ВАРИАНТЫ И ПРИЛОЖЕНИЯ ЕМ-АЛГОРИТМА

Сергей Николенко

СПбГУ — Санкт-Петербург

04 марта 2025 г.

Random facts:

- 4 марта 1628 г. королевскую хартию получила колония Массачусетского залива, 4 марта 1681 г. Уильям Пенн получил хартию на будущую Пенсильванию, а уже 4 марта 1789 г. первое заседание Конгресса США в Нью-Йорке ввело в действие конституцию США
- 4 марта 1733 г. был издан указ Анны Иоанновны «Об учреждении полиции в городах», согласно которому в 23 городах России были созданы полицмейстерские конторы
- 4 марта 1762 г. император Пётр III подписал указ «О свободной для всех торговле»
- 4 марта 1803 г. император Александр I издал «Указ о вольных хлебопашцах», по которому землевладельцы могли освобождать крестьян, наделяя их землёй
- 4 марта 1837 г. Михаил Лермонтов был арестован за стихотворение «Смерть поэта»
- 4 марта 1990 г. на 5-летний срок был избран Съезд народных депутатов РСФСР, высший законодательный орган РСФСР; он был распущен указом Бориса Ельцина 21 сентября 1993 г. и разогнан вооружённой силой 4 октября 1993 г.
- 4 марта 2012 г. прошли выборы президента Российской Федерации; Владимир Путин был в третий раз избран президентом России

Свойства и расширения EM

Требования к X и Y

- Что требуется, чтобы EM-алгоритм работал?
- Неформально нужно, чтобы $p(X \mid \theta)$ было легко максимизировать.
- А формально нужно, чтобы $p(\mathbf{y} \mid \mathbf{x}, \theta) = p(\mathbf{y} \mid \mathbf{x})$, т.е. чтобы выполнялось марковское свойство для $\theta \to \mathbf{x} \to \mathbf{y}$.
- На самом деле обычно EM применяется тогда, когда $\mathbf{y} = f(\mathbf{x})$ для детерминированной функции f, и это свойство тривиально выполняется.
- Более того, обычно EM применяется, когда f — это просто проекция, т.е. когда $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, как мы изначально и рассматривали.

Разложение Q-функции по точкам

- Важное простое свойство — если данные состоят из независимо порождённых $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ (как всегда и бывает), то

$$Q(\theta \mid \theta^{(m)}) = \mathbb{E}_{X|Y,\theta^{(m)}} \left[\log \prod_{n=1}^{N} p(\mathbf{x}_n \mid \theta) \right] =$$

$$= \mathbb{E}_{X|Y,\theta^{(m)}} \left[\sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \theta) \right] = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}_n|\mathbf{y}_n,\theta^{(m)}} \left[\log p(\mathbf{x}_n \mid \theta) \right],$$

  потому что $p(\mathbf{x}_n \mid Y, \theta) = p(\mathbf{x}_n \mid \mathbf{y}_n, \theta)$

- Упражнение: докажите это!

Generalized EM

- Другие обобщения могут пригодиться, если всё-таки подсчитать $\mathbb{E}_{X|Y,\theta^{(m)}}[\log p(X \mid \theta)]$ или оптимизировать $p(X \mid \theta)$ нелегко.
- Обобщённый EM (Generalized EM, GEM): вместо $\arg\max_\theta Q(\theta, \theta^{(m)})$ нам достаточно просто выбирать такую $\theta^{(m+1)}$, чтобы

$$Q\left(\theta^{(m+1)}, \theta^{(m)}\right) > Q\left(\theta^{(m)}, \theta^{(m)}\right).$$

Stochastic EM

- Стохастический EM (Stochastic EM): если Q-функцию не получается посчитать в замкнутой форме, но и просто максимум брать не хочется, как в Classification EM, можно попробовать брать $\mathbf{x}$ случайным образом на E-шаге:

$$\mathbf{x}^{(m)} \sim p(\mathbf{x} \mid \mathbf{y}, \theta^{(m)}),$$

а потом использовать его на M-шаге как обычно:

$$\theta^{(m+1)} = \arg\max_{\theta} p(\mathbf{x}^{(m)} \mid \theta^{(m)}).$$

Monte Carlo EM

- Монте-Карло EM (Monte Carlo EM): саму Q-функцию тоже можно попытаться приблизить, если подсчитать сложно; можно использовать приближение ожидания в Q-функции через сэмплирование:

$$Q(\theta \mid \theta^{(m)}) \approx \frac{1}{R} \sum_{r=1}^{R} \log p(\mathbf{x}^{(m,r)} \mid \theta), \text{ где } \mathbf{x}^{(m,r)} \sim p(\mathbf{x} \mid \mathbf{y}, \theta^{(m)}).$$

- Впрочем, в таких случаях часто можно вообще забить на EM и аппроксимировать напрямую апостериорное распределение.

EM с априорным распределением

- А можно и априорное распределение в EM добавить, конечно же:

$$\theta_{MAP} = \arg\max_\theta \log p(\theta \mid Z) = \arg\max_\theta \left(\log p(Y \mid \theta) + \log p(\theta)\right).$$

- При этом базовая схема особо не меняется:

$$Q(\theta \mid \theta^{(m)}) = \mathbb{E}_{X|Y,\theta^{(m)}} \left[\log p(X \mid \theta)\right],$$
$$\theta^{(m+1)} = \arg\max_\theta \left(Q(\theta \mid \theta^{(m)}) + \log p(\theta)\right).$$

- Например, так можно избежать вырожденных случаев (кластер из одной точки).

Semi-supervised clustering

- И EM, и k–means хорошо обобщаются на случай частично обученных кластеров.
- То есть про часть точек уже известно, какому кластеру они принадлежат.
- Как это учесть?

Semi-supervised clustering

- И EM, и k–means хорошо обобщаются на случай частично обученных кластеров.
- То есть про часть точек уже известно, какому кластеру они принадлежат.
- Как это учесть?
- Чтобы учесть информацию о точке $\mathbf{x}_i$, достаточно для EM положить скрытую переменную g_{nc} равной тому кластеру, которому нужно, с вероятностью 1, а остальным — с вероятностью 0, и не пересчитывать.
- Для k–means то же самое, но для clust_i.

Пример: модели Брэдли--Терри

Модели Брэдли--Терри

- Другой подход к рейтинг–системам — модели Брэдли–Терри (Bradley–Terry).
- Модель предполагает, что для участников $1, \dots, n$ можно подобрать такие рейтинги γ_i, $i = 1..n$, что вероятность победы участника i над участником j равна

$$p(i \text{ побеждает } j) = \frac{\gamma_i}{\gamma_i + \gamma_j}.$$

- Основная задача заключается в том, чтобы найти $\gamma = (\gamma_1, \dots, \gamma_m)$ максимального правдоподобия из имеющихся данных D.

Модели Брэдли--Терри

- Если принять априорное распределение равномерным, можно просто максимизировать правдоподобие

$$p(D|\gamma) = \prod_{i=1}^{m} \prod_{j=1}^{m} \left(\frac{\gamma_i}{\gamma_i + \gamma_j} \right)^{w_{ij}},$$

где w_{ij} — то, сколько раз x_i обыграл x_j при их попарном сравнении ($w_{ii} = 0$ по определению).

- Взяв логарифм, будем максимизировать

$$l(\gamma) = \sum_{i=1}^{m} \sum_{j=1}^{m} \left(w_{ij} \log \gamma_i - w_{ij} \log(\gamma_i + \gamma_j) \right).$$

Модели Брэдли--Терри

- Максимизировать будем классическим MM-алгоритмом (minorization–maximization), фактически вариационным приближением.
- Заметим, что

$$1 + \log \frac{x}{y} - \frac{x}{y} \leq 0.$$

Упражнение. Докажите это.

Модели Брэдли--Терри

- Рассмотрим вспомогательную функцию

$$Q(\gamma, \gamma^{(k)}) = \sum_{i,j} w_{ij} \left[\log \gamma_i - \frac{\gamma_i + \gamma_j}{\gamma_i^{(k)} + \gamma_j^{(k)}} - \log \left(\gamma_i^{(k)} + \gamma_j^{(k)} \right) + 1 \right].$$

Упражнение. Используя предыдущее неравенство, докажите, что $Q(\gamma, \gamma^{(k)}) \leq l(\gamma)$.

Модели Брэдли--Терри

- Чтобы найти $\max_\gamma Q\left(\gamma, \gamma^{(k)}\right)$, можно просто взять производные $\frac{\partial Q}{\partial \gamma_l}$:

$$\frac{\partial Q}{\partial \gamma_l} = \sum_{i,j} w_{ij} \left[\frac{\delta_{il}}{\gamma_i} - \frac{\delta_{il} + \delta_{jl}}{\gamma_i^{(k)} + \gamma_j^{(k)}} \right] =$$

$$= \frac{1}{\gamma_l} \sum_j w_{lj} - \sum_j \frac{w_{lj}}{\gamma_i^{(k)} + \gamma_j^{(k)}} - \sum_i \frac{w_{il}}{\gamma_i^{(k)} + \gamma_j^{(k)}}.$$

Модели Брэдли--Терри

- Если w_l — общее количество побед игрока l ($w_l = \sum_j w_{lj}$), и N_{ij} — количество встреч между игроками i и j ($N_{ij} = w_{ij} + w_{ji}$), получаем

$$\frac{w_l}{\gamma_l} - \sum_j \frac{N_{lj}}{\gamma_i^{(k)} + \gamma_j^{(k)}} = 0.$$

- В результате правило пересчёта на одной итерации выглядит так:

$$\gamma_l^{(k+1)} := w_l \left[\sum_j \frac{N_{lj}}{\gamma_i^{(k)} + \gamma_j^{(k)}} \right]^{-1}.$$

Модели Брэдли--Терри

- Получили алгоритм оценки рейтингов. Правда, он пока работает только для ситуации, когда игроки встречаются один на один и выигрывают или проигрывают.
- Но даже в шахматах бывают ничьи. Как их учесть?

Модели Брэдли--Терри

- Если возможны ничьи, их вероятность можно описать дополнительным параметром $\theta > 1$, и это приводит к модели, в которой

$$p(i \text{ побеждает } j) = \frac{\gamma_i}{\gamma_i + \theta\gamma_j},$$
$$p(j \text{ побеждает } i) = \frac{\gamma_j}{\theta\gamma_i + \gamma_j},$$
$$p(i \text{ и } j \text{ играют вничью}) = \frac{(\theta^2 - 1)\gamma_i\gamma_j}{(\gamma_i + \theta\gamma_j)(\theta\gamma_i + \gamma_j)}.$$

Модели Брэдли--Терри

- Аналогично можно вводить другие обобщения.
- Например, если результат может зависеть от порядка элементов в паре (скажем, команды проводят «домашние» матчи и «гостевые»), можно ввести дополнительный параметр θ, характеризующий, насколько большое преимущество дают «родные стены», и рассмотреть модель с

$$p(i \text{ побеждает } j) = \begin{cases} \frac{\theta\gamma_i}{\theta\gamma_i+\gamma_j}, & \text{если } i \text{ играет дома,} \\ \frac{\gamma_j}{\theta\gamma_i+\gamma_j}, & \text{если } j \text{ играет дома.} \end{cases}$$

Модели Брэдли--Терри

- Можно даже обобщить на случай, когда в одном турнире встречаются несколько игроков: пусть перестановка π подмножества игроков $A = \{1, \dots, k\}$ (результат турнира) имеет вероятность

$$p_A(\pi) = \prod_{i=1}^{k} \frac{\gamma_{\pi(i)}}{\gamma_{\pi(i)} + \gamma_{\pi(i+1)} + \dots + \gamma_{\pi(k)}}.$$

Модели Брэдли--Терри

- Можно показать, что такая модель эквивалентна весьма естественной «аксиоме Люса»: для любой модели, в которой вероятности игроков обыграть друг друга в любой паре не равны нулю, для любых подмножеств игроков $A \subset B$ и любого игрока $i \in A$

$$p_B(i \text{ побеждает}) =$$
$$= p_A(i \text{ побеждает}) p_B(\text{побеждает кто-то из множества } A).$$

- Но для крупных турниров это перестаёт работать; и совсем трудно что-то осмысленное сделать, если игроки соревнуются не поодиночке, а в командах.

Пример EM: presence-only data

Presence-only data

- Пример из экологии: пусть мы хотим оценить, где водятся те или иные животные.
- Как определить, что суслики тут водятся, понятно: видишь суслика — значит, он есть.
- Но как определить, что суслика нет? Может быть, ты не видишь суслика, и я не вижу, а он есть?..

Presence-only data

- Формально говоря, есть переменные $\mathbf{x}$, определяющие некий регион (квадрат на карте), и мы моделируем вероятность того, что нужный вид тут есть, $p(y = 1 \mid \mathbf{x})$, при помощи логит-функции:

$$p(y = 1 \mid \mathbf{x}) = \sigma(\eta(\mathbf{x})) = \frac{1}{1 + e^{-\eta(\mathbf{x})}},$$

  где $\eta(\mathbf{x})$ может быть линейной (тогда получится логистическая регрессия), но может, в принципе, и не быть.
- Заметим, что даже если бы мы знали настоящие y, это было бы ещё не всё: сэмплирование положительных и отрицательных примеров неравномерно, перекошено в пользу положительных.
- *Важное замечание: prospective vs. retrospective studies, case-control studies.*

Presence-only data

- Это значит, что ещё есть пропорции сэмплирования (sampling rates)

$$\gamma_0 = p(s = 1 \mid y = 0), \quad \gamma_1 = p(s = 1 \mid y = 1),$$

  т.е. вероятности взять в выборку положительный/отрицательный пример.

- Их можно оценить как

$$\gamma_0 = \frac{n_0}{(1-\pi)N}, \quad \gamma_1 = \frac{n_1}{\pi N},$$

  где π — истинная доля положительных примеров (встречаемость, occurrence).

- Кстати, эту π было бы очень неплохо оценить в итоге.

Presence-only data

- Тогда, если знать истинные значения всех y, то в принципе можно обучить:

$$p(y = 1 \mid s = 1, \mathbf{x}) =$$

$$= \frac{p(s = 1 \mid y = 1, \mathbf{x})p(y = 1 \mid \mathbf{x})}{p(s = 1 \mid y = 0, \mathbf{x})p(y = 0 \mid \mathbf{x}) + p(s = 1 \mid y = 1, \mathbf{x})p(y = 1 \mid \mathbf{x})} =$$

$$= \frac{\gamma_1 e^{\eta(\mathbf{x})}}{\gamma_0 + \gamma_1 e^{\eta(\mathbf{x})}} = \frac{e^{\eta^*(\mathbf{x})}}{1 + e^{\eta^*(\mathbf{x})}},$$

где $\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log(\gamma_1/\gamma_0)$, т.е.

$$\eta^*(\mathbf{x}) = \eta(\mathbf{x}) + \log\left(\frac{n_1}{n_0}\right) - \log\left(\frac{\pi}{1-\pi}\right).$$

Presence-only data

- Таким образом, если π неизвестно, то $\eta(\mathbf{x})$ можно найти с точностью до константы.
- А у нас не n_0 и n_1, а naive presence n_p и background n_u, т.е.

$$p(y=1 \mid s=1) = \frac{n_p+\pi n_u}{n_p+n_u}, \quad p(y=1 \mid s=0) = \frac{(1-\pi)n_u}{n_p+n_u},$$
$$\gamma_1 = \frac{p(y=1|s=1)p(s=1)}{p(y=1)} = \frac{n_p+\pi n_u}{\pi(n_p+n_u)}p(s=1),$$
$$\gamma_0 = \frac{p(y=0|s=1)p(s=1)}{p(y=0)} = \frac{n_u}{n_p+n_u}p(s=1),$$

  и в нашей модели $\log \frac{n_1}{n_0} = \log \frac{n_p+\pi n_u}{\pi n_u}$, т.е. всё как раньше, но $n_1 = n_p + \pi n_u$, $n_0 = (1-\pi)n_u$.
- Обучать по y можно так: обучить модель, а потом вычесть из $\eta(\mathbf{x})$ константу $\log \frac{n_p+\pi n_u}{\pi n_u}$.

Presence-only data

- Но у нас нет настоящих данных y, чтобы обучить регрессию, а есть только presence-only z: если $z = 1$, то $y = 1$, но если $z = 0$, то неизвестно, чему равен y.
- (Ward et al., 2009): давайте использовать EM. Правдоподобие:

$$\mathcal{L}(\eta \mid \mathbf{y}, \mathbf{z}, X) = \prod_i p(y_i, z_i \mid s_i = 1, \mathbf{x}_i) = \\ = \prod_i p(y_i \mid s_i = 1, \mathbf{x}_i) p(z_i \mid y_i, s_i = 1, \mathbf{x}_i).$$

- В нашем случае при n_p положительных примеров и n_u фоновых (неизвестных)

$$\begin{aligned} p(z_i = 0 \mid y_i = 0, s_i = 1, \mathbf{x}_i) &= 1, \\ p(z_i = 1 \mid y_i = 1, s_i = 1, \mathbf{x}_i) &= \frac{n_p}{n_p + \pi n_u}, \\ p(z_i = 0 \mid y_i = 1, s_i = 1, \mathbf{x}_i) &= \frac{\pi n_u}{n_p + \pi n_u}. \end{aligned}$$

Presence-only data

- А максимизировать нам надо сложное правдоподобие, в котором значения $\mathbf{y}$ неизвестны:

$$L(\eta \mid \mathbf{z}, X) = \prod_i p(z_i \mid s_i = 1, \mathbf{x}) =$$

$$= \prod_i \left(\frac{\frac{n_p}{\pi n_u} e^{\eta(\mathbf{x}_i)}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) e^{\eta(\mathbf{x}_i)}} \right)^{z_i} \left(\frac{1 + e^{\eta(\mathbf{x}_i)}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) e^{\eta(\mathbf{x}_i)}} \right)^{1 - z_i}.$$

- Для этого и нужен EM.

Presence-only data

- E-шаг здесь в том, чтобы заменить y_i на его оценку

$$\hat{y}_i^{(k)} = \mathbb{E}\left[y_i \mid \eta^{(k)}\right] = \frac{e^{\eta^{(k)}} + 1}{1 + e^{\eta^{(k)}} + 1}.$$

- M-шаг мы уже видели, это обучение параметров логистической модели с целевой переменной $\mathbf{y}^{(k)}$ на данных X.

(1) Chose initial estimates: $\hat{y}_i^{(0)} = \pi$ for $z_i = 0$.
(2) Repeat until convergence:

- *Maximization step:*
  - Calculate $\hat{\eta}^{*(k)}$ by fitting a logistic model of $\hat{\mathbf{y}}^{(k-1)}$ given X.
  - Calculate $\hat{\eta}^{(k)} = \hat{\eta}^{*(k)} - \log\left(\frac{n_p + \pi n_u}{\pi n_u}\right)$.
- *Expectation step:*

$$\hat{y}_i^{(k)} = \frac{e^{\hat{\eta}^{(k)}}}{1 + e^{\hat{\eta}^{(k)}}} \text{ for } z_i = 0 \quad \text{and} \quad \hat{y}_i^{(k)} = 1 \text{ for } z_i = 1$$

Presence-only data

- У Ward et al. получалось хорошо, но тут вышел любопытный спор.
- Ward et al. писали так: хотелось бы, чтобы можно было оценить π, но "π is identifiable only if we make unrealistic assumptions about the structure of $\eta(\mathbf{x})$ such as in logistic regression where $\eta(\mathbf{x})$ is linear in $\mathbf{x}$: $\eta(\mathbf{x}) = \mathbf{x}^\top \beta$".
- Через пару лет вышла статья Royle et al. (2012), тоже очень цитируемая, в которой говорилось: "logistic regression... is hardly unrealistic... such models are the most common approach to modeling binary variables in ecology (and probably all of statistics)... the logistic functions... is customarily adopted and widely used, and even books have been written about it".

Presence-only data

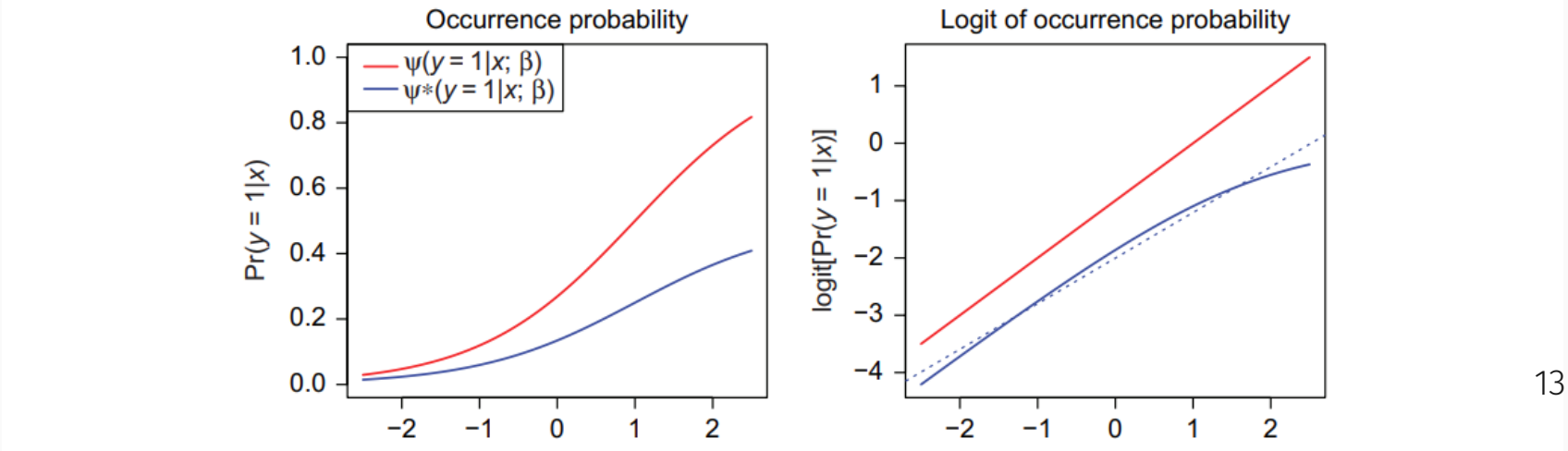
- Они предложили процедуру для оценки встречаемости π:
  - для признаков $\mathbf{x}$ у нас $p(y = 1 \mid \mathbf{x}) = \frac{p(y=1)\pi_1(\mathbf{x})}{p(y=1)\pi_1(\mathbf{x})+(1-p(y=1))\pi_0(\mathbf{x})}$;
  - данные – это выборка из $\pi_1(\mathbf{x})$ и отдельно выборка из $\pi(\mathbf{x})$;
  - как видно, даже если знать π и π_1 полностью, остаётся свобода: надо оценить $p(y = 1 \mid \mathbf{x})$, а $\pi_0(\mathbf{x})$ мы не знаем;
  - Royle et al. вводят предположения для $p(y = 1 \mid \mathbf{x})$ в виде логистической регрессии: $p(y = 1 \mid \mathbf{x}) = \sigma(\beta^\top \mathbf{x})$;
  - тогда действительно можно записать $p(y = 1)\pi_1(\mathbf{x}) = p(y = 1 \mid \mathbf{x})\pi(\mathbf{x}) = p(y = 1, \mathbf{x})$, и если $\pi(\mathbf{x})$ равномерно (это логично), то

$$\pi_1(\mathbf{x}_i) = \frac{p(y_i = 1 \mid \mathbf{x}_i)}{\sum_{\mathbf{x}} p(y = 1 \mid \mathbf{x})};$$

    если подставить сюда логистическую регрессию, то можно оценить β максимального правдоподобия, и не надо знать $p(y = 1)$!
- Что тут не так?

Presence-only data

- Всё так, но предположение о логистической регрессии здесь выполняет слишком много работы.
- Если рассмотреть две кривые, у которых $p^*(y = 1 \mid \mathbf{x}, \beta) = \frac{1}{2} p(y = 1 \mid \mathbf{x}, \beta)$, то у них будет $p^*(y = 1) = \frac{1}{2} p(y = 1)$, но общее правдоподобие $\pi_1(\mathbf{x}_i)$ будет в точности одинаковое, $\frac{1}{2}$ сократится.
- Дело в том, что модель p^* не будет логистической регрессией, с β произойдёт что-то нелинейное; но откуда у нас настолько сильное предположение? Как отличить синюю кривую справа от пунктирной прямой?



Пример: рейтинг спортивного ЧГК

Рейтинг спортивного ЧГК

- Пример из практики: в какой-то момент я хотел сделать рейтинг спортивного «Что? Где? Когда?»:
  - участвуют команды по ≤ 6 человек, причём часто встречаются неполные команды;
  - игроки постоянно переходят между командами (поэтому TrueSkill);
  - в одном турнире могут участвовать до тысячи команд (синхронные турниры);
  - командам задаётся фиксированное число вопросов $(36, 60, 90)$, т.е. в крупных турнирах очень много команд делят одно и то же место.

Рейтинг спортивного ЧГК

- Первое решение — система TrueSkill
- Расскажу о ней потом, когда будем говорить об Expectation Propagation
- У неё есть проблемы в постановке спортивного ЧГК, мы когда-то их отчасти решили, была сложная и интересная модель, она работала лучше базового TrueSkill
- Но...

Рейтинг спортивного ЧГК

- В какой-то момент база турниров ЧГК стала собирать повопросные результаты.
- Мы теперь знаем, на какие именно вопросы ответила та или иная команда.
- Так что когда я вернулся к задаче построения рейтинга ЧГК, задача стала существенно проще.

Рейтинг спортивного ЧГК

- Пример аналогичного приложения:
  - есть набор вопросов для теста из большого числа вопросов (например, IQ-тест или экзамен по какому-то предмету);
  - участники отвечают на случайное подмножество вопросов;
  - надо оценить участников, но уровень сложности вопросов нельзя заранее точно сбалансировать.
- «Что? Где? Когда?» — это оно и есть, только теперь участники объединяются в команды.

Baseline: логистическая регрессия

- Baseline — логистическая регрессия:
  - каждый игрок i моделируется скиллом s_i,
  - каждый вопрос q моделируется сложностью («лёгкостью») c_q,
  - добавим глобальное среднее μ,
  - обучим логистическую модель

$$p(x_{tq} \mid s_i, c_q) \sim \sigma(\mu + s_i + c_q)$$

для каждого игрока $i \in t$ команды-участницы $t \in T^{(d)}$ и каждого вопроса $q \in Q^{(d)}$, где $\sigma(x) = 1/(1 + e^x)$ — логистический сигмоид, x_{tq} — ответила ли команда t на вопрос q.

Модель со скрытыми переменными

- Логистическая модель предполагает фактически, что каждый игрок ответил на каждый вопрос, который взяла команда.
- Это неправда, мы не знаем, кто ответил, только знаем, что кто-то это сделал; а если команда не ответила, то никто.
- Это по идее похоже на presence-only data models (Ward et al., 2009; Royle et al., 2012).

Модель со скрытыми переменными

- Поэтому давайте сделаем модель со скрытыми переменными.
- Для каждой пары игрок-вопрос, добавим переменную z_{iq}, которая означает, что «игрок i ответил на вопрос q».
- На эти переменные есть такие ограничения:
  - если $x_{tq} = 0$, то $z_{iq} = 0$ для каждого игрока $i \in t$;
  - если $x_{tq} = 1$, то $z_{iq} = 1$ для по крайней мере одного игрока $i \in t$.

Модель со скрытыми переменными

- Параметры модели те же — скилл и сложность вопросов:

$$p(z_{iq} \mid s_i, c_q) \sim \sigma(\mu + s_i + c_q).$$

- Обучаем EM-алгоритмом:
  - E-шаг: зафиксируем все s_i и c_q, вычислим ожидания скрытых переменных z_{iq} как

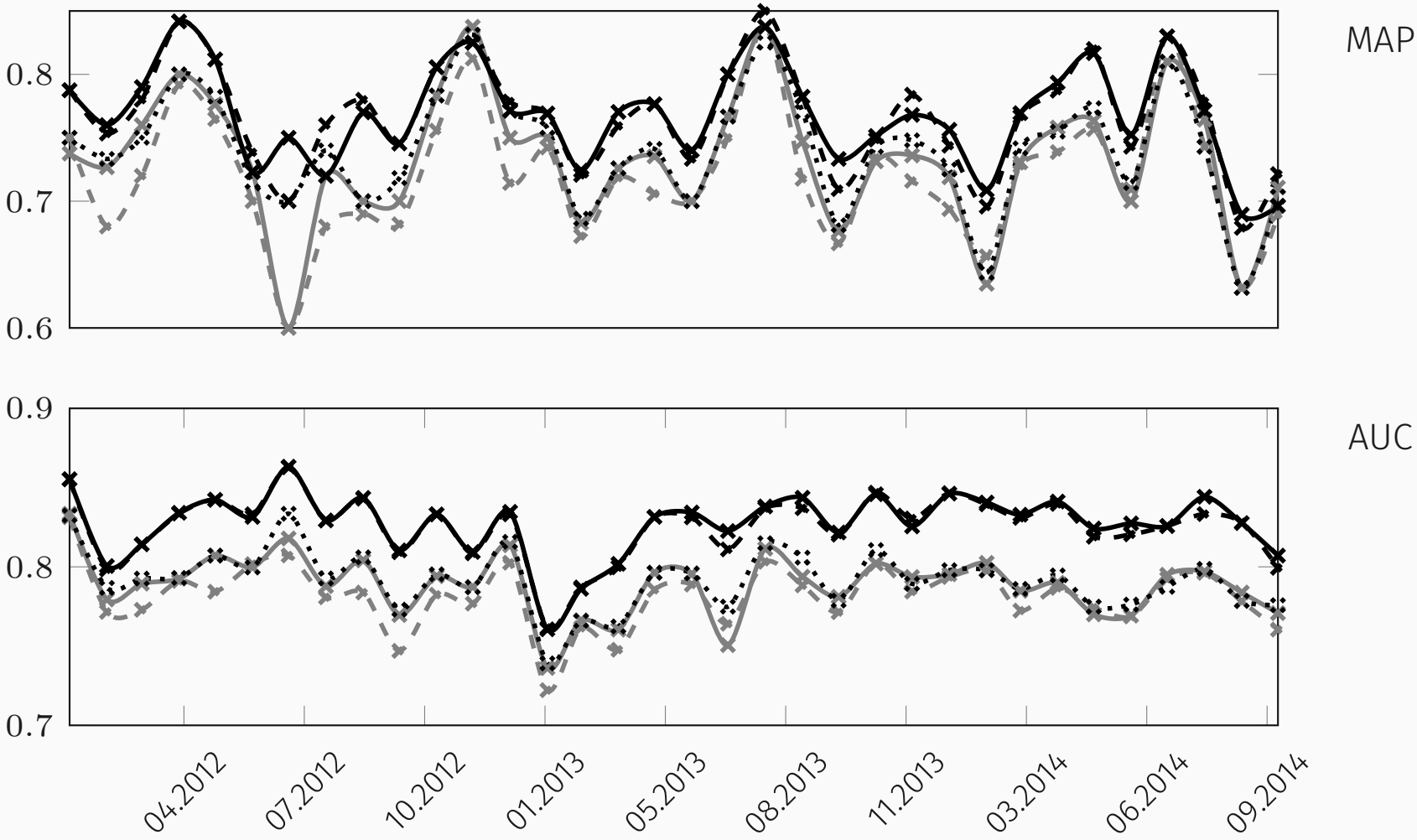
$$\mathbb{E}\left[z_{iq}\right] = \begin{cases} 0, & \text{если } x_{tq} = 0, \\ p(z_{iq} = 1 \mid \exists j \in t\ z_{jq} = 1) = \frac{\sigma(\mu+s_i+c_q)}{1-\prod_{j\in t}\left(1-\sigma(\mu+s_j+c_q)\right)}, & \text{если } x_{tq} = 1; \end{cases}$$

  - M-шаг: зафиксируем $\mathbb{E}\left[z_{iq}\right]$, обучим логистическую модель

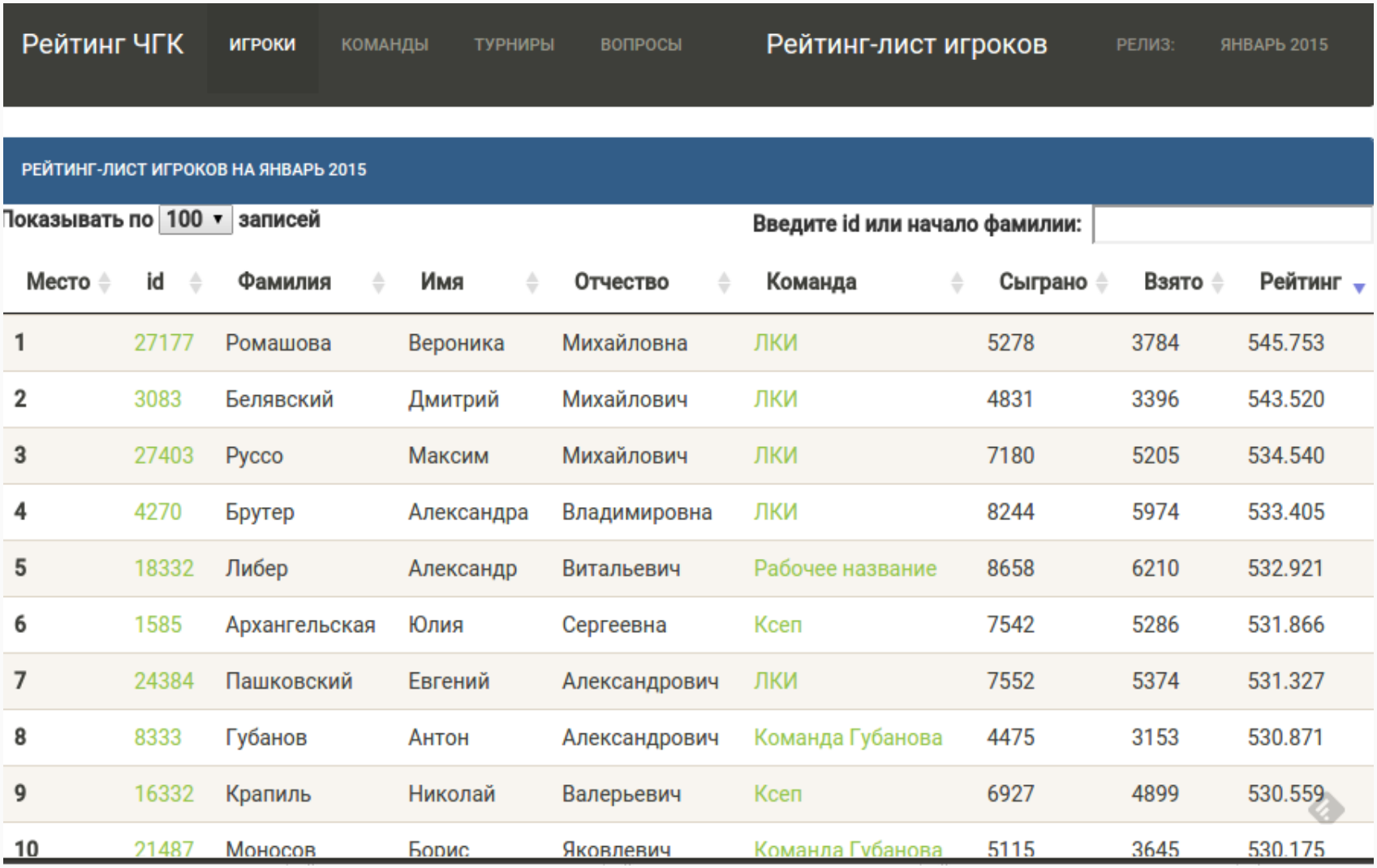
$$\mathbb{E}\left[z_{iq}\right] \sim \sigma(\mu + s_i + c_q).$$

Результаты

- Ну и, конечно, работает хорошо.



Пример



Рейтинг ЧГК | ИГРОКИ | КОМАНДЫ | ТУРНИРЫ | ВОПРОСЫ | Рейтинг-лист игроков | РЕЛИЗ: ЯНВАРЬ 2015

РЕЙТИНГ-ЛИСТ ИГРОКОВ НА ЯНВАРЬ 2015

Показывать по 100 записей

Введите id или начало фамилии:

Место	id	Фамилия	Имя	Отчество	Команда	Сыграно	Взято	Рейтинг
1	27177	Ромашова	Вероника	Михайловна	ЛКИ	5278	3784	545.753
2	3083	Белявский	Дмитрий	Михайлович	ЛКИ	4831	3396	543.520
3	27403	Руссо	Максим	Михайлович	ЛКИ	7180	5205	534.540
4	4270	Брутер	Александра	Владимировна	ЛКИ	8244	5974	533.405
5	18332	Либер	Александр	Витальевич	Рабочее название	8658	6210	532.921
6	1585	Архангельская	Юлия	Сергеевна	Ксеп	7542	5286	531.866
7	24384	Пашковский	Евгений	Александрович	ЛКИ	7552	5374	531.327
8	8333	Губанов	Антон	Александрович	Команда Губанова	4475	3153	530.871
9	16332	Крапиль	Николай	Валерьевич	Ксеп	6927	4899	530.559
10	21487	Моносов	Борис	Яковлевич	Команда Губанова	5115	3645	530.175

Пример

Рейтинг ЧГК ИГРОКИ КОМАНДЫ ТУРНИРЫ ВОПРОСЫ Александр Друзь vs. Максим Поташев

Ссылка на сайт рейтинга МАК

Друзь
Александр Абрамович

Поташев
Максим Оскарович

Ссылка на сайт рейтинга МАК

ИСТОРИЯ РЕЙТИНГА | ИСТОРИЯ ТУРНИРОВ

ИСТОРИЯ РЕЙТИНГОВ

Рейтинг	Команда	Место	Дата	Место	Команда	Рейтинг
450.138	Транссфера	303	Январь 2015	31	Афина	514.643
446.669	Транссфера	336	Декабрь 2014	29	Афина	514.697
442.559	Транссфера	348	Ноябрь 2014	25	Афина	513.155
437.999	Транссфера	400	Октябрь 2014	25	Афина	512.083
445.791	Транссфера	306	Сентябрь 2014	23	Афина	515.737
441.473	Транссфера	372	Август 2014	23	Афина	516.513
446.730	Транссфера	284	Июль 2014	22	Афина	516.755
448.598	Транссфера	322	Июнь 2014	24	Афина	516.603

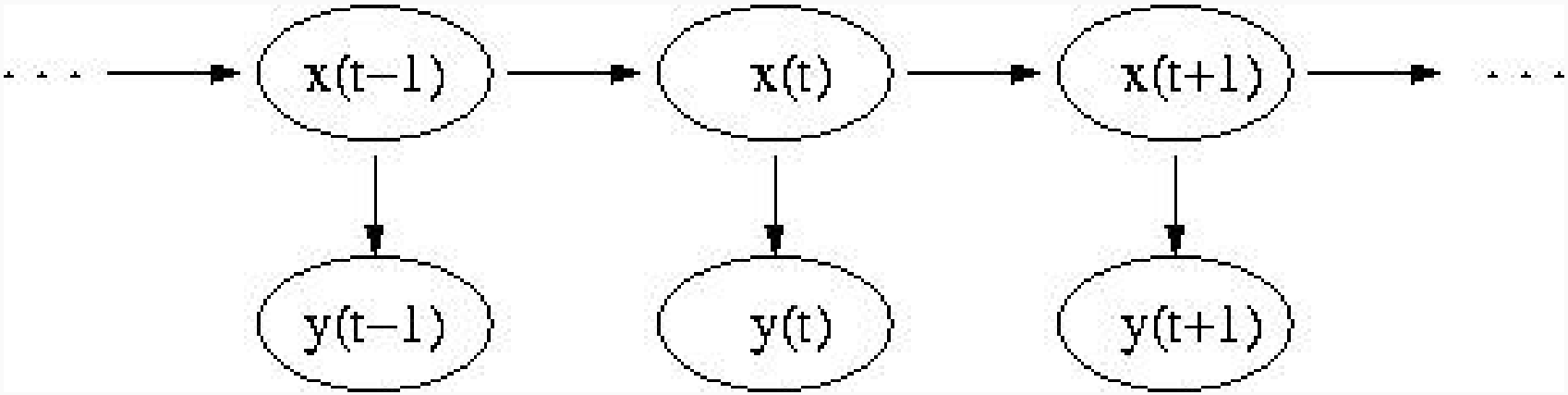
Марковские цепи

Марковские цепи

- Марковская цепь задаётся начальным распределением вероятностей $p^0(x)$ и вероятностями перехода $T(x'; x)$.
- $T(x'; x)$ — это распределение следующего элемента цепи в зависимости от следующего; распределение на $(t+1)$–м шаге равно
$$p^{t+1}(x') = \int T(x'; x) p^t(x) dx.$$
- В дискретном случае $T(x'; x)$ — это матрица вероятностей $p(x' = i | x = j)$.

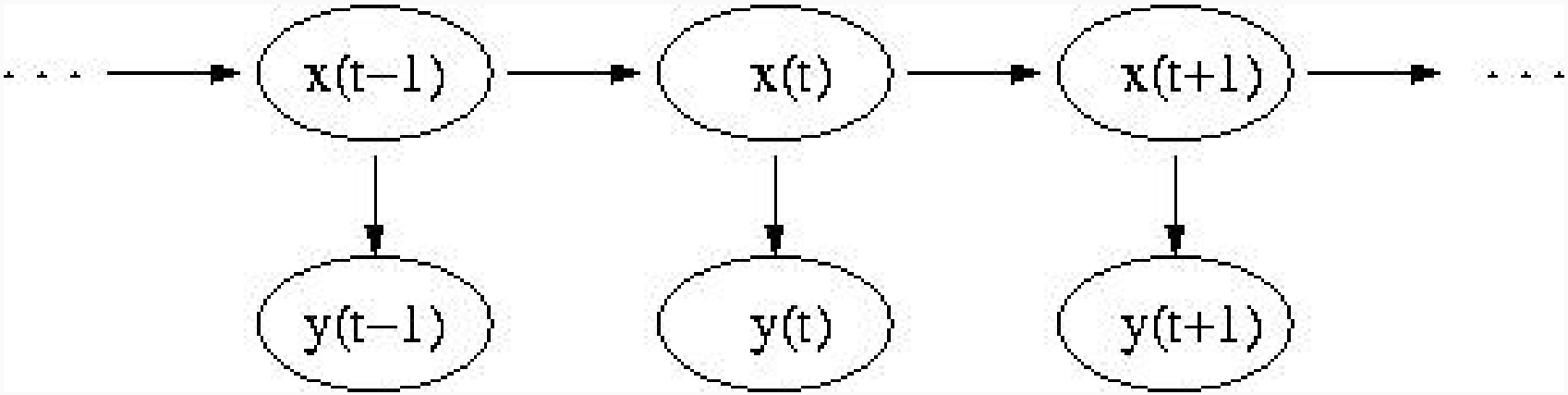
Дискретные марковские цепи

- Мы будем находиться в дискретном случае.
- Марковская модель — это когда мы можем наблюдать какие-то функции от марковского процесса.



Дискретные марковские цепи

- Здесь $x(t)$ — сам процесс (модель), а $y(t)$ — то, что мы наблюдаем.
- Задача — определить скрытые параметры процесса.



Дискретные марковские цепи

- Главное свойство — следующее состояние не зависит от истории, только от предыдущего состояния.

$$p(x(t) = x_j | x(t-1) = x_{j_{t-1}}, \dots, x(1) = x_{j_1}) = \\ = p(x(t) = x_j | x(t-1) = x_{j_{t-1}}).$$

- Более того, эти вероятности $a_{ij} = p(x(t) = x_j | x(t-1) = x_i)$ ещё и от времени t не зависят.
- Эти вероятности и составляют матрицу перехода $A = (a_{ij})$.

Вероятности перехода

- Естественные свойства:
- $a_{ij} \geq 0$.
- $\sum_j a_{ij} = 1$.

Прямая задача

- Естественная задача: с какой вероятностью выпадет та или иная последовательность событий?
- Т.е. найти нужно для последовательности $Q = q_{i_1} \dots q_{i_k}$

$$p(Q|\text{модель}) = p(q_{i_1})p(q_{i_2}|q_{i_1}) \dots p(q_{i_k}|q_{i_{k-1}}).$$

- Казалось бы, это тривиально.
- Что же сложного в реальных задачах?

Скрытые марковские модели

- А сложно то, что никто нам не скажет, что модель должна быть именно такой.
- И, кроме того, мы обычно наблюдаем не $x(t)$, т.е. реальные состояния модели, а $y(t)$, т.е. некоторую функцию от них (данные).
- Пример: распознавание речи.

Спасибо за внимание!