ЕМ-АЛГОРИТМ II

Сергей Николенко

СПбГУ — Санкт-Петербург

22 октября 2020 г.

Random facts:

- 22 октября — Международный день заикающихся людей, отмечающийся с 1998 года при поддержке Международной ассоциации заикающихся людей (International Stuttering Association, ISA)
- 22 октября 1797 г. Андре Жак Гарнерен совершил первый в истории прыжок с парашютом с летательного аппарата (воздушного шара) с высоты около 1000 м над парижским парком Монсо; а 22 октября 1909 г. первая женщина — француженка Элиза Дерош — совершила одиночный полёт на самолёте
- 22 октября 1895 г. произошло знаменитое крушение на вокзале Монпарнас, когда пассажирский поезд выбил путевой упор, выехал на перрон вокзала, пробил стену здания и рухнул с высоты на улицу
- 22 октября 1938 г. Честер Карлсон продемонстрировал свой аппарат для получения копий бумажных документов методом сухой фотографии
- 22 октября 1964 г. Жан-Поль Сартр отказался от Нобелевской премии, а 22 октября 1987 г. она была присуждена Иосифу Бродскому
- 22 октября 1980 г. папа римский отменил вердикт 1633 года, осуждающий Галилея

Обоснование алгоритма EM

Обоснование алгоритма EM

- Дадим формальное обоснование алгоритма EM.
- Мы решаем задачу максимизации правдоподобия по данным $\mathbf{y} = \{y_1, \cdots, y_N\}$.

$$L(\theta \mid Y) = p(Y \mid \theta) = \prod p(y_i \mid \theta)$$

  или, что то же самое, максимизации $\ell(\theta \mid Y) = \log L(\theta \mid Y)$.
- EM может помочь, если этот максимум трудно найти аналитически.

Обоснование алгоритма EM

- Давайте предположим, что в данных есть *скрытые компоненты*, такие, что если бы мы их знали, задача была бы проще.
- Замечание: совершенно не обязательно эти компоненты должны иметь какой-то физический смысл. :) Может быть, так просто удобнее.
- В любом случае, получается набор данных $X = (Y, Z)$ с совместной плотностью

$$p(\mathbf{x} \mid \theta) = p(\mathbf{y}, \mathbf{z} \mid \theta) = p(\mathbf{z} \mid \mathbf{y}, \theta) p(\mathbf{y} \mid \theta).$$

- Получается полное правдоподобие $L(\theta \mid X) = p(Y, Z \mid \theta)$. Это случайная величина (т.к. Z неизвестно).

Обоснование алгоритма EM

- Заметим, что настоящее правдоподобие $L(\theta) = E_Z\left[p(Y, Z \mid \theta) \mid Y, \theta\right]$.
- E-шаг алгоритма EM вычисляет условное ожидание (логарифма) полного правдоподобия при условии Y и текущих оценок параметров θ_n:

$$Q(\theta, \theta_n) = E\left[\log p(Y, Z \mid \theta) \mid Y, \theta_n\right].$$

- Здесь θ_n — текущие оценки, а θ — неизвестные значения (которые мы хотим получить в конечном счёте); т.е. $Q(\theta, \theta_n)$ — это функция от θ.

Обоснование алгоритма EM

- E-шаг алгоритма EM вычисляет условное ожидание (логарифма) полного правдоподобия при условии Y и текущих оценок параметров θ:

$$Q(\theta, \theta_n) = E\left[\log p(Y, Z \mid \theta) \mid Y, \theta_n\right].$$

- Условное ожидание — это

$$E\left[\log p(Y, Z \mid \theta) \mid Y, \theta_n\right] = \int_{\mathbf{z}} \log p(Y, \mathbf{z} \mid \theta) p(\mathbf{z} \mid Y, \theta_n) \mathrm{d}\mathbf{z},$$

  где $p(\mathbf{z} \mid Y, \theta_n)$ — маргинальное распределение скрытых компонентов данных.
- EM лучше всего применять, когда это выражение легко подсчитать, может быть, даже аналитически.
- Вместо $p(\mathbf{z} \mid Y, \theta_n)$ можно подставить $p(\mathbf{z}, Y \mid \theta_n) = p(\mathbf{z} \mid Y, \theta_n) p(Y \mid \theta_n)$, от этого ничего не изменится.

Обоснование алгоритма EM

- В итоге после E-шага алгоритма EM мы получаем функцию $Q(\theta, \theta_n)$.
- На M-шаге мы максимизируем

$$\theta_{n+1} = \arg\max_\theta Q(\theta, \theta_n).$$

- Затем повторяем процедуру до сходимости.
- В принципе, достаточно просто находить θ_{n+1}, для которого $Q(\theta_{n+1}, \theta_n) > Q(\theta_n, \theta_n)$ — Generalized EM.
- Осталось понять, что значит $Q(\theta, \theta_n)$ и почему всё это работает.

Обоснование алгоритма EM

- Мы хотели перейти от θ_n к θ, для которого $\ell(\theta) > \ell(\theta_n)$.

$$\ell(\theta) - \ell(\theta_n) =$$

$$= \log\left(\int_{\mathbf{z}} p(Y \mid \mathbf{z}, \theta) p(\mathbf{z} \mid \theta) \mathrm{d}\mathbf{z}\right) - \log p(Y \mid \theta_n) =$$

$$= \log\left(\int_{\mathbf{z}} p(\mathbf{z} \mid Y, \theta_n) \frac{p(Y \mid \mathbf{z}, \theta) p(\mathbf{z} \mid \theta)}{p(\mathbf{z} \mid Y, \theta_n)} \mathrm{d}\mathbf{z}\right) - \log p(Y \mid \theta_n) \geq$$

$$\geq \int_{\mathbf{z}} p(\mathbf{z} \mid Y, \theta_n) \log\left(\frac{p(Y \mid \mathbf{z}, \theta) p(\mathbf{z} \mid \theta)}{p(\mathbf{z} \mid Y, \theta_n)}\right) \mathrm{d}\mathbf{z} - \log p(Y \mid \theta_n) =$$

$$= \int_{\mathbf{z}} p(\mathbf{z} \mid Y, \theta_n) \log\left(\frac{p(Y \mid \mathbf{z}, \theta) p(\mathbf{z} \mid \theta)}{p(Y \mid \theta_n) p(\mathbf{z} \mid Y, \theta_n)}\right) \mathrm{d}\mathbf{z}.$$

Обоснование алгоритма EM

- Получили

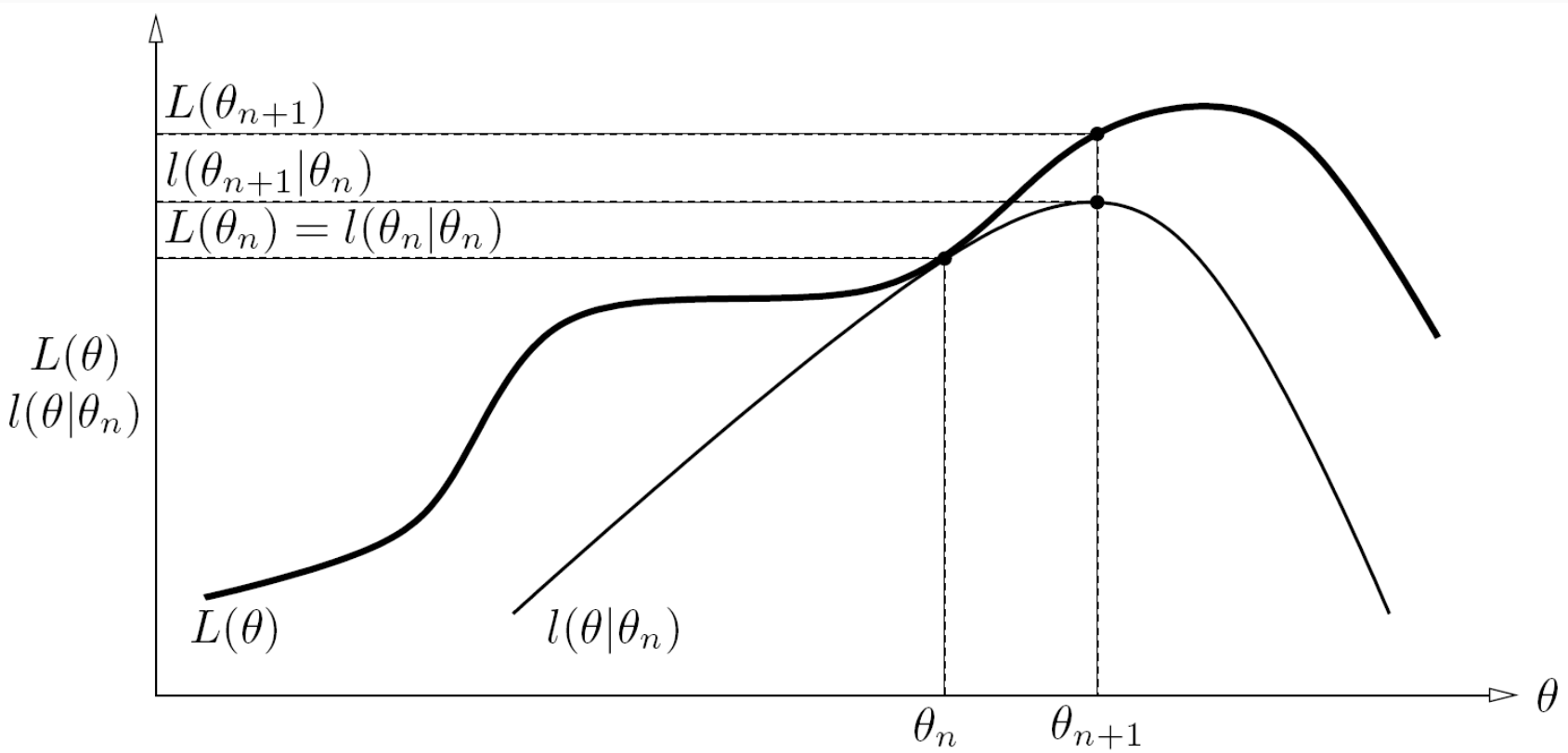
$$\ell(\theta) \geq \mathcal{L}(\theta, \theta_n) =$$
$$= \ell(\theta_n) + \int_{\mathbf{z}} p(\mathbf{z} \mid Y, \theta_n) \log \left(\frac{p(Y \mid \mathbf{z}, \theta) p(\mathbf{z} \mid \theta)}{p(Y \mid \theta_n) p(\mathbf{z} \mid Y, \theta_n)} \right) \mathrm{d}\mathbf{z}.$$

Упражнение. Докажите, что $\mathcal{L}(\theta_n, \theta_n) = \ell(\theta_n)$.

Обоснование алгоритма EM

- Иначе говоря, мы нашли нижнюю оценку на $\ell(\theta)$ везде, касание происходит в точке θ_n.
- Т.е. мы нашли нижнюю оценку для правдоподобия и смещаемся в точку, где она максимальна (или хотя бы больше текущей).
- Такая общая схема называется *MM-алгоритм* (minorization-maximization). Мы к ним ещё вернёмся.

Обоснование алгоритма EM



Обоснование алгоритма EM

- Осталось только понять, что максимизировать можно Q.

$$\theta_{n+1} = \arg\max_\theta l(\theta, \theta_n) = \arg\max_\theta \{\ell(\theta_n) + \\ + \int_{\mathbf{z}} f(y \mid X, \theta_n) \log \left(\frac{p(X \mid y, \theta) f(y \mid \theta)}{p(X \mid \theta_n) f(y \mid X, \theta_n)} \right) \mathrm{d}\mathbf{z} \Bigg\} = \\ = \arg\max_\theta \left\{ \int_{\mathbf{z}} p(\mathbf{z} \mid X, \theta_n) \log \left(p(X \mid y, \theta) p(\mathbf{z} \mid \theta) \right) \mathrm{d}\mathbf{z} \right\} = \\ = \arg\max_\theta \left\{ \int_{\mathbf{z}} p(\mathbf{z} \mid X, \theta_n) \log p(X, y \mid \theta) \mathrm{d}\mathbf{z} \right\} = \\ = \arg\max_\theta \{Q(\theta, \theta_n)\},$$

а остальное от θ не зависит. Вот и получился EM.

История и первые применения

История EM-алгоритма

- Всё это появилось в работе Dempster–Laird–Rubin; доклад на Royal Statistical Society в 1976, статья вышла в 1977

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

Keywords: Maximum likelihood; incomplete data; EM algorithm; posterior mode

- Но были и более ранние аналоги (сами DLR тоже пишут, что было много примеров раньше, и ссылаются на них)...

История EM-алгоритма

- (Ceppellini et al., 1955): подсчёт частот генов в популяции:
  - в популяции есть k аллелей с частотами $p_1, \dots, p_k$ и генами $G_1, \dots, G_k$;
  - например, в рассмотренной MN-системе есть два аллеля $G_1 = M$ и $G_2 = N$, и у человека диплоидные клетки, т.е. бывают варианты MM, MN и NN; если бы мы умели различать все три варианта, то было бы легко подсчитать гены каждого типа;
  - но что если M доминантный, а N рецессивный, т.е. гомозиготные MM- и гетерозиготные MN-организмы неразличимы?
  - есть закон Харди-Вайнберга, который говорит, что гомозиготы и гетерозиготы встречаются с частотами $\frac{p^2}{p^2+2pq}$ и $\frac{2pq}{p^2+2pq}$, где p и $q = 1 - p$ — частоты генов;
  - можно было бы всё подсчитать, но получается замкнутый круг: нужно знать частоты генов, чтобы посчитать долю MM и MN, но чтобы посчитать частоты, нужно знать долю MM и MN...

История EM-алгоритма

- И тут Ceppellini et al. говорят:

> by counting genes we can obtain estimates p' and q' of the gene frequencies. Unfortunately, this argument is still circular, since it presupposes a knowledge of the gene frequencies in order to obtain the estimate. But if any value is provisionally assumed for p, say $p(1)$, the new estimate $p' = p(2)$ obtained by gene counting will be rather more accurate, since the number of genes in the recessive individuals is known for certain, and the provisional value $p(1)$ is used only in estimating the number of genes among the dominants. This new value $p(2)$ can be taken as a new provisional value, and a further estimate $p'(2) = p(3)$ obtained by gene counting. The last process can be continued, giving a series of values $p(1)$, $p(2)$, $p(3)$, ..., each more accurate than the last; when two successive values are equal to the order of accuracy desired their common value can be taken as the final estimate.

- Это, видимо, одно из самых ранних применений EM-алгоритма, и такие применения актуальны, конечно, до сих пор.

Свойства и простые 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 $\mathbb{E}_{X|Y,\theta^{(m)}}\left[\log p(X \mid \theta)\right]$ 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.

Спасибо!

Спасибо за внимание!