Формулы для обучения и инференса моделей

Если $(x,\, t_x,\, n)$ - история активности пользователя, то можно выписать его правдоподобие по p и θ в контексте **BdW / BB** модели:

$$
\begin{aligned}
&L(p,\theta,c|x,t_x,n) = p(x,t_x,n|p,\theta,c) = \\
&= p^x(1-p)^{n-x}(1-\theta)^{n^c} + \sum_{i=0}^{n-t_x-1} p^x(1-p)^{t_x-x+i}\big((1-\theta)^{(t_x+i)^c} - (1-\theta)^{(t_x+i+1)^c}\big) \qquad (1)
\end{aligned}
$$

Тогда $L(w,c|x,t_x,n)$ получается взятием ожидания (1) по q_w,
где w-параметры распределения $(p,\theta) \sim q_w(p,\theta)$:

$$
\begin{aligned}
L(w,c|x,t_x,n) = p(x,tx,n|w,c) &= \int p(x,t_x,n,p,\theta|w,c)dpd\theta = \\
&= \int p(x,t_x,n|p,\theta,c)p(p,\theta|w)dpd\theta = \\
&= \int L(p,\theta,c|x,t_x,n)q_w(p,\theta)dpd\theta \qquad (2)
\end{aligned}
$$

В случае **BdW/BB** модели, где $w = (\alpha,\beta,\gamma,\delta)$, а $q_w(p,\theta) = B(p|\alpha,\beta)B(\theta|\gamma,\delta)$, (2) выражается явно:

$$
\begin{aligned}
&L(\alpha,\beta,\gamma,\delta,c|x,t_x,n) = \frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\frac{B(\gamma,\delta+n^c)}{B(\gamma,\delta)} + \\
&+ \sum_{i=0}^{n-t_x-1} \frac{B(\alpha+x,\beta+t_x-x+i)}{B(\alpha,\beta)}\Big(\frac{B(\gamma,\delta+(t_x+i)^c) - B(\gamma,\delta+(t_x+i+1)^c)}{B(\gamma,\delta)}\Big)
\end{aligned}
$$

В случае **S**$_{BB}$**-G/B** модели и аналогичных у нас нет замкнутой формулы для вычисления, но мы можем считать правдоподобие методом Монте-Карло.

Пусть R_t - событие возвращения пользователя на t-ый период ($t > n$), а A_t - пользователь жив в t-ый период. Тогда

$$
\begin{aligned}
p(R_t|x,t_x,n,p,\theta,c) &= p(R_t|p,\theta,c)\frac{p(x,t_x,n|R_t,p,\theta,c)}{p(x,t_x,n|p,\theta,c)} \\
&= p(R_t|p,\theta,c)\frac{p(x,t_x,n|A_n,p)}{L(p,\theta,c|x,t_x,n)} \qquad (3)
\end{aligned}
$$

(второе равенство следует из свойств вероятностного процесса индивидуального пользователя) где

$$p(R_t|p,\theta,c) = p(1-\theta)^{t^c}$$
$$p(x,t_x,n|A_n,p) = p^x(1-p)^{n-x}$$

Тогда, чтобы получить $p(R_t|x,t_x,n,w,c)$ нужно взять ожидание (3) по

$$q(p,\theta|x,t_x,n,w,c) = \frac{p(x,t_x,n|p,\theta,c)}{p(x,t_x,n|w,c)} q_w(p,\theta) =$$
$$= \frac{L(p,\theta,c|x,t_x,n)}{L(w,c|x,t_x,n)} q_w(p,\theta)$$

Итого получаем формулу:

$$p(R_t|x,t_x,n,w,c) =$$
$$= L(w,c|x,t_x,n)^{-1} \int p(R_t|p,\theta,c) p(x,t_x,n|A_n,p) q_w(p,\theta) dp d\theta$$

В случае **BdW/BB** модели получаем:

$$p(R_t|x,t_x,n,\alpha,\beta,\gamma,\delta,c) =$$
$$= L(\alpha,\beta,\gamma,\delta,c|x,t_x,n)^{-1} \frac{B(\alpha+x+1,\beta+n-x)}{B(\alpha,\beta)} \frac{B(\gamma,\delta+t^c)}{B(\gamma,\delta)}$$

Опять же в случае **S**$_{BB}$**-G/B** модели и аналогичных у нас нет замкнутой формулы для вычисления, но мы можем считать эту вероятность методом Монте-Карло.

Также имеем в общем виде и в случае для **BdW/BB** модели формулу Retention:

$$p(R_t|w,c) = \int p(1-\theta)^{t^c} q_w(p,\theta) dp d\theta =$$
$$= \frac{\alpha}{\alpha+\beta} \frac{B(\gamma,\delta+t^c)}{B(\gamma,\delta)}$$