

Лекция 6.

Скрытые марковские модели

Скрытые марковские модели (СММ), спецификация которых была опубликована еще в конце 60-х годов, в последнее время стали очень популярны. Во-первых, математическая структура СММ очень богата и позволяет решать математические проблемы различных областей науки. Во-вторых, грамотно спроектированная модель дает на практике хорошие результаты работы. В этом руководстве мы рассмотрим скрытые марковские модели и их применение в отдельных аспектах распознавания речи.

Введение

Происходящие явления можно описывать как сигналы. Сигналы могут быть дискретными, как письменная речь, или непрерывными, как фонограмма или кардиограмма. Сигналы с постоянными статистическими свойствами называются стабильными (стационарными), а с меняющимися — нестабильными (нестационарными). Сигнал может быть чистым, а может и искаженным, с помехами или посторонними сигналами.

Для описания сигналов часто нужны математические модели. В модели сигнала на основе его характеристик может быть предусмотрен определенный механизм обработки, который позволяет получить желаемый выход при анализе сигнала. Например, если надо облагородить сигнал, искаженный и зашумлённый при передаче, мы можем смоделировать его и рассмотреть эту модель отвлечённо от искажений и шумов в сигнале. Модели позволяют также генерировать и исследовать сигнал без его источника. В этом случае, имея под рукой хорошую модель, мы можем имитировать сигнал и изучить его по этой имитации.

Модели очень успешно применяются на практике, позволяя создавать эффективные рабочие системы: системы прогноза, распознавания, идентификации. Грубо все модели можно разделить на детерминистические и статистические. Детерминистические используются, если известны фундаментальные характеристики сигнала: что сигнал — это синусоидальная волна или, например, сумма экспонент. В таком случае достаточно просто описать подобную модель сигнала — для этого нужно всего лишь подобрать (вычислить) параметры этой модели: для синусоидальной волны — это амплитуда, частота, фаза. Второй класс — это статистические модели, которые, в соответствии со своим названием, базируются на статистических характеристиках сигнала. Эти модели описывают Гауссовы, пуассоновские, марковские процессы, а также подобные им процессы. В общем, статистические модели описывают сигнал как определённый случайный процесс, параметры которого могут быть качественно определены.

В области распознавания речи используются оба типа моделей, но в этом руководстве мы обсудим только одну, статистическую модель, а именно — скрытую марковскую модель (СММ).

Теория скрытых марковских моделей не нова. Её основы опубликовал Баум и его коллеги в конце 60-х — начале 70-х годов. Тогда же, в начале 70-х, Бейкер и Джелинек с коллегами из IBM применили СММ в распознавании речи.

Тем не менее, широкое распространение СММ получили совсем недавно:

- основы теории скрытых марковских моделей были опубликованы в журналах для математиков, не очень популярных среди инженеров, занимающихся распознаванием речи;
- опубликованная теория не содержала соответствующих обучающих материалов, которые бы объяснили возможности и способы применения СММ в различных прикладных областях.

В результате несколько вышедших подробных обучающих материалов о скрытых марковских моделях инициировали исследования по их применению в области распознавания речи..

Марковские процессы

Рассмотрим систему, которую в любой момент времени можно описать одним из N состояний, S_1, S_2, \dots, S_N , (рис. 1), где для простоты $N = 5$.

Через определенный промежуток времени система может изменить свое состояние или остаться в прежнем состоянии согласно вероятностям, указанным для данных состояний. Моменты времени, когда мы регистрируем состояние системы, обозначим как $t = 1, 2, \dots$, а состояние в момент времени t обозначим q_t . Полное описание рассмотренной выше системы должно содержать текущее состояние (в момент времени t) и последовательность всех предыдущих состояний, через которые прошла система.

В отдельных случаях описание системы сводится к указанию текущего и предыдущего состояния, то есть

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots] = P[q_t = S_j | q_{t-1} = S_i] \quad (1)$$

Кроме того, мы также полагаем что процессы, протекающие в системе, не зависят от времени, о чем нам говорит правая часть формулы (1). Таким образом, систему можно описать множеством вероятностей a_{ij} в виде

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i], \quad 1 \leq i, j \leq N \quad (2)$$

где a_{ij} — это вероятность перехода из состояния S_i в состояние S_j в данный момент времени. Поскольку эти вероятности характеризуют случайный процесс, они имеют обычные свойства, то есть

$$a_{ij} \geq 0 \quad (3 - 1)$$

$$\sum_{j=1}^N a_{ij} = 1 \quad (3 - 2)$$

Описанный выше случайный процесс можно назвать открытой марковской моделью, поскольку выходной сигнал модели — это последовательность состояний, регистрируемых во времени. Каждое состояние соответствует определённому (наблюдаемому) событию.

Для того, чтобы лучше понять все вышесказанное, рассмотрим простую марковскую модель погоды, у которой будет всего три состояния. Предполагается, что мы один раз в день (например, в полдень) смотрим в окно и регистрируем в журнале текущее состояние погоды. Мы условились, что лишь одно из трех ниженазванных состояний в день t мы записываем в журнал:

- Состояние № 1: дождь (или снег)
- Состояние № 2: пасмурно
- Состояние № 3: ясно

Матрица вероятностей изменения погоды A имеет вид

$$A = \{a_{ij}\} = \begin{vmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{vmatrix}$$

Исходя из того, что погода в первый день ($t = 1$) ясная (состояние 3), мы можем задать себе вопрос: какова вероятность (согласно нашей модели), что следующие 7 дней будет именно «ясно — ясно — дождь — дождь — ясно — пасмурно — ясно»? Точнее сказать, для данной последовательности состояний O , где $O = \{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$ соответствует $t = 1, 2, \dots, 8$, мы хотим на основе данной модели определить вероятность наблюдения последовательности O .

Эта вероятность может быть выражена (и вычислена) следующим образом

$$P(O | Model) = P[S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3 | Model] =$$

$$P[S_3] \cdot P[S_3|S_3] \cdot P[S_3|S_3] \cdot P[S_1|S_3] \cdot P[S_1|S_1] \cdot P[S_3|S_1] \cdot P[S_2|S_3] \cdot P[S_3|S_2] =$$

$$\pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23} =$$

$$1 \cdot (0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2) = 1,536 \times 10^{-4},$$

где

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \quad (4),$$

это вероятность того, что начальное состояние системы будет S_i .

Есть и другой интересный вопрос, ответ на который нам даст эта модель: какова вероятность того, что модель сохранит свое состояние в течение ровно d дней? Эта вероятность может быть вычислена как вероятность наблюдения следующей последовательности

$$O = \left\{ \begin{matrix} S_i, & S_i, & S_i, & \dots, & S_i, & S_j & \neq S_i \\ 1 & 2 & 3 & & d & d+1 \end{matrix} \right\}$$

дает модель, в которой

$$P(O|Model, q_1 = S_i) = (a_{ii})^{d-1}(1 - a_{ii}) = p_i(d). \quad (5)$$

Величина $p_i(d)$ — это вероятность того, что система будет находиться в состоянии i ровно d раз подряд. Соответственно есть функция распределения вероятности для продолжительности пребывания системы в одном состоянии, которая является характеристикой сохранения состояния для марковской цепи. Зная величины $p_i(d)$ мы можем вычислить среднее время, в течение которого система сохранит свое состояние (используем формулу математического ожидания):

$$\bar{d}_i = \sum_{d=1}^{\infty} d \cdot p_i(d) \quad (6-1)$$

$$= \sum_{d=1}^{\infty} d(a_{ii})^{d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}} \quad (6-2)$$

Ожидается, что солнечная погода скорее всего простоит $\frac{1}{0.2} = 5$ дней, пасмурная — 2,5 дня, а вот дождливая погода, согласно нашей модели, вероятнее всего продержится 1,67 дня.

Переход к скрытым марковским моделям

В вышеописанной марковской модели каждому физическому явлению соответствовало определенное состояние модели. Эта модель, к сожалению, слишком ограничена и ей не под силу решение многих актуальных проблем. В этом разделе мы рассмотрим марковские модели, в которых наблюдаемая последовательность — это результат переходов в соответствии с обозначенными вероятностями. В данном случае модель (скрытая марковская модель) — это результат двух случайных процессов. Первый — скрытый процесс — его никак нельзя зарегистрировать, но его можно охарактеризовать с помощью другого случайного процесса, который предоставляет нам набор сигналов — наблюдаемую последовательность. Проиллюстрируем это описание на примере подбрасывания монеты.

Пример подбрасываемой монеты. Действуем по следующему сценарию. Вы находитесь в комнате, а за перегородкой — в другой комнате — находится человек, который подбрасывает монету. Он не говорит, как именно он подбрасывает монету, а может он её вообще ленится подбрасывать. Он лишь говорит вам результат каждого падения монеты: орел или решка. В этом и заключается суть скрытого процесса (вы не знаете что происходит с монетой), когда о процессе вы можете судить лишь по наблюдаемой последовательности

$O = O_1 O_2 O_3 \dots O_T = \mathcal{H} \mathcal{H} \mathcal{T} \mathcal{T} \mathcal{T} \mathcal{H} \mathcal{T} \mathcal{T} \mathcal{H} \dots \mathcal{H}$, где \mathcal{H} — это орел, а \mathcal{T} — это решка.

Как же построить скрытую марковскую модель, соответствующую этой ситуации? Первый вопрос: сколько состояний будет у модели и что означает каждое состояние такой модели? Предположим, что мы подбрасываем одну единственную монету и других у нас нет. Тогда выбор мы остановим на модели с двумя состояниями, где одно состояние означает выпадение орла, другое — решки.

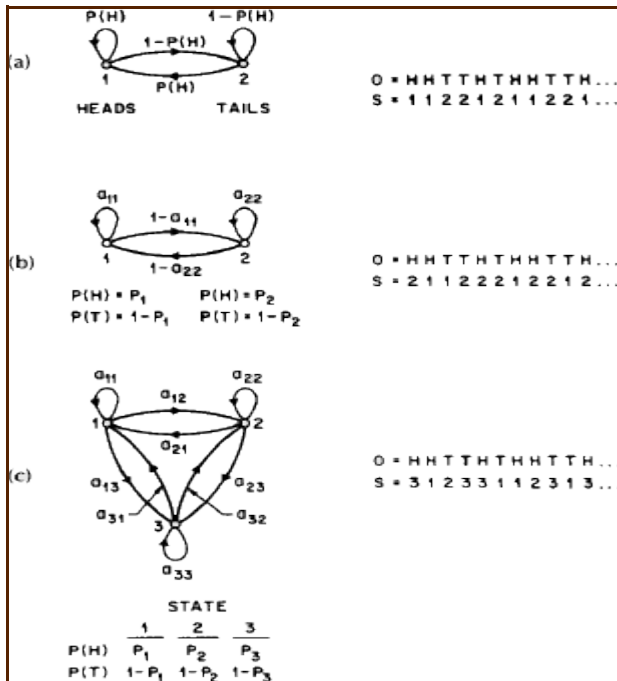


Рис. 2. Три примерных марковских модели, которые могут описать эксперимент со скрыто подбрасываемой монетой. (а) 1 монета участвует в подбрасывании, (2) 2 — монеты, (3) — три монеты.

Эта модель изображена на рисунке 2(а). В этом случае марковская модель является *открытой* и единственное, что мы можем сделать с этой моделью — это оптимизировать вероятность смены состояния. Следует заметить, что скрытая марковская модель, являющаяся аналогом модели, изображенной на рис. 2(а), будет представлять собой модель одного состояния. В этой модели единственное состояние означает, что подбрасывается всего лишь одна монета.

На следующем рисунке 2(б) изображена СММ двух состояний. В этом случае каждое состояние соответствует различным монетам, которые подбрасываются в ходе эксперимента (напр. 1 копейка и 5 рублей). Каждому состоянию соответствует *распределение вероятностей* между выпадением орла и решки, а также *матрицей вероятностей переходов* (**матрицей переходов**), указывающей вероятность перехода из одного состояния в другое. Переход из состояния в состояние согласно заданным вероятностям из *матрицы переходов* может осуществляется на основе того же подбрасывания монеты или на основе любого другого случайного события.

На третьем рисунке 2(в) представлена модель, учитывающая тот факт, что подбрасываются три различных монеты, причем выбор между ними осуществляется опять же на основе какого-либо случайного события.

Здесь, как и каждый раз при проектировании мы задаемся вопросом: какая из трех моделей наилучшим образом подходит для описания наблюдаемой последовательности? Хорошо видно, что первая модель (рис. 2(а)) имеет всего лишь 1 неизвестный параметр. Модель для двух монет (рис. 2(б)) имеет 4 неизвестных параметра. И наконец, модель для трех монет (рис. 3(в)) имеет 9 неизвестных параметров. Таким образом, СММ с большим количеством степеней свободы по существу более работоспособна, чем ее меньшие аналоги. Также теоретически доказано (и это мы увидим далее), что в современных условиях существуют ограничения на размер моделей. Более того, может оказаться, что в случае, когда человек за стеной подбрасывает одну единственную монету, мы выберем модель трех состояний. В таком случае выясняется, что состояния системы не соответствуют реальным состояниям за стеной; и, следовательно, мы используем избыточную модель.

Пример шариков в вазах. Сейчас мы дополним СММ новыми структурными элементами, для того чтобы она могла решать ряд более сложных задач. Поможет нам в этом пример с шариками в вазах (рис. 3).

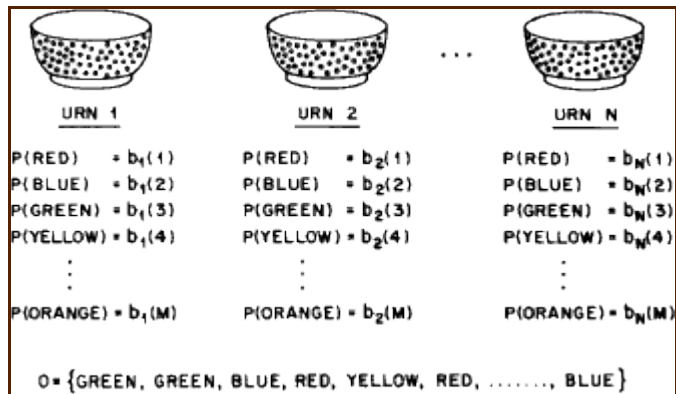


Рис. 3. Модель с N состояниями (вазами) и шариками, цвета которых обозначают элементы наблюдаемой последовательности.

Допустим, у нас есть N стеклянных прозрачных ваз. В каждой вазе — большое число шариков разного цвета. Полагаем, что у нас в корзине лежат шарики M различных цветов. Физически это можно представить следующим образом. Человек находится в комнате с вазами. Каким-либо случайным образом он выбирает любую вазу, засовывает руку поглубже, и вытаскивает шар. Цвет шара записывается в журнал показаний — наблюдаемую последовательность, и человек кладет шар обратно в эту вазу. Потом наш человек выбирает новую корзину, идет к ней, и вытаскивает оттуда новый шар, и так далее. В результате мы получаем последовательность цветов — результат работы СММ — наблюдаемую последовательность.

Очевидно, что пример шариков в вазах соответствует скрытой марковской модели, где каждое состояние модели — это выбранная ваза, причем у различных ваз различная вероятность вытащить шарик красного (или другого) цвета, что соответствует различному распределению вероятностей для каждого состояния. То, какая ваза будет выбрана следующей, зависит от матрицы переходов СММ, то есть зависит и от того, у какой вазы мы сейчас находимся.

Элементы скрытой марковской модели

Приведенные выше примеры дают неплохое представление о СММ, и о возможных сферах их применения. Сейчас мы дадим формальное определение элементам СММ и объясним, как модель генерирует наблюдаемую последовательность. СММ определяется следующими элементами:

1. N — общее количество *состояний* в модели. Несмотря на то что состояния в СММ являются скрытыми, во многих случаях есть соответствие между состоянием модели и реальным состоянием процесса. В примере с подбрасыванием монеты каждое состояние соответствовало выбранной монете, а в примере с шариками в вазах состояние модели соответствовало выбранной вазе. В общем, переход в любое выбранное состояние возможен из любого состояния всей системы (в том числе и само в себя); с другой стороны, и это мы увидим впоследствии, лишь определенные пути переходов представляют интерес в каждой конкретной модели. Мы обозначим совокупность состояний модели множеством $S = \{S_1, S_2, \dots, S_N\}$, а текущее состояние в момент времени t как q_t .

2. M , количество возможных *символов* в наблюдаемой последовательности, размер *алфавита* наблюдаемой последовательности. В случае с подбрасыванием монеты — это 2 символа: орел и решка; в случае с шариками — это количество цветов этих самых шариков. *Алфавит* наблюдаемой последовательности мы обозначим как $V = \{v_1, v_2, \dots, v_M\}$.

3. Матрица вероятностей переходов (или матрица переходов) $A = \{a_{ij}\}$, где

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \leq i, j \leq N, \quad (7)$$

то есть это вероятность того, что система, находящаяся в состоянии S_i , перейдет в состояние S_j . Если для любых двух состояний в модели возможен переход из одного состояния в другое, то $a_{ij} > 0$ для любых i, j . В остальных СММ для некоторых i, j у нас вероятность перехода $a_{ij} = 0$.

4. Распределение вероятностей появления символов в j -том состоянии, $B = \{b_j(k)\}$, где

$$b_j(k) = P[v_k | q_t = S_j] \quad 1 \leq j \leq N, \quad 1 \leq k \leq M. \quad (8)$$

$b_j(k)$ — вероятность того, что в момент времени t , система, находящаяся в j -ом состоянии (состояние S_j), выдаст k -тый символ (символ v_k) в наблюдаемую последовательность.

5. Распределение вероятностей начального состояния $\pi = \{\pi_i\}$, где

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N, \quad (9)$$

то есть вероятность того, S_i — это начальное состояние модели.

Совокупность значений N, M, A, B и π — это скрытая марковская модель, которая может сгенерировать *наблюдаемую последовательность*

$$O = O_1 O_2 \dots O_T \quad (10)$$

(где O_t — один из символов алфавита V , а T — это количество элементов в наблюдаемой последовательности).

СММ строит наблюдаемую последовательность по следующему алгоритму

1. Выбираем начальное состояние $q_1 = S_i$ в соответствии с распределением π
2. Устанавливаем $t = 1$.
3. Выбираем $O_t = v_k$ в соответствии с распределением $b_j(k)$ в состоянии (S_i) .
4. Переводим модель в новое состояние $q_{t+1} = S_j$ в соответствии с матрицей переходов a_{ij} с учетом текущего состояния S_i .
5. Устанавливаем время $t = t + 1$; возвращаемся к шагу 3, если $t < T$; иначе — заканчиваем выполнение.

Подводя итог, заметим, что *полное* описание СММ состоит из двух параметров модели (N и M), описания символов наблюдаемой последовательности и трех массивов вероятностей — A, B , и π . Для удобства мы используем следующую запись

$$\lambda = (A, B, \pi) \quad (11)$$

для обозначения *достаточного* описания параметров модели.

Три основных задачи СММ

Согласно описанию скрытой марковской модели, изложенному в предыдущем разделе, существует три основных задачи, которые должны быть решены для того, чтобы модель могла успешно решать поставленные перед ней задачи.

Задача 1

Дано: наблюдаемая последовательность $O = O_1, O_2, \dots, O_T$ и модель $\lambda = (A, B, \pi)$. Необходимо вычислить вероятность $P(O | \lambda)$ — вероятность того, что данная наблюдаемая последовательность построена именно для данной модели.

Задача 2

Дано: наблюдаемая последовательность $O = O_1, O_2, \dots, O_T$ и модель $\lambda = (A, B, \pi)$. Необходимо подобрать последовательность состояний системы $Q = q_1, q_2, \dots, q_T$, которая лучше всего соответствует наблюдаемой последовательности, то есть «объясняет» наблюдаемую последовательность.

Задача 3

Подобрать параметры модели $\lambda = (A, B, \pi)$ таким образом, чтобы максимизировать $P(O | \lambda)$.

Задача 1 — это задача оценки модели, которая заключается в вычислении вероятности того, что модель соответствует заданной наблюдаемой последовательности. К сути этой задачи можно подойти и с другой стороны: насколько выбранная СММ соответствует заданной наблюдаемой последовательности. Такой подход имеет большую практическую ценность. Например, если у нас стоит вопрос выбора наилучшей модели из набора уже существующих, то решение первой задачи дает нам ответ на этот вопрос.

Задача 2 — это задача, в которой мы пытаемся понять, что же происходит в скрытой части модели, то есть найти «правильную» последовательность, которую проходит модель. Совершенно ясно, что абсолютно точно нельзя определить эту последовательность. Здесь можно говорить лишь о предположениях с соответственной степенью достоверности. Тем не менее для приближенного решения этой проблемы мы обычно будем пользоваться некоторыми оптимальными показателями, критериями. Далее мы увидим, что, к сожалению, не существует единого критерия оценки для определения последовательности состояний. При решении второй задачи необходимо каждый раз принимать решение о том, какие показатели использовать. Данные, полученные при решении этой задачи используются для изучения поведения построенной модели, нахождения оптимальной последовательности её состояний, для статистики и т. п.

Решение задачи 3 состоит в оптимизации модели таким образом, чтобы она как можно лучше описывала *реальную* наблюдаемую последовательность. Наблюдаемая последовательность, по которой оптимизируется СММ, принято называть обучающей последовательностью, поскольку с помощью нее мы «обучаем» модель. Задача обучения СММ — это важнейшая задача для большинства проектируемых СММ, поскольку она заключается в оптимизации параметров СММ на основе обучающей наблюдаемой последовательности, то есть создается модель, наилучшим образом описывающая реальные процессы.

Для лучшего понимания рассмотрим все вышесказанное на примере системы, предназначенной для распознавания речи. Для каждого слова из словаря W мы спроектируем СММ с N состояниями. Каждое слово в частности мы представим как последовательность спектральных векторов. Обучение мы будем считать завершенным, когда модель с высокой точностью будет воспроизводить ту самую последовательность спектральных векторов, которая использовалась для обучения модели. Таким образом каждая отдельная СММ будет обучаться воспроизводить какое-либо одно слово, но обучать эту модель следует на нескольких вариантах произнесения этого слова; то есть например три человека (каждый по-своему) проговаривают слово «собака», а затем каждое сказанное слово конвертируется в упорядоченный по времени набор спектральных векторов, и модель обучается на основе этих трех наборов. Для каждого отдельного слова проектируются соответствующие модели. Сперва решается 3-я задача СММ: каждая модель настраивается на «произнесение» определенного слова из словаря W , согласно заданной точности.

Для того чтобы интерпретировать каждое состояние спроектированных моделей мы решаем 2-ую задачу, а затем выделяем те свойства спектральных векторов, которые имеют наибольший вес для определенного состояния. Это момент тонкой настройки модели. А уже после того, как набор моделей будет спроектирован, оптимизирован и обучен, следует оценить модель на предмет ее способности распознавать слова в реальной жизни. Здесь мы уже решаем 1-ую задачу CMM. Нам дается тестовое слово, представленное, разумеется, в виде наблюдаемой последовательности спектральных векторов. Далее мы вычисляем функцию соответствия этого тестового слова для каждой модели. Модель, для которой эта функция будет иметь наибольшее значение, будет считаться моделью названного слова.

В следующем разделе мы дадим четкое формальное решение трем задачам CMM.

Решение трех задач СММ

Решение 1-ой задачи

Нам необходимо вычислить вероятность того, что последовательность наблюдений $O = O_1, O_2, \dots, O_t$ принадлежит модели λ , то есть вычислить $P(O | \lambda)$. В первую очередь в голову приходит решение подсчитать вероятность появления последовательности наблюдений для каждой возможной последовательности состояний модели. Рассмотрим такой способ на примере одной последовательности состояний

$$Q = q_1, q_2, \dots, q_T \quad (12)$$

где q_1 — это начальное состояние модели.

Вероятность появления последовательности наблюдений O для последовательности состояний (12) равна

$$P(O|Q, \lambda) = \prod_{t=1}^T P(O_t|q_t, \lambda) \quad (13 - 1)$$

где мы подразумеваем статистическую независимость наблюдений. Отсюда получаем

$$P(O|Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdot \dots \cdot b_{q_T}(O_T) \quad (13 - 2)$$

Вероятность, что в модели состояния пройдут последовательность Q равна

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdot \dots \cdot a_{q_{T-1} q_T} \quad (14)$$

Вероятность совмещения O и Q , то есть вероятность одновременного их проявления, выражается произведением

$$P(O, Q|\lambda) = P(O|Q, \lambda)P(Q, \lambda) \quad (15)$$

Вероятность появления O — это сумма вероятностей совмещения (15) по всем возможным комбинациям состояний состояний q системы:

$$P(O|\lambda) = \sum_{all Q} P(O|Q, \lambda) P(Q|\lambda) \quad (16)$$

$$= \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \dots a_{q_{T-1} q_T} b_{q_T}(O_T) \quad (17)$$

Объяснить это можно так. Сперва (в момент времени $t = 1$) мы выбираем начальное состояние q_1 в соответствии с вероятностью π_{q_1} , и генерируем символ O_1 (в этом состоянии) с вероятностью $b_{q_1}(O_1)$. Далее переходим к следующему моменту времени $t + 1$ ($t = 2$) и выполняем переход в состояние q_2 с вероятностью $a_{q_1 q_2}$; после чего генерируем символ O_2 с вероятностью $b_{q_2}(O_2)$. Этот процесс повторяется, пока мы не достигнем времени $t = T$. В конце мы переведем систему из состояния q_{T-1} в q_T с вероятностью $a_{q_{T-1} q_T}$ и сгенерируем символ O_T с вероятностью $b_{q_T}(O_T)$.

Следует отметить, что прямое вычисление $P(O | \lambda)$ по формуле (17) требует произвести порядка $2T \cdot N^T$ вычислений, поскольку для каждого времени $t = 1, 2, \dots, T$ существует N возможных состояний системы, то есть N^T возможных вариантов последовательности состояний; и для каждого варианта около $2T$ вычислений — для каждого слагаемого суммы в формуле (17). Для абсолютной точности скажем, что нам необходимо произвести $(2T - 1)N^T$ умножений и $N^T - 1$ сложений. Подобные вычисления невыполнимы даже для малых значений N и T ; то есть для $N = 5$ (состояний), $T = 100$ (наблюдений) количество вычислений будет порядка $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$! Совершенно ясно, что для решения 1-ой задачи СММ требуется гораздо более эффективный алгоритм. К счастью существуют даже два таких алгоритма и называются они алгоритм прямого хода и алгоритм обратного хода.

Алгоритмы прямого и обратного хода. Введем прямую переменную $\alpha_t(i)$ и определим ее как

$\alpha_t(i) = P(O_1 O_2 \dots O_t, q_t = S_i | \lambda)$ (18) то есть вероятность того что для заданной модели λ к моменту времени t наблюдалась последовательность $O_1 O_2 \dots O_t$, и в момент t система находится в состоянии S_i . Значение $\alpha_t(i)$ мы можем найти методом индукции по следующему алгоритму:

1) Инициализация:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N. \quad (19)$$

2) Индукция:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), \quad 1 \leq t \leq T-1. \quad 1 \leq j \leq N$$

3) Завершение:

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i). \quad (21)$$

На шаге 1) подсчитываются вероятности совмещения состояния S_j и первого наблюдения O_1 . Индукция является центральной частью вычисления; её схема показана на рис. 4(а). На этой схеме видно, каким путем система в момент времени $t + 1$ приходит в состояние S_j из N возможных состояний предыдущего момента времени. Поскольку $\alpha_t(i)$ — совмещенная вероятность проявления наблюдений $O_1 O_2 \dots O_t$ и нахождения системы в состоянии S_i в момент времени t .