Maximum-Entropy Markov Model

Владислав Самсонов

Moscow Institute of Physics and Technology

vvladxx@yandex-team.ru

November 27, 2016

Maximum Entropy Markov Models

Дано

Упорядоченная последовательность наблюдений (observations) $\{o_t\}_{t=1}^n : o_1, o_2, ..., o_n$ и **конечное** множество ответов S.
Примеры: текст, ДНК, значение амплитуды в момент времени t.

Задача

Присвоить каждому наблюдению o_t ответ $s_t \in S$, максимизирующий условную вероятность $P(s_1, ..., s_n | o_1, ..., o_n)$:

$$\{s_1, ..., s_n\} = \text{argmax}_{s_1, ..., s_n} P(s_1, ..., s_n | o_1, ..., o_n)$$

Примеры: определить часть речи для слов из текста.

Maximum Entropy Markov Models

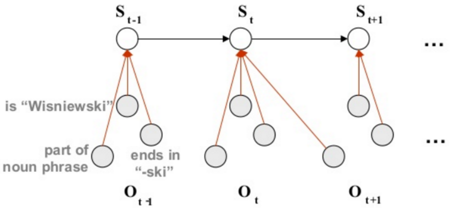
Идея

Заменить генеративную модель в HMM моделью максимальной энтропии.

Maximum Entropy Markov Models

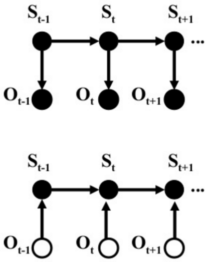
Предположение о марковости: переходные вероятности зависят только от текущего наблюдения и прошлого ответа.

$$P(s_1, \ldots, s_n | o_1, \ldots, o_n) = \prod_{t=1}^{n} P(s_t | s_{t-1}, o_t)$$

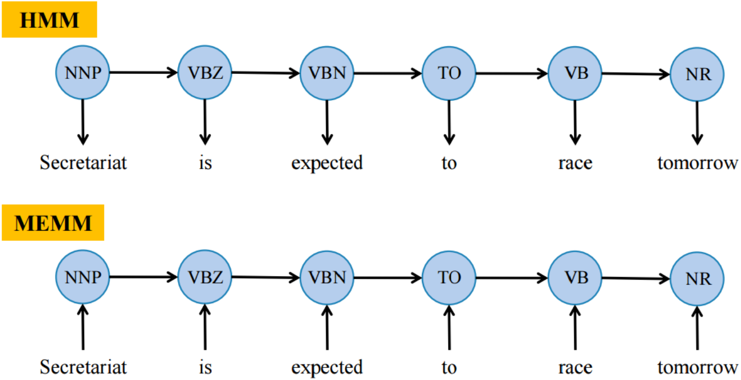


MEMM vs HMM

- HMM: максимиз. $P(s_1, ..., s_n, o_1, ..., o_n) = \prod_{t=1}^{n} P(s_t|s_{t-1})P(o_t|s_{t-1})$
- MEMM: максимизируем $P(s_1, ..., s_n|o_1, ..., o_n) = \prod_{t=1}^{n} P(s_t|s_{t-1}, o_t)$
- HMM: пытаемся предсказывать наблюдения и ответы (но наблюдения нам известны, нужны только ответы).
- MEMM: пытаемся предсказать только ответы.



MEMM vs HMM

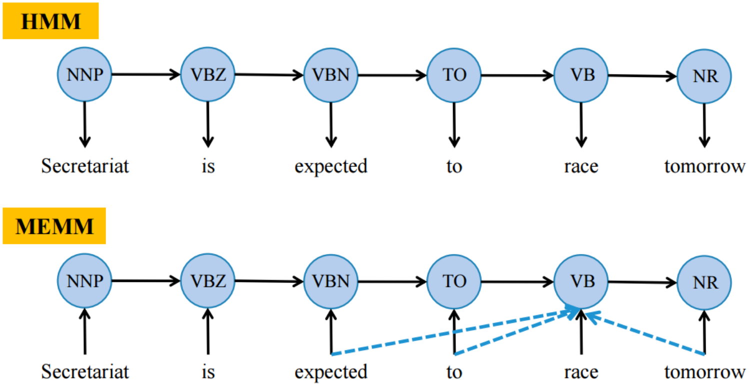

HMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow
MEMM
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow

Недостатки HMM

- Сильные предположения о данных: требуется независимость наблюдений $O \implies$ не учитываются взаимодействия между признаками (в MEMM нет такого предположения).
- В каждом состоянии учитывается только одно наблюдение.
- Размерность вектора $o_t \in O$ фиксирована.
- Нет свободы в выборе признаков. На практике, при попытке учитывать сложные глобальные и даже локальные признаки сильно возрастает размерность вектора o_t.
  Если попробовать учесть несколько соседних наблюдений вместо одного, то размерность растет.
  Под глобальные признаки вроде количества точек в тексте нужно увеличивать размерность.

Достоинства MEMM

- Не требуется независимость наблюдений.
- MEMM: прямое моделирование $P(s_t|s_{t-1}, o_t)$ позволяет вычислить $P(s_1, ..., s_n|o_1, ..., o_n)$.
- Количество наблюдений в момент времени t может быть произвольным.

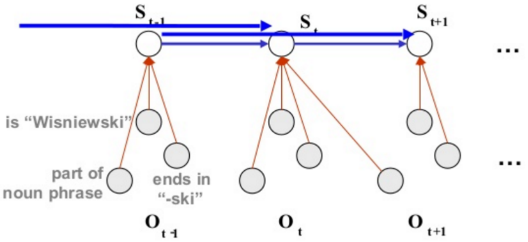


Недостатки MEMM

- Семейства априорных распределений, которые можно использовать для S, не отличаются разнообразием.
- Невозможно получить в некоторый момент времени вероятность произвольного наблюдения, т.к. мы изначально не учитывали это в модели. Есть возможность получить только вероятность ответа.

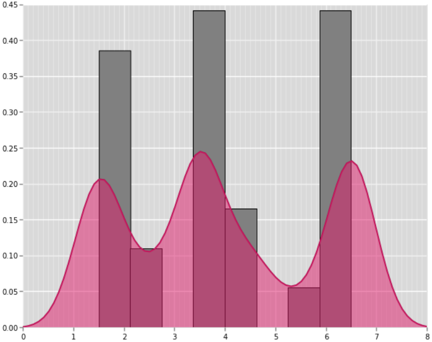
Примеры признаков для MEMM

- Само слово.
- Начинается ли с заглавной буквы.
- Является ли ссылкой.
- Является ли именем собственным.
- Количество букв в слове.
- Количество слов в тексте.
- ...



Проблемы

Эмпирические оценки низкочастотных признаков ненадежны и могут приводить к переобучению. Выбрасывание низкочастотных признаков не всегда помогает, т.к. эти признаки могут быть важны.
Решение: гауссовское сглаживание. Зададим априорное гауссовское распределение на параметры (оценка апостериорного максимума, maximum a posteriori estimation, MAP inference).



Метод максимальной энтропии

Метод максимальной энтропии – это метод для оценки распределения по данным.
Идея: принимаем гипотезу, которая максимизирует энтропию с дополнительными ограничениями.
Мотивация: распределение должно удовлетворять данным и делать как можно меньше предположений о данных (как можно ближе к равномерному распределению).

Метод максимальной энтропии

Задача

Хотим найти распределение, которое максимизирует энтропию при заданных линейных ограничениях.

$$\begin{aligned} \underset{p}{\text{maximize}} \quad & \left(- \sum_{x \in A \times B} p(x) \log p(x) \right) \\ \text{subject to} \quad & f_i(x) = c_i \end{aligned}$$

Идея решения

Это задача нелинейного программирования с линейными ограничениями.
Применить метод множителей Лагранжа и теорему Каруша-Куна-Таккера $\implies$ получить задачу оптимизации без ограничений $\implies$ посчитать производную, приравнять 0 $\implies$ profit.

Метод максимальной энтропии

Закодируем все наблюдения o_t с помощью бинарных предикатов: $b_t^i : O \to \{0, 1\}$, где $i = 1..B$.
Определим функции $f_t^i : O \times S \to \{0, 1\}$ как

$$f_t^{i,s}(o_t, s_t) = \begin{cases} 1, & \text{если } b_t^i(o_t) = 1 \text{ и } s = s_t \\ 0, & \text{иначе} \end{cases}$$

Введём следующие ограничения:

$$\mathop{\mathbb{E}}_{(o_t, s_t) \sim Z} [f_t^{i,s}(o_t, s_t)] = \frac{1}{n} \sum_{t=1}^{n} f_t^{i,s}(o_t, s_t)$$

т.е. матожидание предиката в искомом распределении должно быть равно его выборочному среднему.

Метод максимальной энтропии

Итоговая задача оптимизации

$$\begin{aligned} \underset{p}{\text{maximize}} \quad & H(p) \\ \text{subject to} \quad & E^{i,s} = F^{i,s} \end{aligned}$$

где

$$H(p) = -\frac{1}{n} \sum_{t=1}^{n} \sum_{s \in S} p(s|o_t) \log p(s|o_t)$$

$$E^{i,s} = \sum_{t=1}^{n} \sum_{s' \in S} p(o_t, s') f_t^{i,s}(o_t, s')$$

$$F^{i,s} = \frac{1}{n} \sum_{t=1}^{n} f_t^{i,s}(o_t, s_t)$$

Выписываем функцию Лагранжа:

$$L(p, \lambda) = H(p) + \sum_{i,s} \lambda^{i,s}(F^{i,s} - E^{i,s})$$

Находим решение:

$$(\hat{p}, \hat{\lambda}) = \underset{p,\lambda}{\text{argmax}}\ L(p, \lambda)$$

Решение существует и единственно (Della Pietra and Lafferty, 1997)

$$p_\lambda(s|o) = \frac{1}{Z_\lambda(o)} \exp\left(\sum_{i,s,t} \lambda^{i,s} f_t^{i,s}(o, s)\right)$$

где $Z_\lambda(o)$ – нормировочная константа, определяемая условием $\sum_{s \in S} p_\lambda(s|o) = 1$, а λ – параметры для обучения.

Обучение (поиск параметров)

Как искать λ?
Функции $\lambda^{i,s}(\cdot)$ гладкие и выпуклые. Можно применять почти любой численный метод!
Есть метод, специально придуманный для этой задачи: Generalised Iterative Scaling (GIS). Он применим, если все функции f неотрицательны: $f_t^{i,s}(o_t, s_t) \geq 0$.
GIS:

- Задать начальное приближение для λ.
- Повторять до сходимости:

$$\lambda_{(j+1)}^{i,s} = \lambda_{(j)}^{i,s} + \frac{1}{\max\limits_{o,s} \sum\limits_{t,i,s} f_t^{i,s}(o,s)} \log \frac{E^{i,s}}{F^{i,s}}$$

Улучшения

- Можно обучаться даже если ответов S нет (unsupervised) с помощью EM-алгоритма.
- E-шаг: можно посчитать вероятность ответа, поэтому находим наиболее вероятную последовательность ответов и вычисляем $F^{i,s}$.
- M-шаг: Generalised Iterative Scaling.

Е-шаг: Алгоритм Витерби

А как быстро находить вероятности ответов p, имея λ?
Ответ: алгоритм Витерби.
Идея: простое динамическое программирование. Сложность: $O(n|S|^2)$.
$\alpha_t(s)$ – вероятность быть в вершине s во время t при условии $o_1, ..., o_t$.

$$\alpha_{t+1}(s) = \sum_{s' \in S} \alpha_t(s') P_{s'}(s|o_{t+1})$$

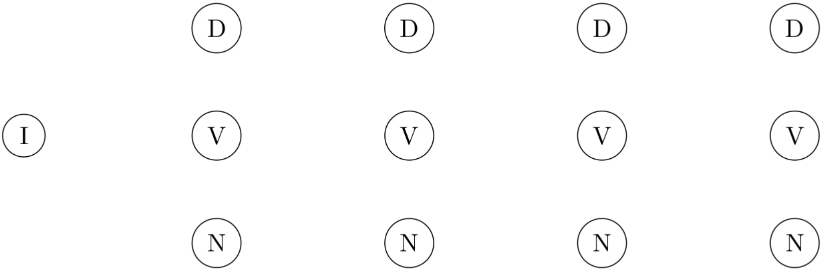
$\delta_t(s)$ – вероятность лучшего пути, который доходит до s во время t при условии $o_1, ..., o_t$.

$$\delta_{t+1}(s) = \max_{s' \in S} \delta_t(s') P_{s'}(s|o_{t+1})$$

Алгоритм Витерби

“Matt saw the cat”

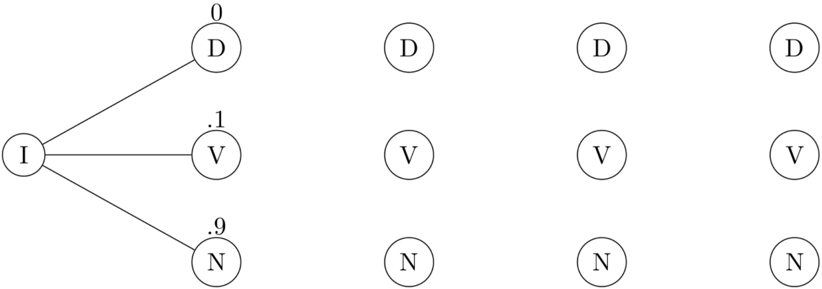
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Алгоритм Витерби

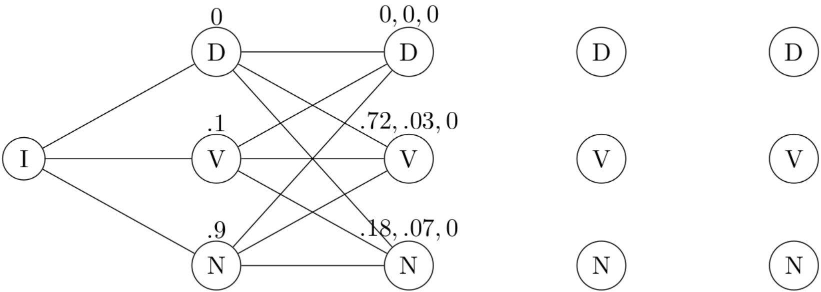
"*Matt* saw the cat"

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



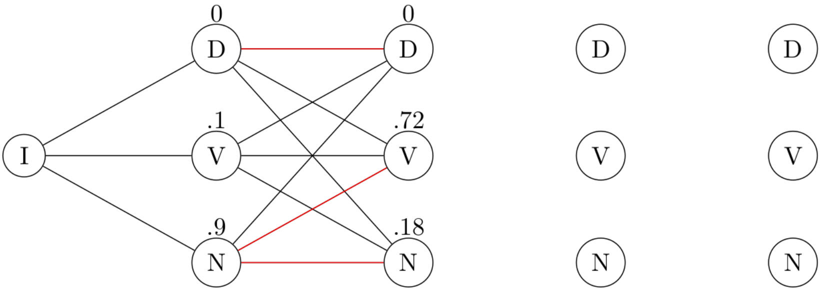
"Matt *saw* the cat"

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



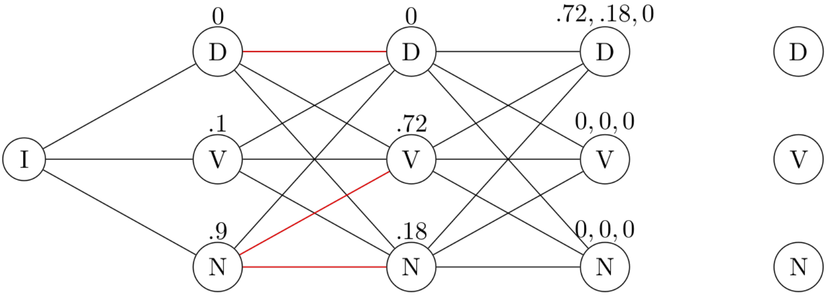
"Matt *saw* the cat"

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



“Matt saw *the* cat”

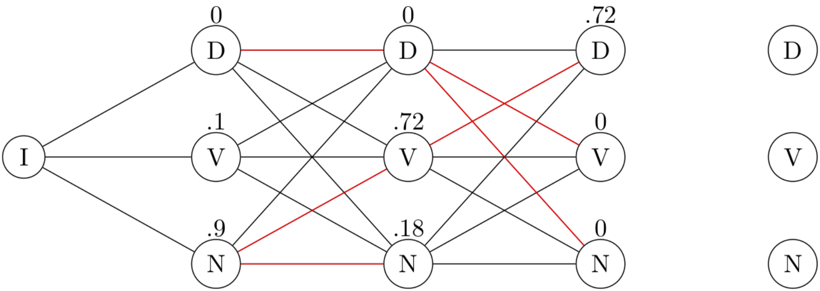
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Алгоритм Витерби

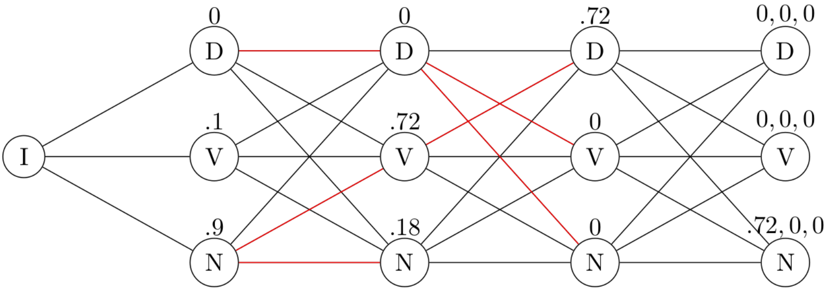
“Matt saw *the* cat”

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



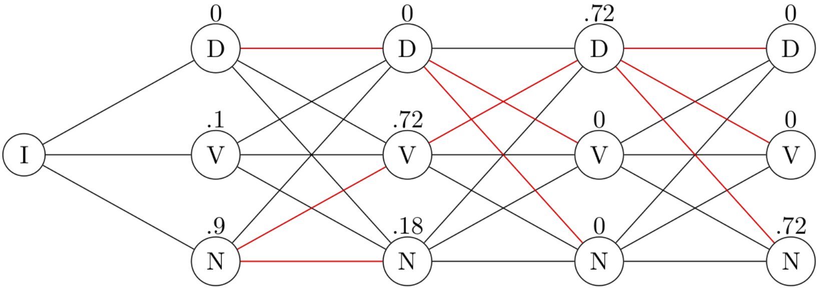
"Matt saw the *cat*"

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



"Matt saw the *cat*"

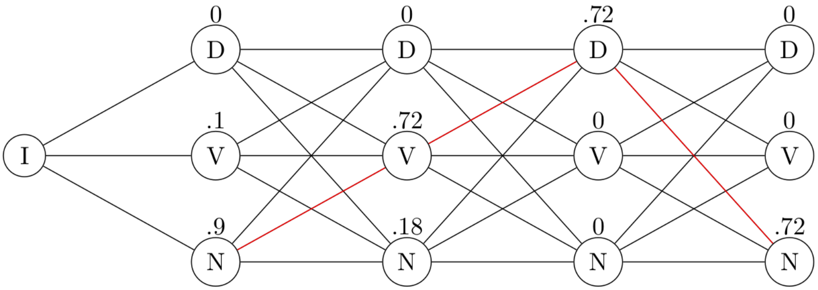
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Алгоритм Витерби

“Matt saw the cat”

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Резюме

Если ответы S даны (**обучение с учителем**), просто применяем алгоритм **Generalized Iterative Scaling**.

.

Если ответов нет (**обучение без учителя**), можно применить ЕМ-алгоритм:

- Задать начальное приближение для λ.
- Е-шаг: вычислить вероятности ответов алгоритмом Витерби, используя $p_\lambda(s|o)$; посчитать $F^{i,s}$.
- М-шаг: обновить переходные вероятности алгоритмом GIS.