

SIR-модели и вариационные приближения

Сергей Николенко

Академия MADE — Mail.Ru

1 июня 2020 г.

Random facts:

- 1 июня 1495 г. датировано первое письменное упоминание о шотландском виски в казначейских отчётах Шотландии; винокурком был монах Джон Кор (John Cor)
- 1 июня 1816 г. доктор Хосе Гаспар Родригес де Франсия, «Великий правитель» и «Отец парагвайской нации», стал Постоянным диктатором Республики
- 1 июня 1938 г. за авторством Джерри Сигела (автор) и Джо Шустера (художник) вышел первый номер журнала *Action Comics*, в котором впервые появился Супермен
- 1 июня 1967 г. вышел *Sgt. Pepper's Lonely Hearts Club Band*, лучший альбом всех времён по версии журнала *Rolling Stone*
- 1 июня 1970 г. чехословацкий Ту-104 упал близ Триполи, 1 июня 1976 г. советский Ту-154 потерпел катастрофу в Экваториальной Гвинее, а 1 июня 2009 г. произошла крупнейшая в истории *Air France* катастрофа над Атлантическим океаном
- 1 июня 2010 г. в Катманду наследный принц Дипендра расстрелял всю свою семью, в том числе короля Бирендру, и застрелился сам

SIR-модели в эпидемиологии

- Прежде чем двигаться дальше — конкретный (и весьма актуальный) пример
- Давайте попробуем применить то, о чём мы говорили, к эпидемиологии
- В модели SIR есть:
 - объекты (люди) $X = \{x_1, \dots, x_N\}$,
 - каждый эволюционирует между тремя состояниями $\mathcal{S} = \{S, I, R\}^N$;
 - S, I, R — ещё общее число объектов в соответствующих состояниях;
 - входные данные — число зарегистрированных случаев заболевания, изменяющееся во времени: $\mathbf{y} = \left(y^{(t)}\right)_{t=1}^T$.

- Введём для каждого объекта *траекторию* (subject-path)
 $\mathbf{x}_j = \left(x_j^{(t)}\right)_{t=1}^T, j = 1, \dots, N.$
- Тогда и общие статистики изменяются во времени: $S^{(t)}, I^{(t)}, R^{(t)}$.
- Неизвестные параметры модели — это $\boldsymbol{\theta} = \{\beta, \mu, \rho, \boldsymbol{\pi}\}$:
 - $\boldsymbol{\pi}$ — начальное распределение заболевших, $x_j^{(1)} \sim \boldsymbol{\pi}$;
 - ρ — вероятность обнаружить инфицированного в общей популяции, то есть вероятность того, что человек x_j в момент t , когда $x_j^{(t)} = I$, будет обнаружен тестированием и зачислен в данные $y^{(t)}$; тогда $y_t | I^{(t)}, \rho \sim \text{Binom}(I^{(t)}, \rho)$;
 - μ — вероятность для заболевшего выздороветь, то есть вероятность перехода из состояния I в состояние R ;
 - β — самый интересный параметр, вероятность заразиться за один отсчёт времени *от одного инфицированного человека*; будем предполагать самую простую модель, в которой вероятность заразиться от одного инфицированного равна β и все эти события независимы, а значит, вероятность остаться здоровым равна $(1 - \beta)^{I^{(t)}}$.

- Обозначим вектор состояний всех людей, кроме x_j , через \mathbf{x}_{-j} (и остальные величины так же).
- Вероятности перехода из $x_j^{(t-1)}$ в $x_j^{(t)}$:

$$\begin{aligned} p(x_j^{(t)} = S | x_j^{(t-1)} = S, \mathbf{x}_{-j}^{(t-1)}) &= (1 - \beta)^{I_{-j}^{(t-1)}}, \\ p(x_j^{(t)} = I | x_j^{(t-1)} = S, \mathbf{x}_{-j}^{(t-1)}) &= 1 - (1 - \beta)^{I_{-j}^{(t-1)}}, \\ p(x_j^{(t)} = R | x_j^{(t-1)} = I, \mathbf{x}_{-j}^{(t-1)}) &= \mu, \\ p(x_j^{(t)} = I | x_j^{(t-1)} = I, \mathbf{x}_{-j}^{(t-1)}) &= 1 - \mu, \\ p(x_j^{(t)} | x_j^{(t-1)}, \mathbf{x}_{-j}^{(t-1)}) &= 0 \quad \text{во всех остальных случаях.} \end{aligned}$$

- Скрытые переменные — те же самые траектории \mathbf{x} (не зря же мы их вводили).

- Тогда полное правдоподобие $\mathcal{L}(X, Y | \theta)$ получается как

$$\begin{aligned}\mathcal{L}(X, Y | \theta) &= p(Y | X, \rho) p(X^{(1)} | \pi) p(X | X^{(1)}, \beta, \mu) \\ &= \left[\prod_{t=1}^T \binom{I^{(t)}}{Y^{(t)}} \rho^{Y^{(t)}} (1 - \rho)^{I^{(t)} - Y^{(t)}} \right] \times \\ &\quad \times \left[\pi_S^{S^{(1)}} \pi_I^{I^{(1)}} \pi_R^{R^{(1)}} \right] \cdot \left[\prod_{t=2}^T \prod_{j=1}^N p(x_j^t | x_{-j}^{t-1}, \theta) \right],\end{aligned}$$

где $p(x_j^t | x_{-j}^{t-1}, \theta)$ определено матрицей вероятностей переходов.

- Апостериорное распределение, которое нам нужно:

$$p(\theta | Y) \propto p(\theta) p(Y | \theta) = \int \mathcal{L}(Y | X, \theta) p(X | \theta) p(\theta) dX,$$

и этот интеграл, конечно, никак не подсчитать. Что же делать?

- На помощь приходит алгоритм Метрополиса-Гастингса, точнее, сэмплирование по Гиббсу.
- Будем сэмплировать траектории \mathbf{x}_j последовательно, зафиксировав все остальные \mathbf{x}_{-j} , данные \mathbf{y} и параметры модели $\boldsymbol{\theta}$:

$$\mathbf{x}_j \sim p(\mathbf{x}_j | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}).$$

- Для этого нужно сначала понять, как выглядит распределение на траектории \mathbf{x}_j .
- Очевидно, её элементы $x_j^{(t)}$ нельзя считать независимыми, ведь человек проходит цепочку состояний $S \rightarrow I \rightarrow R$ только один раз и слева направо (если проходит вообще). Всё это на первый взгляд опять выглядит сложно...

- ...но здесь получается модель, которая нам уже хорошо знакома: последовательность случайных переменных $x_j^{(t)}$ образует марковскую цепь, а если добавить ещё известные нам данные, то получится скрытая марковская модель.
- Выбросим x_j из множества траекторий, получив статистики по всей остальной популяции $S_{-j}^{(t)}$, $I_{-j}^{(t)}$ и $R_{-j}^{(t)}$. Тогда параметры скрытой марковской модели таковы:
 - скрытые состояния $x_j^{(t)}$ с множеством возможных значений $\{S, I, R\}$;
 - матрица вероятностей перехода $p(x_j^t | x_{-j}^{t-1}, \theta)$, определённая выше;
 - наблюдаемые y , вероятности получить которые зависят от того, заражён ли человек x_j в момент времени t :

$$p(y^{(t)} | x_j^{(t)}) = \text{Binom} \left(I_{-j}^{(t)} + [x_j^{(t)} = I], \rho \right).$$

- Чтобы сэмплировать одну траекторию x_j при условии фиксированных остальных траекторий x_{-j} , нужно сэмплировать траекторию вдоль скрытых состояний марковской модели.
- Здесь x_j будет эволюционировать от состояния S к состоянию R последовательно, с вероятностями перехода x_j на каждом шаге от S к R

$$p(x_j^{(t)} = I | x_j^{(t-1)} = S, x_{-j}) = 1 - (1 - \beta)^{I_{-j}^{(t-1)}},$$

а вероятность перехода от I к R фиксирована и равна μ .

- Стохастический алгоритм Витерби: два прохода по НММ слева направо и справа налево.
- На прямом проходе подсчитываем матрицы совместных вероятностей пар последовательных состояний

$$Q_j^{(t)} = \left(q_{j,s',s}^t \right)_{s',s \in \{S,I,R\}}, \quad \text{где}$$

$$q_{j,s',s}^t = p(x_j^{(t)} = s, x_j^{(t-1)} = s' | Y, \mathbf{x}_{-j}, \boldsymbol{\theta}).$$

- Фактически в нашей модели возможных пар таких состояний всего шесть (остальные переходы запрещены), и все матрицы Q выглядят как

$$Q_j^{(t)} = \begin{pmatrix} q_{j,S,S}^{(t)} & q_{j,S,I}^{(t)} & 0 \\ 0 & q_{j,I,I}^{(t)} & q_{j,I,R}^{(t)} \\ 0 & 0 & q_{j,R,R}^{(t)} \end{pmatrix}.$$

- Чтобы вычислить $q_{j,s',s}^{(t)}$, нужно подсчитать

$$\begin{aligned}
 q_{j,s',s}^{(t)} &= p(x_j^{(t)} = s, x_j^{(t-1)} = s' | \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}) \\
 &\propto p(x_j^{(t-1)} = s' | \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}) p(x_j^{(t)} = s | x_j^{(t-1)} = s' \\
 &\quad = s', \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}) p(y_t | x_j^{(t)} = s, \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}) = \\
 &= \left[\sum_{s''} q_{j,s'',s'}^{(t-1)} \right] \cdot p(x_j^{(t)} = s | x_j^{(t-1)} = s', \mathbf{x}_{-j}, \boldsymbol{\theta}) \times \\
 &\quad \times p_{\text{Binom}} \left(y^{(t)} \mid I_{-j}^{(t)} + [x_j^{(t)} = I], \rho \right),
 \end{aligned}$$

где $p(x_j^{(t)} = s | x_j^{(t-1)} = s', \mathbf{x}_{-j}, \boldsymbol{\theta})$ — это те самые вероятности перехода в нашей модели, подсчитанные по статистикам $S_{-j}^{(t-1)}$, $I_{-j}^{(t-1)}$ и $R_{-j}^{(t-1)}$, а p_{Binom} — вероятность по биномиальному распределению.

- Потом нужно нормировать, учитывая, что $\sum_{s,s'} q_{j,s',s}^{(t)} = 1$.

- Когда все матрицы $Q_j^{(t)}$ подсчитаны, их можно использовать для того, чтобы сэмплировать целые последовательности скрытых состояний. Для этого нужно разложить $p(\mathbf{x}_j | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta})$ не с начала времён, а с конца:

$$p(\mathbf{x}_j | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) = p(x_j^{(T)} | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) p(x_j^{(T-1)} | x_j^{(T)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) \times \dots \\ \dots \times p(x_j^{(2)} | x_j^{(3)}, \dots, x_j^{(T)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) p(x_j^{(1)} | x_j^{(2)}, \dots, x_j^{(T)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}).$$

- И можно сэмплировать справа налево по матрицам Q .

- Последнее состояние сэмплируется из сумм по строкам последней матрицы $Q_j^{(T)}$:

$$\begin{aligned}x_j^{(T)} \sim p(x_j^{(T)} = s | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) &= \sum_{s'} p(x_j^{(T)} = s, x_j^{(T-1)} = s' | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) = \\ &= \sum_{s'} q_{j,s',s}^{(T)}.\end{aligned}$$

- А дальше достаточно, по марковскому свойству последовательности \mathbf{x}_j , сэмплировать при условии следующего состояния, то есть использовать распределение

$$\begin{aligned}x_j^{(t)} \sim p(x_j^{(t)} = s | x_j^{(t+1)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) &\propto \\ \propto p(x_j^{(t)} = s, x_j^{(t+1)} = s' | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) &= q_{j,s,s'}^{(t+1)}.\end{aligned}$$

- Так мы получим новую траекторию \mathbf{x}_j , и её можно подставить в X на место старой траектории и продолжать процесс сэмплирования: выбрать новый индекс j и повторить всё заново.
- В какой-то момент надо будет остановиться и обновить значения параметров.
- Теоретически можно даже сделать полноценный байесовский вывод, пересчитав параметры сопряжённых априорных распределений.
- Три основных параметра β , ρ и μ — это три монетки, а оставшийся параметр π — кубик с тремя гранями. Поэтому сопряжёнными априорными распределениями будут

$$\begin{aligned} p(\beta) &= \text{Beta}(a_\beta, b_\beta), & p(\mu) &= \text{Beta}(a_\mu, b_\mu), \\ p(\rho) &= \text{Beta}(a_\rho, b_\rho), & p(\pi) &= \text{Dir}(\mathbf{a}_\pi). \end{aligned}$$

- Чтобы пересчитать их апостериорные значения, нужно аналогично обычным HMM подсчитать «статистику» того, сколько раз соответствующие монетки и кубики «бросали» и чем они «выпадали» в текущем наборе скрытых переменных (траекторий) X :
 - к параметрам \mathbf{a}_π добавляются статистики того, в каких состояниях начинаются траектории:

$$a_{\pi,s} := a_{\pi,s} + \sum_{j=1}^N \mathbb{1}[X_j^{(1)} = s];$$

- Чтобы пересчитать их апостериорные значения, нужно аналогично обычным НММ подсчитать «статистику» того, сколько раз соответствующие монетки и кубики «бросали» и чем они «выпадали» в текущем наборе скрытых переменных (траекторий) X :
 - параметры a_μ и b_μ обновляются в зависимости от того, каково было ожидаемое число переходов из состояния I в состояние R (выздоровлений) и сколько всего времени люди провели в состоянии I (проболели):

$$a_\mu := a_\mu + \sum_{t=1}^{T-1} \sum_{j=1}^N [X_j^{(t)} = I, X_j^{(t+1)} = R],$$

$$b_\mu := b_\mu + \sum_{t=1}^T I^{(t)} - \sum_{t=1}^{T-1} \sum_{j=1}^N [X_j^{(t)} = I, X_j^{(t+1)} = R].$$

- Чтобы пересчитать их апостериорные значения, нужно аналогично обычным НММ подсчитать «статистику» того, сколько раз соответствующие монетки и кубики «бросали» и чем они «выпадали» в текущем наборе скрытых переменных (траекторий) X :
 - аналогично, параметры a_ρ и b_ρ получаются из статистики выявленных случаев, попавших в y , по сравнению со случаями, которые оказались только в $I^{(t)}$:

$$a_\rho := a_\rho + \sum_{t=1}^T y^{(t)}, \quad b_\rho := b_\rho + \sum_{t=1}^T (I^{(t)} - y^{(t)});$$

- Параметры a_β и b_β самые интересные: нужно подсчитать ожидаемое число «возможностей заразиться», которые реализовались и не реализовались для всех людей в популяции:

$$p(x_j \text{ заразился при одном контакте} | x_j \text{ заразился}) = \frac{\beta}{1 - (1 - \beta)^{I^{(t)}}},$$

а значит,

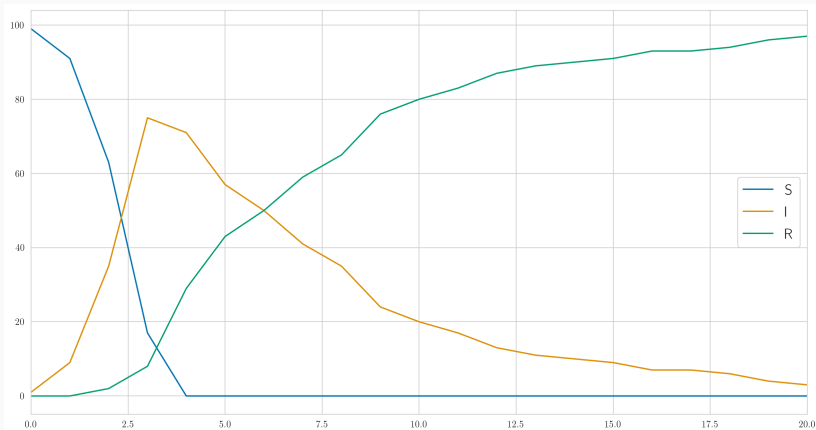
$$a_\beta := a_\beta + \sum_{t,j: x_j^{(t)}=S, x_j^{(t+1)}=I} \frac{\beta I^{(t)}}{1 - (1 - \beta)^{I^{(t)}}},$$

$$b_\beta := b_\beta + \sum_{t,j: x_j^{(t)}=S, x_j^{(t+1)}=S} I^{(t)} + \sum_{t,j: x_j^{(t)}=S, x_j^{(t+1)}=I} \left(I^{(t)} - \frac{\beta I^{(t)}}{1 - (1 - \beta)^{I^{(t)}}} \right).$$

- Итого получили все компоненты нашей (сильно упрощённой!) SIR-модели: скрытые переменные в виде траекторий элементов популяции, алгоритм для сэмплирования по Гиббсу, который сэмплирует одну траекторию при условии всех остальных, и правила обновления параметров, которыми можно воспользоваться после того, как марковская цепь сэмплирования достаточно долго поработала.
- Давайте теперь посмотрим на практику...

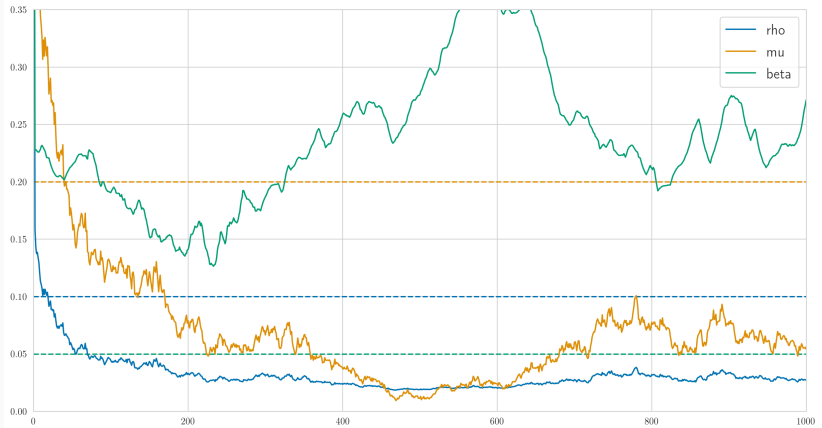
SIR-модели

- Пример визуализации статистик заражения при параметрах $N = 100$, $T = 20$, $\rho = 0.1$, $\beta = 0.05$, $\mu = 0.1$:



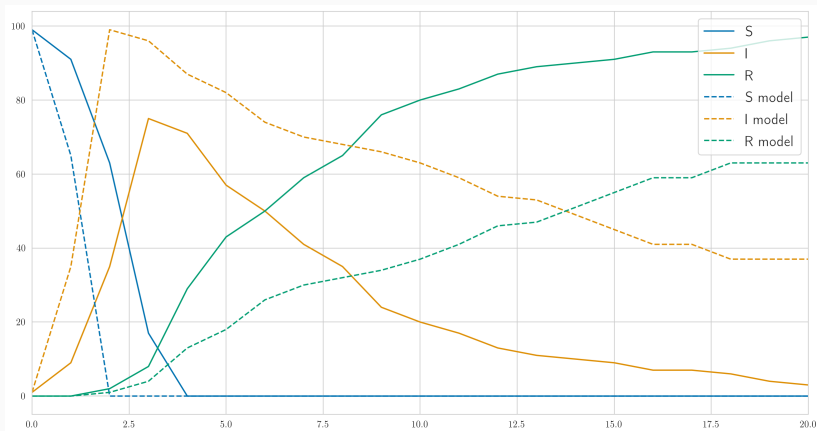
SIR-модели

- Пример обучения параметров модели SIR:



SIR-модели

- И если посэмплировать популяции из полученных параметров и из настоящих, получится совсем одно и то же:



- Какие выводы? Как это использовать на практике?

Спасибо!

Спасибо за внимание!