MCMC и коронавирус: модель SIR

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург

23 мая 2020 г.

Random facts:

- 23 мая — Всемирный день черепахи, учреждённый в 2000 году по инициативе Американского общества спасения черепах
- 23 мая 1430 г. бургундцы при Компьене захватили Жанну д'Арк, а 23 мая 1498 г. во Флоренции сожгли Савонаролу
- 23 мая 1873 г. в Москве состоялась премьера оперы Римского-Корсакова «Снегурочка», а 23 мая 1969 г. *The Who* выпустили первую в истории рок-оперу *Tommy*
- 23 мая 1980 г. вышел фильм *The Shining*, 23 мая 1994 г. *Pulp Fiction* завоевал «Золотую пальмовую ветвь», а 23 мая 2000 г. Бьорк получила в Каннах звание лучшей актрисы за дебютную роль в *Dancer in the Dark*
- 23 мая 1988 г. Мишель Платини попрощался с большим футболом
- 23 мая 1995 г. вышла первая версия языка программирования *Java*

SIR-модели в эпидемиологии

SIR-модели

- 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) $X = \{x_1, \ldots, x_N\}$,
    - 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: $\mathbf{y} = \left(y^{(t)}\right)_{t=1}^{T}$.

SIR-модели

- Введём для каждого объекта *траекторию* (subject-path) $\mathbf{x}_j = \left(x_j^{(t)}\right)_{t=1}^{T}$, $j = 1, \ldots, N$.
- 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 $y_t \mid I^{(t)}, \rho \sim \mathrm{Binom}(I^{(t)}, \rho)$;
    - 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 $\mathcal{L}(X, Y \mid \boldsymbol{\theta})$ получается как

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

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

- Апостериорное распределение, которое нам нужно:

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

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

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

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

- Для этого нужно сначала понять, как выглядит распределение на траектории x_j.
- Очевидно, её элементы $x_j^{(t)}$ нельзя считать независимыми, ведь человек проходит цепочку состояний $S \to I \to R$ только один раз и слева направо (если проходит вовсе). 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 $\mathbf{x}_j$ 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 $p(x_j^t | \mathbf{x}_{-j}^{t-1}, \boldsymbol{\theta})$, определённая выше;
    - наблюдаемые $\mathbf{y}$, вероятности получить которые зависят от того, заражён ли человек x_j в момент времени t:

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

SIR-модели

- 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

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

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

SIR-модели

- Стохастический алгоритм Витерби: два прохода по HMM 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 $q^{(t)}_{j,s',s}$, нужно подсчитать

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

  где $p(x^{(t)}_j = s | x^{(t-1)}_j = s', \mathbf{x}_{-j}, \boldsymbol{\theta})$ — это те самые вероятности перехода в нашей модели, подсчитанные по статистикам $S^{(t-1)}_{-j}$, $I^{(t-1)}_{-j}$ и $R^{(t-1)}_{-j}$, а p_{Binom} — вероятность по биномиальному распределению.
- Потом нужно нормировать, учитывая, что $\sum_{s,s'} q^{(t)}_{j,s',s} = 1$.

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

SIR-модели

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

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

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

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

- 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 $\boldsymbol{\pi}$ — кубик с тремя гранями. Поэтому сопряжёнными априорными распределениями будут

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

- 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 $\mathbf{a}_{\boldsymbol{\pi}}$ добавляются статистики того, в каких состояниях начинаются траектории:

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

SIR-модели

- 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:
    - 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):

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

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

- 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:
  - аналогично, параметры a_ρ и b_ρ получаются из статистики выявленных случаев, попавших в $\mathbf{y}$, по сравнению со случаями, которые оказались только в $I^{(t)}$:

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

- 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,

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

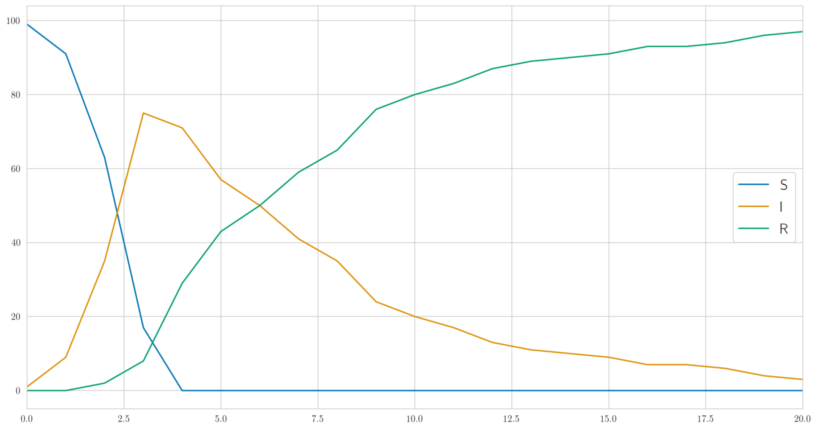
$$b_\beta := b_\beta + \sum_{t,j:\ x_j^{(t)}=S, x_j^{(t+1)}=S} I^{(t)} + \sum_{t,j:\ x_j^{(t)}=S, x_j^{(t+1)}=I} \left(I^{(t)} - \frac{\beta I^{(t)}}{1-(1-\beta)^{I^{(t)}}} \right).$$

SIR-модели

- 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...

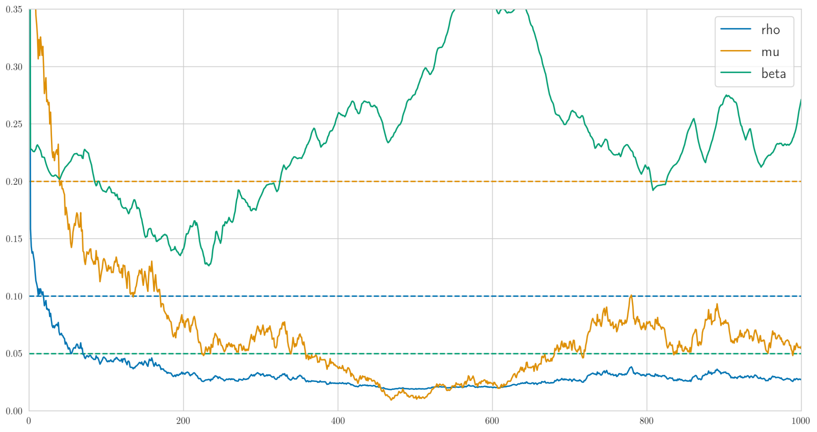
SIR-модели

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



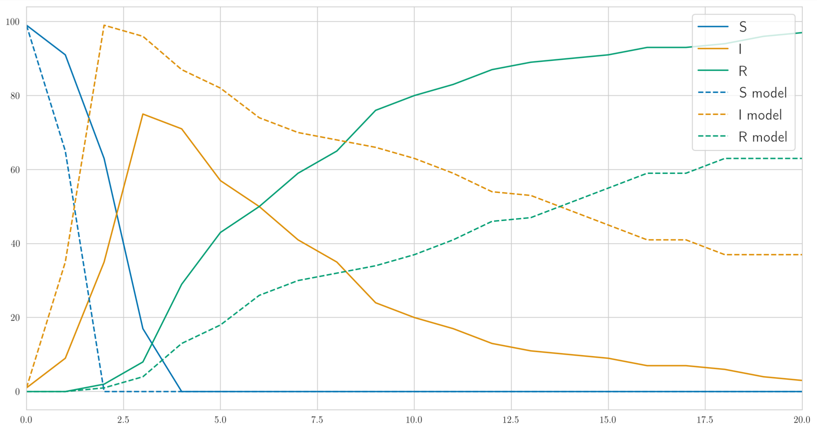
SIR-модели

- 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!