Вариационный вывод в графических моделях

Задача приближённого вывода в вероятностных моделях

Рассмотрим задачу вывода в вероятностных моделях в следующей постановке. Пусть имеется некоторое вероятностное распределение, известное с точностью до нормировочной константы:

$$p(\boldsymbol{x}) = \frac{1}{Z_p}\tilde{p}(\boldsymbol{x}),$$

т.е. величина $\tilde{p}(\boldsymbol{x})$ может быть вычислена в произвольной точке $\boldsymbol{x}$, а точное значение нормировочной константы $Z_p = \int \tilde{p}(\boldsymbol{x})d\boldsymbol{x}$ является недоступным. Тогда задачей вывода назовём задачу оценки некоторой статистики $f(\boldsymbol{x})$ для распределения $p(\boldsymbol{x})$, т.е. величины $\mathbb{E}_p f(\boldsymbol{x}) = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$.

Рассмотрим несколько примеров задачи вывода. Пусть имеется некоторая вероятностная модель, задаваемая совместным распределением $p(X,T)$. Здесь X – известные переменные, T – скрытые переменные. Тогда вывод относительно переменных T соответствует вычислению апостериорного распределения

$$p(T|X) = \frac{p(X,T)}{p(X)} = \frac{p(X,T)}{\int p(X,\tilde{T})d\tilde{T}}.$$

Для получения точечной оценки на T можно взять мат.ожидание апостериорного распределения $\hat{T} = \mathbb{E}_{T|X}T$[1]. В этом случае в качестве ненормированной плотности $\tilde{p}$ выступает совместное распределение $p(X,T)$, в качестве недоступной для вычисления нормировочной константы – $p(X)$, а в качестве статистики $f(T) = T$.

Другой пример задачи вывода связан с EM-алгоритмом для оценки параметров вероятностных моделей со скрытыми переменными. Пусть вероятностная модель задаётся распределением $p(X,T|\Theta)$, где X – наблюдаемый набор переменных, T – набор скрытых переменных, а Θ – параметры модели. Тогда задачу оценки Θ с помощью максимизации неполного правдоподобия

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

можно решать с помощью итерационного EM-алгоритма, в котором на E-шаге при текущих параметрах Θ вычисляется апостериорное распределение на скрытые переменные $q(T) = p(T|X,\Theta)$, а затем на M-шаге новые значения параметров Θ оцениваются путем максимизации функционала $\mathbb{E}_q \log p(X,T|\Theta)$ по Θ. С точки зрения общей постановки задачи вывода здесь в качестве ненормированного распределения $\tilde{p}$ выступает совместное распределение $p(X,T|\Theta)$, в качестве недоступной нормировочной константы – неполное правдоподобие $p(X|\Theta)$, а искомой статистикой $f(T)$ является величина $\log p(X,T|\Theta)$.

Задача вывода $\mathbb{E}_p f(\boldsymbol{x}) = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ требует интегрирования по многомерному пространству $\boldsymbol{x}$. Во многих практических ситуациях такой интеграл не может быть вычислен аналитически или эффективно оценён численно с помощью квадратурных формул. В этом случае возникает задача приближённого вывода, т.е. построения приближённой оценки для $\mathbb{E}_p f(\boldsymbol{x})$. В рамках подхода детерминированной аппроксимации здесь для распределения $p(\boldsymbol{x})$ сначала находится приближение $q(\boldsymbol{x})$ в некотором простом семействе распределений, а затем требуемая статистика оценивается как

$$\mathbb{E}_p f(\boldsymbol{x}) \approx \mathbb{E}_q f(\boldsymbol{x}).$$

[1] Строго говоря, вид оцениваемой статистики определяется задачей минимизации функционала среднего риска с выбранной функцией потерь.

Дивергенция Кульбака-Лейблера

Пусть $g(t) : \mathbb{R} \to \mathbb{R}$ – произвольная строго вогнутая функция, т.е. $\forall t_1, t_2$ и $\forall \alpha \in [0, 1]$ выполняется неравенство

$$g(\alpha t_1 + (1 - \alpha)t_2) \geq \alpha g(t_1) + (1 - \alpha)g(t_2),$$

причем равенство достигается только в случае $t_1 = t_2$ или $\alpha = 0, 1$. Нетрудно показать (по индукции), что это неравенство остаётся справедливым и для большего числа точек $t_1, \ldots, t_n$:

$$g\left(\sum_{i=1}^{n} \alpha_i t_i\right) \geq \sum_{i=1}^{n} \alpha_i g(t_i) \ \forall t_1, \ldots, t_n, \ \forall \boldsymbol{\alpha} : \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0. \tag{1}$$

Как и раньше, равенство здесь достигается только при $t_1 = t_2 = \cdots = t_n$ или $\alpha_i = 1$ для некоторого i. Результат (1) известен как неравенство Йенсена. Используя предельный переход, неравенство (1) можно сформулировать в интегральной форме

$$g\left(\int \alpha(\boldsymbol{x}) t(\boldsymbol{x}) d\boldsymbol{x}\right) \geq \int \alpha(\boldsymbol{x}) g(t(\boldsymbol{x})) d\boldsymbol{x}, \ \forall \alpha(\boldsymbol{x}) : \int \alpha(\boldsymbol{x}) d\boldsymbol{x} = 1, \ \alpha(\boldsymbol{x}) \geq 0.$$

Рассмотрим два произвольных вероятностных распределения $p(\boldsymbol{x})$ и $q(\boldsymbol{x})$. Дивергенцией Кульбака-Лейблера (сокращённо KL-дивергенцией) между распределениями $q(\boldsymbol{x})$ и $p(\boldsymbol{x})$ назовём величину

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}.$$

Эта характеристика обладает следующими свойствами:

1. Неотрицательность: $\mathrm{KL}(q||p) \geq 0$ и равенство достигается тогда и только тогда, когда $q(\boldsymbol{x}) \equiv p(\boldsymbol{x})$;
2. Несимметричность: $\mathrm{KL}(q||p) \neq \mathrm{KL}(p||q)$.

Свойство неотрицательности непосредственно вытекает из неравенства Йенсена. Действительно,

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \geq \{\text{Н-во Йенсена для логарифма}\} \geq -\log\left(\int q(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}\right) =$$
$$= -\log\left(\int p(\boldsymbol{x}) d\boldsymbol{x}\right) = -\log(1) = 0.$$

Равенство в этой цепочке рассуждений достигается только в случае $p(\boldsymbol{x})/q(\boldsymbol{x}) = \mathrm{const}$, т.е. с учётом условия нормировки для плотностей $p(\boldsymbol{x}) \equiv q(\boldsymbol{x})$.

Благодаря свойству неотрицательности KL-дивергенцию можно рассматривать как несимметричную меру отклонения между двумя вероятностными распределениями. В рамках упомянутого выше подхода детерминированной аппроксимации необходимо искать приближение $q(\boldsymbol{x})$ для распределения $p(\boldsymbol{x})$. Будем решать эту задачу с помощью функционала KL-дивергенции. С учетом несимметричности KL-дивергенции здесь можно сформулировать две оптимизационные задачи:

$$\mathrm{KL}(q||p) \to \min_q \ – \text{минимизация прямой дивергенции,}$$
$$\mathrm{KL}(p||q) \to \min_q \ – \text{минимизация обратной дивергенции.}$$

Решение этих двух задач соответствует поиску аппроксимаций с различными свойствами (см. рис. 1). Прямая дивергенция $\mathrm{KL}(q||p)$ принимает большие значения там, где, во-первых, плотности $p(\boldsymbol{x})$ и $q(\boldsymbol{x})$ значительно отличаются, а, во-вторых, плотность $q(\boldsymbol{x})$ существенно отлична от нуля. В результате носитель аппроксимации $q(\boldsymbol{x})$, как правило, оказывается подмножеством носителя исходного распределения $p(\boldsymbol{x})$[2]. Это может приводить к недооценке статистик распределения $p(\boldsymbol{x})$, связанных с разбросом. Минимизация обратной дивергенции $\mathrm{KL}(p||q)$ не ведёт к сужению носителя аппроксиманта $q(\boldsymbol{x})$, но при этом может приводить к переоценке некоторых статистик распределения $p(\boldsymbol{x})$.

В дальнейшем будем рассматривать только задачу минимизации прямой дивергенции. Минимизация обратной дивергенции реализуется в подходе Expectation Propagation, который здесь затрагиваться не будет.

[2]Здесь под носителем понимается область значений $\boldsymbol{x}$, для которых плотность распределения значимо отличается от нуля.

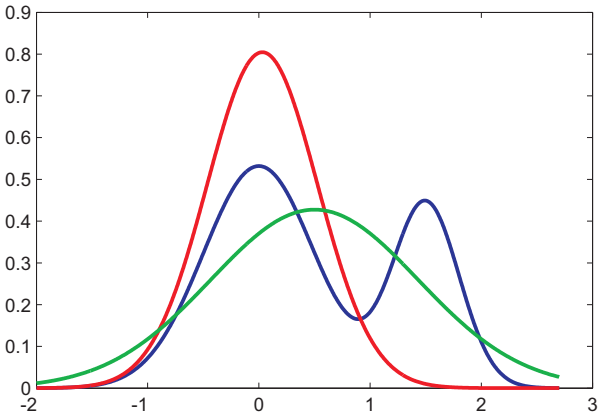

Рис. 1: Приближение двумодального распределения (синяя кривая) с помощью одномодального распределения путем минимизации прямой и обратной KL-дивергенции. Минимизация прямой KL-дивергенции (красная кривая) соответствует поиску приближения на подмножестве носителя исходного распределения. Минимизация обратной KL-дивергенции (зеленая кривая) соответствует поиску приближения на полном носителе исходного распределения.

Решение задачи минимизации прямой KL-дивергенции

Рассмотрим решение задачи $\mathrm{KL}(q||p) \to \min_q$ для распределения $p(\boldsymbol{x})$, известного с точностью до нормировочной константы Z_p. Заметим, что из-за неизвестности Z_p значение оптимизируемого функционала $\mathrm{KL}(q||p)$ не может быть вычислено напрямую. Для произвольного распределения $q(\boldsymbol{x})$ верна следующая цепочка равенств:

$$\log Z_p = \int q(\boldsymbol{x}) \log Z_p d\boldsymbol{x} = \int q(\boldsymbol{x}) \log \frac{\tilde{p}(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} = \int q(\boldsymbol{x}) \log \frac{\tilde{p}(\boldsymbol{x})}{q(\boldsymbol{x})} \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x} =$$

$$= \underbrace{\int q(\boldsymbol{x}) \log \frac{\tilde{p}(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}}_{\mathcal{L}\{q\}} \underbrace{- \int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}}_{\mathrm{KL}(q||p)} = \mathcal{L}\{q\} + \mathrm{KL}(q||p). \quad (2)$$

В силу неотрицательности KL-дивергенции отсюда следует, что $\log Z_p \geq \mathcal{L}\{q\}$, причем равенство достигается тогда и только тогда, когда $q(\boldsymbol{x}) \equiv p(\boldsymbol{x})$. Кроме того, величина $\log Z_p$ не зависит от q. Поэтому

$$\mathrm{KL}(q||p) \to \min_q \;\Leftrightarrow\; \mathcal{L}\{q\} \to \max_q .$$

Таким образом, задача минимизации прямой KL-дивергенции сведена с эквивалентной задаче максимизации функционала $\mathcal{L}\{q\}$. В отличие от $\mathrm{KL}(q||p)$ значение $\mathcal{L}\{q\}$ может быть вычислено, т.к. зависит от известных величин $\tilde{p}(\boldsymbol{x})$ и $q(\boldsymbol{x})$. Кроме того, данное значение после оптимизации является нижней оценкой для $\log Z_p$.

Рассмотрим теперь задачу $\mathcal{L}\{q\} \to \max_q$ для семейства полностью факторизованных распределений q:

$$q(\boldsymbol{x}) = \prod_i q_i(x_i),$$

где x_i – отдельные переменные. Будем искать решение задачи с помощью покоординатной оптимизации, т.е. зафиксируем все факторы q_i за исключением одного фактора q_j и рассмотрим оптимизацию

$\mathcal{L}\{q\}$ по q_j:

$$\mathcal{L}\{q\} = \int \log(\tilde{p}(\boldsymbol{x})) \prod_i q_i(x_i) dx_i - \int \log\left(\prod_i q_i(x_i)\right) \prod_k q_k(x_k) dx_k =$$

$$= \int \underbrace{\left[\int \log(\tilde{p}(\boldsymbol{x})) \prod_{i\neq j} q_i(x_i) dx_i\right]}_{\log \tilde{r}_j(x_j)} q_j(x_j) dx_j - \sum_i \int \log(q_i(x_i)) \prod_k q_k(x_k) dx_k =$$

$$= \int \log(\tilde{r}_j(x_j)) q_j(x_j) dx_j - \int \log(q_j(x_j)) q_j(x_j) dx_j \underbrace{- \sum_{i\neq j} \int \log(q_i(x_i)) q_i(x_i) dx_i}_{\text{не зависит от } q_j(x_j)} =$$

$$= \underbrace{\int q_j(x_j) \log \frac{\tilde{r}_j(x_j)}{q_j(x_j)} dx_j}_{\mathcal{L}_j\{q_j\}} + \text{const} \to \max_{q_j}.$$

Здесь через $\tilde{r}_j(x_j)$ обозначена неотрицательная величина $\exp\left(\int \log(\tilde{p}(\boldsymbol{x})) \prod_{i\neq j} q_i(x_i) dx_i\right)$. Максимизация функционала $\mathcal{L}_j\{q_j\}$ эквивалентна минимизации $\mathrm{KL}(q_j||r_j)$ для распределения $r_j(x_j) \propto \tilde{r}_j(x_j)$. Таким образом,

$$q_j(x_j) = r_j(x_j) = \frac{\exp\left(\int \log(\tilde{p}(\boldsymbol{x})) \prod_{i\neq j} q_i(x_i) dx_i\right)}{\int \exp\left(\int \log(\tilde{p}(\boldsymbol{x})) \prod_{i\neq j} q_i(x_i) dx_i\right) dx_j}. \tag{3}$$

Итерационный пересчёт по формуле (3) продолжается до сходимости функционала

$$\mathcal{L}\{q\} = \int \log(\tilde{p}(\boldsymbol{x})) \prod_i q_i(x_i) dx_i - \sum_i \int \log(q_i(x_i)) q_i(x_i) dx_i.$$

Заметим, что в процессе итераций значение $\mathcal{L}\{q\}$ монотонно не убывает. С учетом ограничения сверху $\mathcal{L}\{q\} \leq \log Z_p$ итерационный процесс гарантированно сходится. Можно показать, что все приведённые рассуждения остаются в силе и для семейства факторизованных распределений

$$q(\boldsymbol{x}) = \prod_i q_i(\boldsymbol{x}_i),$$

где $\boldsymbol{x}_i$ – подмножество переменных $\boldsymbol{x}$, и подмножества не пересекаются. Представленный способ поиска аппроксимации для распределения p с помощью q из факторизованного семейства получил название вариационного вывода[3].

Задача оптимизации $\mathcal{L}\{q\} \to \max_q$ в семействе произвольных вероятностных распределений q, а также задача $\mathcal{L}_j(q_j) \to \max_{q_j}$ для произвольного распределения q_j являются выпуклыми. Их решения единственны и равны, соответственно, $q(\boldsymbol{x}) = p(\boldsymbol{x})$ и $q_j(x_j) = r_j(x_j)$. Однако, задача $\mathcal{L}\{q\} \to \max_q$ в семействе факторизованных распределений q не обладает свойством выпуклости и, как следствие, может иметь множество локальных оптимумов. Поэтому на практике итерационный пересчёт (3) обычно запускается из нескольких случайных начальных приближений с выбором лучшего по максимуму итогового значения $\mathcal{L}\{q\}$.

В рамках графических моделей исходное распределение p представляется в виде произведения факторов

$$p(\boldsymbol{x}) = \frac{1}{Z_p} \tilde{p}(\boldsymbol{x}) = \frac{1}{Z_p} \prod_k f_k(\boldsymbol{x}_k),$$

где $\boldsymbol{x}_k$ – подмножество переменных $\boldsymbol{x}$, различные подмножества могут пересекаться. Рассмотрим применение вариационного вывода к распределению $p(\boldsymbol{x})$. Будем искать аппроксимацию для p в

[3]Такое название было дано из-за того, что здесь фактически была аналитически решена задача вариационной оптимизации функционала $\mathcal{L}\{q\}$ по распределению q.

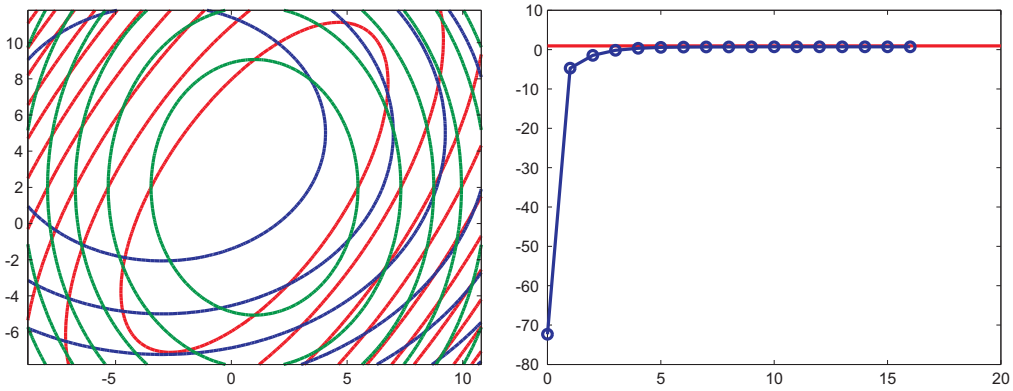

Рис. 2: Пример поиска факторизованного вариационного приближения для двухмерного нормального распределения. Слева показаны линии уровня исходного распределения (красные линии), начальное факторизованное приближение (синие линии) и итоговое факторизованное приближение (зелёные линии). Справа показано истинное значение логарифма нормировочной константы (красная прямая), а также значение её нижней оценки $\mathcal{L}\{q\}$ по итерациям вариационного приближения.

семействе полностью факторизованных распределений

$$q(\boldsymbol{x}) = \prod_i q_i(x_i),$$

где x_i – отдельные переменные. Применяя общую формулу (3), получаем

$$q_j(x_j) \propto \exp\left(\int \log \tilde{p}(\boldsymbol{x}) \prod_{i\neq j} q_i(x_i) dx_i\right) \propto \prod_{k: x_j \in \boldsymbol{x}_k} \exp\left(\int \log f_k(\boldsymbol{x}_k) \prod_{\substack{i\neq j \\ x_i \in \boldsymbol{x}_k}} q_i(x_i) dx_i\right).$$

Таким образом, многомерный интеграл по $\boldsymbol{x}$ в данном случае разбивается на произведение интегралов, каждый из которых зависит только от небольшого подмножества переменных $\boldsymbol{x}_k$.

Пример: факторизованное нормальное распределение

Рассмотрим в качестве примера применения вариационного вывода приближение многомерного нормального распределения

$$p(\boldsymbol{x}) = \frac{1}{Z_p}\tilde{p}(\boldsymbol{x}) = \frac{1}{Z_p}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

с помощью факторизованного распределения

$$q(\boldsymbol{x}) = \prod_i q_i(\boldsymbol{x}_i).$$

Здесь $Z_p = \sqrt{2\pi}^d/\sqrt{\det\Lambda}$, $\Lambda = \Sigma^{-1}$ – матрица точности, набор переменных $\boldsymbol{x}_i$ для различных i не пересекается.

Применяя формулу (3), получаем

$$\log q_j(\boldsymbol{x}_j) = \mathbb{E}_{\prod_{i\neq j} q_i}\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Lambda(\boldsymbol{x}-\boldsymbol{\mu})\right) + \text{const} = \mathbb{E}_{\prod_{i\neq j} q_i}\left(-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_j)^T\Lambda_{jj}(\boldsymbol{x}_j-\boldsymbol{\mu}_j) - \right.$$

$$\left. -(\boldsymbol{x}_j-\boldsymbol{\mu}_j)^T\sum_{i\neq j}\Lambda_{ji}(\boldsymbol{x}_i-\boldsymbol{\mu}_i)\right) + \text{const} = -\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_j)^T\Lambda_{jj}(\boldsymbol{x}_j-\boldsymbol{\mu}_j) - (\boldsymbol{x}_j-\boldsymbol{\mu}_j)^T\sum_{i\neq j}\Lambda_{ji}(\mathbb{E}_{q_i}\boldsymbol{x}_i-\boldsymbol{\mu}_i) + \text{const} =$$

$$= -\frac{1}{2}\boldsymbol{x}_j^T\Lambda_{jj}\boldsymbol{x}_j + \boldsymbol{x}_j^T(\Lambda_{jj}\boldsymbol{\mu}_j - \sum_{i\neq j}\Lambda_{ji}(\mathbb{E}_{q_i}\boldsymbol{x}_i - \boldsymbol{\mu}_i)) + \text{const}. \quad (4)$$

Здесь через const обозначена нормировочная константа распределения $q_j(\boldsymbol{x}_j)$, не зависящая от $\boldsymbol{x}_j$. Выражение (4) имеет вид квадратичной функции от $\boldsymbol{x}_j$. Следовательно, распределение q_j является нормальным со следующими параметрами:

$$\begin{aligned} &q_j(\boldsymbol{x}_j) = \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{m}_j, S_j), \\ &S_j = \Lambda_{jj}^{-1}, \quad \boldsymbol{m}_j = \boldsymbol{\mu}_j - \sum_{i\neq j} \Lambda_{jj}^{-1}\Lambda_{ji}(\boldsymbol{m}_i - \boldsymbol{\mu}_i). \end{aligned} \tag{5}$$

Найдем теперь значение оптимизируемого критерия $\mathcal{L}\{q\}$:

$$\mathcal{L}\{q\} = \mathbb{E}_q \log\tilde{p}(\boldsymbol{x}) - \mathbb{E}_q \log q = \mathbb{E}_q\left(-\frac{1}{2}\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_i)^T\Lambda_{ii}(\boldsymbol{x}_i - \boldsymbol{\mu}_i) - \sum_{i\neq k}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\Lambda_{ik}(\boldsymbol{x}_k - \boldsymbol{\mu}_k)\right) -$$

$$-\sum_i \mathbb{E}_{q_i}\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m}_i)^T S_i^{-1}(\boldsymbol{x}_i - \boldsymbol{m}_i) - \frac{1}{2}\log\det S_i - \frac{d_i}{2}\log 2\pi\right) = \mathbb{E}_q\left(-\frac{1}{2}\sum_i (\boldsymbol{x}_i^T\Lambda_{ii}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\Lambda_{ii}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T\Lambda_{ii}\boldsymbol{\mu}_i)\right) -$$

$$-\sum_{i\neq k}(\mathbb{E}_{q_i}\boldsymbol{x}_i - \boldsymbol{\mu}_i)^T\Lambda_{ik}(\mathbb{E}_{q_k}\boldsymbol{x}_k - \boldsymbol{\mu}_k) - \sum_i\left(-\frac{1}{2}\operatorname{tr} S_i^{-1}S_i - \frac{1}{2}\log\det S_i - \frac{d_i}{2}\log 2\pi\right) =$$

$$= -\frac{1}{2}\sum_i \operatorname{tr}\Lambda_{ii}S_i - \frac{1}{2}(\boldsymbol{m} - \boldsymbol{\mu})^T\Lambda(\boldsymbol{m} - \boldsymbol{\mu}) + \frac{1}{2}\sum_i\left(d_i(\log 2\pi + 1) + \log\det S_i\right). \tag{6}$$

Здесь через d_i обозначена длина вектора $\boldsymbol{x}_i$, а через $\boldsymbol{m}$ – конкатенация мат.ожиданий $\boldsymbol{m}_i$ для всех i. Заметим, что выражение (6) не следует пытаться упрощать дальше, подставляя вместо S_i и $\boldsymbol{m}_i$ их значения из формулы (5). В текущем виде выражение для $\mathcal{L}\{q\}$ является корректным для произвольного факторизованного нормального распределения q. Для экономии вычислительных затрат данное выражение можно вычислять, например, только на каждой 10-й итерации.

На рис. (2) показан пример применения вариационного приближения для нормального распределения в двухмерном случае. Заметим, что факторы $q_i(\boldsymbol{x}_i)$ не являются маргинальными распределениями для $p(\boldsymbol{x})$. В частности, матрица ковариации для маргинала $p(\boldsymbol{x}_i)$ равна Σ_{ii}, а для q_i – Λ_{ii}^{-1}. Можно показать, что $\Lambda_{ii}^{-1} = \Sigma_{ii} - \Sigma_{i,\setminus i}\Sigma_{\setminus i,\setminus i}^{-1}\Sigma_{\setminus i,i}$, где $\Sigma_{\setminus i,\setminus i}$ – это матрица Σ с исключением строк и столбцов, входящих в группу i. Таким образом, дисперсии компонент q_i меньше, чем дисперсии компонент маргиналов. Этот результат согласуется с общим тезисом о том, что при минимизации прямой дивергенции возможна недооценка значений статистик, связанных с разбросом.

Вариационный EM-алгоритм

Вариационный вывод позволяет сформулировать одно из наиболее сильных обобщений EM-алгоритма – вариационный EM-алгоритм. Рассмотрим задачу обучения параметров вероятностной модели со скрытыми переменными $p(X, T|\Theta)$ с помощью метода максимального правдоподобия:

$$p(X|\Theta) = \int p(X, T|\Theta)dT \to \max_\Theta.$$

Здесь X – набор наблюдаемых переменных, T – набор скрытых переменных, а Θ – набор параметров. Применяя общий результат (2) для ненормированного по T распределения $p(X, T|\Theta)$, получаем, что для произвольного распределения $q(T)$ справедливо разложение

$$\log p(X|\Theta) = \int \log p(X, T|\Theta)q(T)dT - \int \log(q(T))q(T)dT + \mathrm{KL}(q||p(T|X, \Theta)) = \mathcal{L}\{q\} + \mathrm{KL}(q||p(T|X, \Theta)).$$

Выполняя покомпонентную максимизацию данного выражения по q и Θ, получаем классические формулы EM-алгоритма:

$$\begin{aligned} &\text{E-шаг: } q(T) = p(T|X, \Theta), \\ &\text{M-шаг: } \mathbb{E}_q \log p(X, T|\Theta) \to \max_\Theta. \end{aligned}$$

В силу того, что после E-шага $\log p(X|\Theta) = \mathcal{L}\{q\}$, значение неполного правдоподобия $p(X|\Theta)$ монотонно не убывает в EM-итерациях. Для многих вероятностных моделей точное вычисление условного распределения $p(T|X, \Theta)$ на E-шаге или, эквивалентно, точная оценка $\mathbb{E}_q \log p(X, T|\Theta)$, требуемая

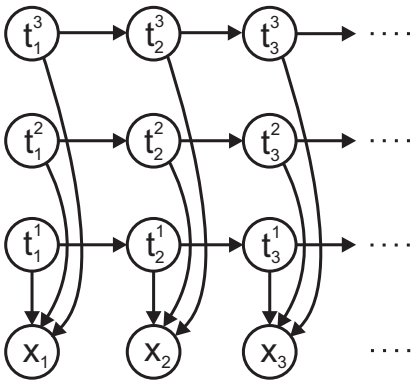

Рис. 3: Граф для факториальной скрытой марковской модели.

на M-шаге, являются недоступными. В этом случае можно рассмотреть задачу максимизации нижней границы $\mathcal{L}\{q\}$ по параметрам Θ и по распределению q в рамках факторизованного семейства распределений:

$$\text{E-шаг: } \mathcal{L}\{q\} \to \max_{q(T)=\prod_i q_i(T_i)},$$
$$\text{M-шаг: } \mathbb{E}_q \log p(X,T|\Theta) \to \max_{\Theta}.$$

Здесь задача оптимизации на E-шаге решается с помощью описанной выше процедуры вариационного вывода. В отличие от классической схемы EM-алгоритма, вариационный EM-алгоритм не гарантирует монотонного увеличения функционала $\log p(X|\Theta)$ в итерациях. В данном случае гарантируется лишь монотонное возрастание нижней оценки правдоподобия $\mathcal{L}\{q\}$.

Факториальная скрытая марковская модель

Рассмотрим в качестве примера применения вариационного EM-алгоритма задачу обучения т.н. факториальной скрытой марковской модели. Данная модель является обобщением классической скрытой марковской модели (СММ) и представляет собой байесовскую сеть с графом, показанном на рисунке 3. Здесь наблюдаемые переменные $\boldsymbol{x}_n \in \mathbb{R}^d$ зависят от нескольких скрытых марковских процессов с дискретными переменными $t_n^m \in \{1,\dots,K\}$:

$$p(X,T|\{\boldsymbol{\pi}^m, A^m\}_{m=1}^M, \{\boldsymbol{\mu}_k^m\}_{k,m=1}^{K,M}, \Sigma) = p(\boldsymbol{x}_1|t_1^1,\dots,t_1^M) \prod_{m=1}^M p(t_1^m) \prod_{n=2}^N p(x_n|t_n^1,\dots,t_n^M) \prod_{m=1}^M p(t_n^m|t_{n-1}^m),$$
$$p(t_1^m) = \pi_{t_1^m}^m, \quad p(t_n^m|t_{n-1}^m) = A_{t_{n-1}^m, t_n^m}^m,$$
$$p(\boldsymbol{x}_n|t_n^1,\dots,t_n^M) = \mathcal{N}(\boldsymbol{x}_n|\sum_{m=1}^M \boldsymbol{\mu}_{t_n^m}^m, \Sigma).$$

Факториальную СММ можно свести к классической СММ, объединив все состояния в каждый момент времени n. В этом случае получится СММ с числом состояний K^M. Сложность алгоритма Витерби и алгоритма «вперёд-назад» для такой модели составляет $O(NK^{2M})$. Уже при относительно небольших значениях K и M такая сложность является запретительной.

Полная факторизация

Рассмотрим решение задачи обучения параметров факториальной СММ

$$p(X|\Theta) \to \max_{\Theta}, \quad \Theta = \{\{\boldsymbol{\pi}^m, A^m\}_{m=1}^M, \{\boldsymbol{\mu}_k^m\}_{k,m=1}^{K,M}, \Sigma\},$$

с помощью вариационного EM-алгоритма, в котором аппроксимант $q(T)$ для апостериорного распределения $p(T|X,\Theta)$ выбирается в семействе полностью факторизованных распределений

$$q(T) = \prod_{m,n=1}^{M,N} q_{mn}(t_n^m).$$

Применяя общий результат вариационного вывода (3), получаем

$$\log q_{mn}(t_n^m) = \mathbb{E}_{q_{\setminus(mn)}} \log p(X,T|\Theta)+\text{const} = \sum_{t_{n-1}^m} \log A^m_{t_{n-1}^m,t_n^m} q_{m,n-1}(t_{n-1}^m)+\sum_{t_{n+1}^m} \log A^m_{t_n^m,t_{n+1}^m} q_{m,n+1}(t_{n+1}^m)-$$

$$-\frac{1}{2}\mathbb{E}_{q_{\setminus(mn)}}(\boldsymbol{x}_n - \boldsymbol{\mu}^m_{t_n^m} - \sum_{l\neq m}\boldsymbol{\mu}^l_{t_n^l})^T\Sigma^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_{t_n^m} - \sum_{l\neq m}\boldsymbol{\mu}^l_{t_n^l}) + \text{const} = \sum_{t_{n-1}^m} \log A^m_{t_{n-1}^m,t_n^m} q_{m,n-1}(t_{n-1}^m)+$$

$$+\sum_{t_{n+1}^m} \log A^m_{t_n^m,t_{n+1}^m} q_{m,n+1}(t_{n+1}^m)-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}^m_{t_n^m})^T\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}^m_{t_n^m})+(\boldsymbol{x}_n-\boldsymbol{\mu}^m_{t_n^m})^T\Sigma^{-1}\left(\sum_{l\neq m}\sum_{t_n^l}\boldsymbol{\mu}^l_{t_n^l}q_{ln}(t_n^l)\right)+\text{const} =$$

$$= B_{mn,t_n^m} + \text{const}.$$

Здесь через B_{mn,t_n^m} обозначена вычисленная величина. С учётом нормировки для q_{mn}, находим

$$q_{mn}(k) = \frac{\exp(B_{mn,k})}{\sum_j \exp(B_{mn,j})}.$$

Формулы пересчёта для крайних компонент q_{m1} и q_{mN} выводятся аналогично. Сложность одной итерации вариационного вывода составляет $O(MNK^2)$, что значительно меньше, чем $O(NK^{2M})$ для случая сведения к классической CMM. Решая задачу максимизации $\mathbb{E}_q \log p(X,T|\Theta)$ по Θ, получаем

$$\pi_k^m = q_{m1}(k), \quad A^m_{kj} = \frac{\sum_{n=2}^N q_{m,n-1}(k)q_{mn}(j)}{\sum_{i=1}^K\sum_{n=2}^N q_{m,n-1}(k)q_{mn}(i)},$$

$$\boldsymbol{\mu}_k^m = \frac{\sum_{n=1}^N q_{mn}(k)(\boldsymbol{x}_n - \sum_{l\neq m}\sum_{j=1}^K \boldsymbol{\mu}^l_j q_{ln}(j))}{\sum_{n=1}^N q_{mn}(k)},$$

$$\Sigma = \frac{1}{N}\sum_{n=1}^N \left[(\boldsymbol{x}_n - \sum_{m,k=1}^{M,K}\boldsymbol{\mu}^m_k q_{mn}(k))(\boldsymbol{x}_n - \sum_{m,k=1}^{M,K}\boldsymbol{\mu}^m_k q_{mn}(k))^T + \sum_{m,k=1}^{M,K}\boldsymbol{\mu}^m_k(\boldsymbol{\mu}^m_k)^T q_{mn}(k)(1-q_{mn}(k))\right].$$

Факторизация по строкам

Аппроксимация $q(T)$ в семействе полностью факторизованных распределений является достаточно грубой. Эта грубость особенно сказывается при решении задачи вывода относительно скрытых переменных T, т.к. здесь фактически игнорируется связь между состояниями в соседние моменты времени. Рассмотрим вариационный EM-алгоритм для обучения параметров факториальной CMM в рамках семейства факторизаций по строкам:

$$q(T) = \prod_{m=1}^M q_m(T^m).$$

Применяя результат (3), получаем

$$\log q_m(T^m) = \mathbb{E}_{q_{\setminus m}} \log p(X,T|\Theta) + \text{const} = \log \pi^m_{t_1^m} + \sum_{n=2}^N \log A^m_{t_{n-1}^m,t_n^m} +$$

$$+\sum_{n=1}^N\left[-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}^m_{t_n^m})^T\Sigma^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}^m_{t_n^m}) + (\boldsymbol{x}_n - \boldsymbol{\mu}^m_{t_n^m})^T\Sigma^{-1}\left(\sum_{l\neq m}\sum_{t_n^l}\boldsymbol{\mu}^l_{t_n^l}q_l(t_n^l)\right)\right] + \text{const}.$$

Здесь через $q_l(t_n^l)$ обозначен маргинал для n-ой переменной в многомерном распределении $q_l(T^l)$. Таким образом,

$$q_m(T^m) \propto \pi^m_{t_1^m} \xi^m_{1,t_1^m} \prod_{n=2}^{N} \left(A^m_{t^m_{n-1}, t^m_n} \xi^m_{n,t^m_n} \right), \tag{7}$$

$$\xi^m_{n,k} = \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}^m_k)^T \Sigma^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}^m_k) + (\boldsymbol{x}_n - \boldsymbol{\mu}^m_k)^T \Sigma^{-1} \left(\sum_{l \neq m} \sum_{j=1}^{K} \boldsymbol{\mu}^l_j q_l(t^l_n = j) \right) \right).$$

Результат (7) говорит о том, что $q_m(T^m)$ представляет собой СММ, в которой априорные вероятности равны $\boldsymbol{\pi}^m$, матрица перехода – A^m, а вероятности наблюдений в момент времени n в состоянии k определяются величиной $\xi^m_{n,k}$. С помощью алгоритма «вперёд-назад» в данной СММ можно найти все одномерные и двухмерные маргинальные распределения $q_m(t^m_n)$ и $q_m(t^m_{n-1}, t^m_n)$. Сложность этой операции для всех факторов q_m составляет $O(MNK^2)$, что совпадает со сложностью одной итерации вариационного вывода при использовании полной факторизации.

Решая задачу максимизации $\mathbb{E}_q \log p(X, T|\Theta)$ по Θ, получаем

$$\pi^m_k = q_m(t^m_1 = k), \quad A^m_{kj} = \frac{\sum_{n=2}^{N} q_m(t^m_{n-1} = k, t^m_n = j)}{\sum_{i=1}^{K} \sum_{n=2}^{N} q_m(t^m_{n-1} = k, t^m_n = i)},$$

$$\boldsymbol{\mu}^m_k = \frac{\sum_{n=1}^{N} q_m(t^m_n = k)(\boldsymbol{x}_n - \sum_{l \neq m} \sum_{j=1}^{K} \boldsymbol{\mu}^l_j q_l(t^l_n = j))}{\sum_{n=1}^{N} q_m(t^m_n = k)},$$

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} \left[(\boldsymbol{x}_n - \sum_{m,k=1}^{M,K} \boldsymbol{\mu}^m_k q_m(t^m_n = k))(\boldsymbol{x}_n - \sum_{m,k=1}^{M,K} \boldsymbol{\mu}^m_k q_m(t^m_n = k))^T + \sum_{m,k=1}^{M,K} \boldsymbol{\mu}^m_k (\boldsymbol{\mu}^m_k)^T q_m(t^m_n = k)(1 - q_m(t^m_n = k)) \right].$$

Сравнение вариационного вывода и MCMC

В заключение сравним между собой два метода приближённого вывода в графических моделях: вариационный вывод и методы Монте Карло по схеме марковских цепей (MCMC).

В методе MCMC оценка на $\mathbb{E}_p f(\boldsymbol{x})$ является тем точнее, чем больше конфигураций $\boldsymbol{x}_n$ генерируется. В пределе оценка MCMC является точной. В вариационном выводе нет никаких гарантий на близость между $\mathbb{E}_p f(\boldsymbol{x})$ и $\mathbb{E}_q f(\boldsymbol{x})$. Более того, нетрудно привести пример, когда эти две величины будут сколь угодно далеки друг от друга.

В итерациях вариационного вывода происходит максимизация функционала $\mathcal{L}\{q\}$, который является нижней оценкой для логарифма нормировочной константы $\log Z_p$. Как правило, вариационный вывод обеспечивает достаточно точную оценку на значение нормировочной константы даже в ситуациях, когда используется существенно ограниченное семейство распределений q, например, полностью факторизованное семейство распределений. В методе MCMC нормировочная константа распределения не может быть вычислена напрямую, т.к. она является «нулевой» статистикой.

Время работы одной итерации вариационного вывода для полностью факторизованного семейства распределений и одной итерации схемы Гиббса, как правило, очень близки. На практике для сходимости вариационного вывода часто достаточно несколько десятков итераций, в то время как для надежной оценки статистик по схеме Гиббса требуется несколько тысяч итераций. В результате время работы вариационного вывода в разы меньше, чем время работы схемы Гиббса.