ВМК МГУ

EM-алгоритм для разделения смеси распределений

Курс: Практикум на ЭВМ для 317-й группы, осень 2015

EM-алгоритм в общем виде

Для ознакомления с EM-алгоритмом рекомендуется книга [1], глава 9. Пусть имеется вероятностная модель, задаваемая совместным распределением $p(X,T|\Theta)$. Здесь X – набор наблюдаемых переменных, T – набор ненаблюдаемых переменных и Θ – набор параметров модели. Рассмотрим задачу обучения модели (поиск параметров Θ по выборке X) с помощью метода максимального правдоподобия:

$$\log p(X|\Theta) = \log \int p(X,T|\Theta)dT \to \max_{\Theta}. \tag{1}$$

Рассмотрим произвольное вероятностное распределение $q(T)$. Тогда справедлива следующая цепочка равенств:

$$\log p(X|\Theta) = \int q(T)\log p(X|\Theta)dT = \int q(T)\log\frac{p(X,T|\Theta)}{p(T|X,\Theta)}dT = \int q(T)\log\left[\frac{p(X,T|\Theta)}{q(T)}\frac{q(T)}{p(T|X,\Theta)}\right]dT =$$
$$= \underbrace{\int q(T)\log p(X,T|\Theta)dT - \int q(T)\log q(T)dT}_{\mathcal{L}(q,\Theta)} \underbrace{- \int q(T)\log\frac{p(T|X,\Theta)}{q(T)}dT}_{\mathrm{KL}(q(T)\,||\,p(T|X,\Theta))}.$$

Дивергенция $\mathrm{KL}(q(T)\,||\,p(T|X,\Theta)) \geq 0$, следовательно

$$\log p(X|\Theta) \geq \mathcal{L}(q,\Theta). \tag{2}$$

EM-алгоритм для решения задачи (1) представляет собой покомпонентную максимизацию нижней границы $\mathcal{L}(q,\Theta)$. Пусть фиксировано некоторое значение параметров Θ_{old}. Сначала (на E-шаге) функционал $\mathcal{L}(q,\Theta_{old})$ максимизируется по распределению $q(T)$. Эта задача имеет аналитическое решение, т.к. КЛ-дивергенция обнуляется для тождественных распределений:

$$q(T) = p(T|X,\Theta_{old}) = \frac{p(X,T|\Theta_{old})}{\int p(X,T|\Theta_{old})dT}. \tag{3}$$

При таком выборе $q(T)$ нижняя оценка (2) является точной при $\Theta = \Theta_{old}$. Затем (на M-шаге) при фиксированном $q(T)$ новые значения параметров Θ находятся путем максимизации нижней границы $\mathcal{L}(q,\Theta)$, что эквивалентно решению задачи

$$\mathbb{E}_{q(T)}\log p(X,T|\Theta) \to \max_{\Theta}, \tag{4}$$

так как энтропия $-\mathbb{E}_{q(T)}\log q(T)$ распределения $q(T)$ не зависит от Θ. Шаги E и M повторяются в цикле до сходимости. Очевидно, что в процессе EM-итераций нижняя оценка (2), а также значение правдоподобия $p(X|\Theta)$ не убывают.

Итерационный процесс в EM-алгоритме проиллюстрирован на рис. 1. С учётом того, что правдоподобие $p(X|\Theta)$ не является, вообще говоря, выпуклой или унимодальной функцией, EM-алгоритм позволяет находить только локальный максимум правдоподобия. Поэтому на практике EM-алгоритм запускают несколько раз из различных начальных приближений с выбором наилучшего решения по значению найденного правдоподобия.

Заметим, что во многих практических случаях решение задачи (4) намного проще, чем решение задачи (1). В частности, в рассматриваемой ниже задаче разделения гауссовской смеси задача оптимизации на M шаге решается аналитически.

Вычисление значения функции правдоподобия $p(X|\Theta)$ в фиксированной точке Θ требует интегрирования по пространству T и в ряде случаев может представлять собой вычислительно трудоемкую задачу. Заметим, что эта величина правдоподобия необходима также для вычисления апостериорного распределения $p(T|X,\Theta_{old})$ на E шаге. Однако, распределение $p(T|X,\Theta_{old})$ используется затем только для вычисления математического ожидания логарифма полного правдоподобия на M шаге. Как правило, здесь не требуется знать все апостериорное

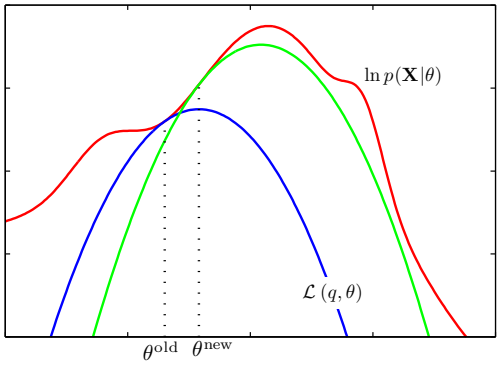

Рис. 1: Иллюстрация итерационного процесса в EM-алгоритме. Нижняя оценка (2) обозначена через $\mathcal{L}(q,\theta)$.

распределение целиком, а достаточно знать лишь несколько статистик этого распределения (например, только мат.ожидания отдельных компонент $\mathbb{E}_{T|X,\Theta_{old}} t_n$ и парные ковариации $\mathbb{E}_{T|X,\Theta_{old}} t_n t_k$). Поэтому EM-алгоритм может быть вычислительно эффективен даже в тех случаях, когда вычисление значения правдоподобия $p(X|\Theta)$ в одной точке затруднено.

Как было отмечено выше, EM-итерации соответствуют монотонному увеличению значения правдоподобия $p(X|\Theta)$. Однако, данное обстоятельство само по себе ещё не гарантирует сходимость EM-итераций к локальному оптимуму правдоподобия. Например, при максимизации одномерной функции $-(x+1)^2$ последовательность точек $x_k = 1/k$ обеспечивает монотонный рост значения оптимизируемой функции, но при этом не обеспечивает сходимость к её локальному максимуму. Тем не менее, в случае EM-алгоритма можно доказать сходимость к локальному максимуму правдоподобия при минимальных требованиях на вероятностную модель $p(X|\Theta)$ [2].

EM-алгоритм можно применять также для решения задачи обучения вероятностной модели со скрытыми переменными $p(X,T|\Theta)$ с помощью максимизации апостериорного распределения:

$$p(\Theta|X) \to \max_{\Theta} \;\Leftrightarrow\; \log p(X|\Theta) + \log p(\Theta) \to \max_{\Theta}.$$

Здесь $p(\Theta)$ — априорное распределение на параметры модели. В этом случае нижней оценкой для оптимизируемого функционала является следующая величина:

$$\log p(X|\Theta) + \log p(\Theta) \geq \mathcal{L}(q,\Theta) + \log p(\Theta).$$

Итерационная оптимизация данной нижней границы по распределению $q(T)$ и по параметрам Θ приводит к E шагу (3) и M шагу

$$\mathbb{E}_{T|X,\Theta_{old}} \log p(X,T|\Theta) + \log p(\Theta) \to \max_{\Theta}.$$

EM-алгоритм для разделения гауссовской смеси

Рассмотрим вероятностную модель смеси нормальных распределений:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k),\; \sum_{k=1}^{K} w_k = 1,\; w_k \geq 0. \tag{5}$$

Модель смеси распределений (не обязательно нормальных) можно рассматривать как модель со скрытой переменной t, которая обозначает номер компоненты смеси:

$$p(t=k) = w_k, \tag{6}$$
$$p(\boldsymbol{x}|t=k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k). \tag{7}$$

Легко показать, что маргинальное распределение $p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}|t=k)p(t=k)$ в этой модели совпадает с распределением (5). В этом смысле модели (6)-(7) и (5) эквивалентны.

Интерпретация вероятностной модели смеси распределений как модели со скрытой переменной позволяет генерировать выборку из модели смеси следующим образом. Сначала с вероятностями, равными $\boldsymbol{w}$, генерируется номер компоненты смеси t, а затем точка $\boldsymbol{x}$ генерируется из компоненты с номером t.

Для аппроксимации выборки $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ с помощью модели смеси из K гауссиан воспользуемся методом максимального правдоподобия:

$$\begin{aligned} &p(X|\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}) = \prod_{n=1}^{N} \left(\sum_k w_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k) \right) \to \max_{\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}}, \\ &\sum_k w_k = 1, \;\; w_k \geq 0, \\ &\Sigma_k = \Sigma_k^T, \;\; \Sigma_k \succ 0. \end{aligned} \tag{8}$$

Данная задача условной оптимизации может быть эффективно решена с помощью EM-алгоритма, описанного выше. Заметим, что количество компонент смеси K не может быть найдено аналогичным образом с помощью максимизации правдоподобия, т.к. значение правдоподобия данных тем выше, чем больше компонент K используется. Для поиска оптимального значения K можно воспользоваться скользящим контролем, где критерием качества аппроксимации тестовых данных является значение правдоподобия.

При использовании EM-алгоритма для решения задачи (8) вероятностная модель смеси распределений интерпретируется как вероятностная модель со скрытыми переменными. Вычислим значение мат.ожидания логарифма полного правдоподобия, необходимого для решения задачи оптимизации на M шаге:

$$\mathbb{E}_q \log p(X, T|\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} q(t_n = k)(\log w_k + \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)). \tag{9}$$

Заметим, что это выражение зависит только от вероятностей отдельных скрытых переменных $q(t_n = k)$. Нетрудно показать, что эти величины можно найти следующим образом:

$$\gamma_{nk} \triangleq q(t_n = k) = p(t_n = k|\boldsymbol{x}_n, \boldsymbol{w}^{old}, \{\boldsymbol{\mu}_k^{old}\}, \{\Sigma_k^{old}\}) = \frac{w_k^{old} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{old}, \Sigma_k^{old})}{\sum_{j=1}^{K} w_j^{old} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{old}, \Sigma_j^{old})}. \tag{10}$$

Также нетрудно показать, что задача максимизации критерия (9) при ограничениях $\sum_{k=1}^{K} w_k = 1, \; w_k \geq 0$ может быть решена аналитически:

$$w_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}, \tag{11}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}, \tag{12}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma_{nk}}. \tag{13}$$

Заметим, что решение для Σ_k (13) удовлетворяет условию симметричности и положительной определенности. Кроме того, формулы (12),(13) соответствуют оценкам максимального правдоподобия для многомерного нормального распределения, в которых каждый объект $\boldsymbol{x}_n$ берется с весом γ_{nk}.

Таким образом, EM-алгоритм для смеси нормальных распределений заключается в итерационном применении формул (10) и (11)-(13). Этот процесс имеет простую интерпретацию. Величина γ_{nk} показывает степень соответствия между объектом $\boldsymbol{x}_n$ и компонентой k (определяет вес объекта $\boldsymbol{x}_n$ для компоненты k). Эти веса затем используются на M шаге для вычисления новых значений параметров компонент. Иллюстрация применения EM-алгоритма для разделения нормальной смеси с двумя компонентами показана на рис. 2.

Одним из применений смеси нормальных распределений является решение задачи кластеризации на K кластеров. В этом случае номер кластера для объекта $\boldsymbol{x}_n$ определяется величиной

$$t_n = \arg\max_k \gamma_{nk}.$$

Такая схема кластеризации является вероятностным обобщением известного метода кластеризации K-средних.

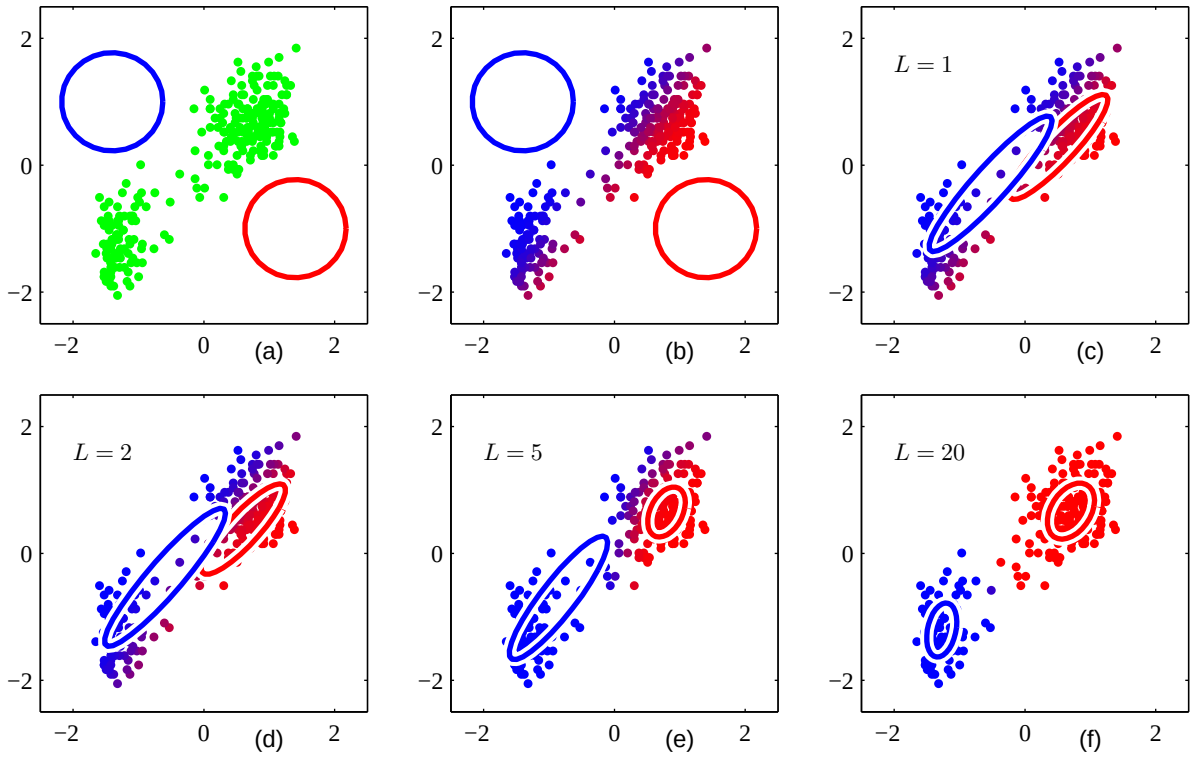

Рис. 2: Иллюстрация применения EM-алгоритма для разделения смеси нормальных распределений с двумя компонентами. На рис. a показана исходная выборка и начальное приближение для двух компонент. На рис. b показан результат E шага. При этом цвета объектов соответствуют значениям γ_{nk}. На рис. c-f показаны результаты вычислений после 1, 2, 5 и 20 итераций.

В заключение заметим, что восстановление плотности по данным (в частности, смеси нормальных распределений) является простейшим способом решения задачи идентификации. Для этого сначала для всех объектов обучающей выборки, обладающих заданным свойством, восстанавливается плотность распределения. Затем, для нового объекта $\boldsymbol{x}$ решение о наличии у него заданного свойства принимается, если значение восстановленной плотности $p(\boldsymbol{x})$ выше определенного порога.

Список литературы

[1] C. Bishop. Pattern recognition and machine learning. Springer, 2006.

[2] F. Vaida. Parameter convergence for EM and MM algorithms // Statistica Sinica, Vol. 15, 2005.