознавание речи (один из первых применённых алгоритмов).
3. Восстановление повреждённого текста.

Скрытая марковская модель в дискретном времени

Заданы множества скрытых состояний $X = \{x_1, x_2, \ldots, x_n\}$, множество наблюдаемых состояний $Y = \{y_1, y_2, \ldots, y_m\}$ и множество шагов времени $\{t_1, t_2, \ldots, t_l\}$ ($timestep$).
Фактические состояния в моменты времени неизвестны, их следует определить, но известны наблюдаемые состояния.
Также заданы матрица вероятности S_{ij} перехода из одного состояния x_i в другое, $P(y_i|x_j)$ — вероятность наблюдения y_i при скрытом состоянии x_j и z — вектор вероятности начальных состояний.

Алгоритм Витерби

Алгоритм позволяет определить последовательность скрытых состояний (путь Витерби), имеющую максимальную вероятность среди всех возможных вариантов последовательности. Порядок вычисления матрицы в алгоритме Витерби следующий:

$$V_{1k} = P(y_1|k)z_k$$
$$V_{tk} = \max_{x \in X}(P(y_t|k)S_{xk}V_{(t-1)x})$$

Недостатки СММ

Скрытые марковские модели имеют следующие недостатки:

- последовательность анализируется только в одном направлении;
- вероятность перехода определяется предысторией единичной длины.

Тематическое моделирование

Тематическое моделирование — приписывание документов из коллекции документов к одному из нескольких тематических кластеров.

1. Позволяет выделять скрытые темы, о существовании которых пользователь не подозревал.
2. Современные алгоритмы предлагают моделировать темы как распределение вероятностей слов в каждом кластере.
3. Порядки слов и документов чаще всего не играют значения. Учитывается только факт употребления слова в документе.

Смесь распределений вероятностей

$$p(x) = \sum_{j=1}^{k} w_j p_j(x)$$

$$\sum_{j=1}^{k} w_j = 1, w_j >= 0$$

$$p_j(x) = \phi(x; \theta_j)$$

w_j — вес компоненты, $\phi(x)$ — заданная функция правдоподобия., $\Theta = \{w_1, w_2, \ldots, w_j; \theta_1, \theta_2, \ldots, \theta_j\}$ — искомые параметры.

Задача EM-алгоритма

Выражение ниже следует максимизировать при всех возможных Θ и при фиксированных x_i, которые являются входными параметрами:

$$\ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i; \theta_j)$$

Выражение получено на основе метода максимального правдоподобия из математической статистики. К сожалению, найти аналитически максимум логарифма суммы невозможно, но это можно сделать численно.

Шаг Expectation

Введём матрицу скрытых переменных:
$H_{ij} = P(\theta_j|x_i)$ — вероятность того, что x_i принадлежит j-ой компоненте. Размер матрицы $m \times k$.
По-формуле Байеса:

$$H_{ij} = \frac{w_j p_j(x_i)}{p(x_i)} = \frac{w_j p_j(x_i)}{\sum\limits_{s=1}^{k} w_s p_s(x_i)}$$

$$\sum_{j=1}^{k} H_{ij} = 1$$

Шаг Maximization

Оптимальные веса w_j вычисляются как среднее арифметическое значение для j:

$$w_j = \frac{1}{m} \sum_{i=1}^{m} H_{ij}$$

Максимизация суммы превращается в максимизацию каждого параметра θ_i по отдельности:

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{m} H_{ij} \ln \phi(x_i; \theta)$$

Это утверждение требует отдельного доказательства, но оно справедливо.

Спасибо за внимание!

Веб-сайт: https://marigostra.ru/
E-mail: msp@luwrain.org