Данг Куинь Ньы
Коган Александра

Семинар 4. EM-алгоритм

17 декабря 2020 г.

Постановка задачи: Сначала вспомним постановку задачи. Мы хотим максимизировать логарифм функции правдоподобия $\ell(x \mid \theta)$, чтобы найти оценку вектора параметров θ. Однако бывает так, что эта функция имеет такой вид, что максимизировать сложно (например, если под логарифмом оказывается сумма). Идея EM-алгоритма состоит в том, чтобы ввести латентные переменные z и использовать совместное распределение $p(X, Z)$ для нахождения $\hat{\theta}$.

EM-алгоритм в общем виде:

- **Инициализация:** задать начальные значения θ^{old}
- **E.1-шаг:** найти условное распределение латентных переменных $p(Z \mid X, \theta^{old})$
- **E.2-шаг:** построить функцию $Q(\theta, \theta^{old}) = \mathbb{E}_{Z|X,\theta}(\ell(x, z \mid \theta) \mid x, \theta^{old})$
- **M-шаг:** максимизировать Q по θ
- Далее повторять E- и M-шаги до сходимости [1].

Обоснование EM-алгоритма: Для обоснования EM-алгоритма и понимания обозначений рассмотрим пример:

Пусть $X_1, X_2, \ldots, X_n \sim f(x \mid \theta)$, где f – какая-то функция плотности, θ – вектор неизвестных параметров этой плотности.

Пусть латентные переменные принимают всего два значения: $Z \in \{0, 1\}$ – с вероятностями $\mathbb{P}(Z = 0) = p_1$, $\mathbb{P}(Z = 1) = 1 - p_1$. Заметим, что

$$\ell(x \mid \theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta) = \sum_{i=1}^{n} \sum_{j=0}^{1} \mathbb{P}(Z = j) \ln f(x_i \mid \theta) = \left[\mathbb{P}(Z = j \mid x_i, \theta) = \frac{f(x_i, Z = j \mid \theta)}{f(x_i \mid \theta)}\right] [2] =$$

$$= \sum_{i=1}^{n} \sum_{j=0}^{1} \mathbb{P}(Z = j) \ln \frac{f(x_i, Z = j \mid \theta)}{\mathbb{P}(Z = j \mid x_i, \theta)} = \sum_{i=1}^{n} \sum_{j=0}^{1} \mathbb{P}(Z = j) \ln \frac{f(x_i, Z = j \mid \theta)\mathbb{P}(Z = j)}{\mathbb{P}(Z = j \mid x_i, \theta)\mathbb{P}(Z = j)} =$$

$$= \sum_{i=1}^{n} \sum_{j=0}^{1} \mathbb{P}(Z = j) \ln \frac{f(x_i, Z = j \mid \theta)}{\mathbb{P}(Z = j)} + \sum_{i=1}^{n} \sum_{j=0}^{1} \mathbb{P}(Z = j) \ln \frac{\mathbb{P}(Z = j)}{\mathbb{P}(Z = j \mid x_i, \theta)} =$$

$$= M(\mathbb{P}(Z = j), \theta) + D_{KL}[\mathbb{P}(Z = j) || \mathbb{P}(Z = j \mid x_i, \theta)]$$

Так как $D_{KL} \geq 0$, то $M(\mathbb{P}(Z = j), \theta)$ является нижней оценкой на логарифм правдоподобия. Идея EM-алгоритма состоит в том, чтобы поочерёдно максимизировать $M(\mathbb{P}(Z = j), \theta)$ по $\mathbb{P}(Z = j)$ (на E-шаге) и по θ (на M-шаге).

[1] например, критерием остановки м.б. $|Q_t - Q_{t-1}| < 1\text{e}{-4}$

[2] по формуле условной вероятности

- **E-шаг** Максимизируем $M(\mathbb{P}(Z=j),\theta)$ по $\mathbb{P}(Z=j)$.

  Так как $\ell(x|\theta)$ не зависит от $\mathbb{P}(Z=j)$, то максимум $M(\mathbb{P}(Z=j),\theta)$ по $\mathbb{P}(Z=j)$ будет достигнут, когда D_{KL} минимальна.

  По определению $D_{KL} \geq 0$, минимальная $D_{KL}(A||B) = 0$ достигается, когда $A = B$. Из этого делаем вывод, что на E-шаге мы устанавливаем

  $$\mathbb{P}(Z=j) := \mathbb{P}(Z=j \mid x_i, \theta^{old}).$$

- **M-шаг** Максимизируем $M(\mathbb{P}(Z=j),\theta)$ по θ. Распишем M ещё раз:

  $$M = \sum_{i=1}^{n}\sum_{j=0}^{1} \mathbb{P}(Z=j)\ln\frac{f(x_i, Z=j \mid \theta)}{\mathbb{P}(Z=j)}$$

  Заметим, что знаменатель подлогарифмического выражения не зависит от θ. Выбросим его[3] и заменим $\mathbb{P}(Z=j)$ на результат, полученный нами на E-шаге:

  $$M = \sum_{i=1}^{n}\sum_{j=0}^{1} \mathbb{P}(Z=j \mid x_i, \theta^{old})\ln f(x_i, Z=j \mid \theta) = \sum_{i=1}^{n} \mathbb{E}_{Z|x_i,\theta^{old}}(\ln f(x_i, Z=j \mid \theta)) := Q(\theta, \theta^{old}).$$

  Далее мы максимизируем Q по по θ, обновляем θ на аргмаксимум Q, и возвращаемся к E-шагу.

Обозначения:
Теперь соотнесём обозначения из общей постановки EM-алгоритма с теми, что мы получили в примере.

- $p(Z)$ – это безусловное распределение Z. По сути, это массив размера $1 \times k$, где k – число значений Z.
- $p(Z \mid X, \theta^{old})$ – это условное распределение Z при условии выборки. По сути, это массив размера $n \times k$, где n – размер выборки, k – число значений Z. Каждая строчка есть вектор вероятностей того, что на данном наблюдении Z равно соответсвующему значению.
- $\mathbb{E}_{Z|X,\theta}(\cdot)$ – это сумма матожиданий $\sum_{i=1}^{n} \mathbb{E}_{Z|x_i,\theta}(\cdot)$ по всем наблюдениям выборки.
- $p(\cdot)$ – распределение того, что стоит в скобках. Эта функция может оказаться функцией вероятности или функцией плотности в зависимости от контекста.

[3] Функция будет максимизироваться по θ, поэтому независящий от θ знаменатель не нужно учитывать

Задача о кластеризации (разделение смесей)

Пусть мы точно знаем, что наблюдения принадлежат одному из двух кластеров. Пусть в первом кластере наблюдения берутся из нормального $\mathcal{N}(\mu_1, \sigma_1^2)$ распределения, а во втором – из нормального $\mathcal{N}(\mu_2, \sigma_2^2)$ распределения. Предположим, что все наблюдения независимы, и вероятность того, что наблюдение относится к первому кластеру, равна p_1.

Решение:

1. **Определение латентных переменных:** $z \in \{1, 2\}$ – номер кластера. Вспомним, что у нас есть обозначение $p_1 = \mathbb{P}(z = 1)$ – вероятность отнести наблюдение к первому кластеру.

2. **E-1 шаг:** Найти условное распределение латентных переменных $p(Z \mid X, \theta^{old})$. Для данной задачи это означает, что мы должны получить массив $n \times 2$ (n наблюдений, 2 значения z). Нам достаточно найти только один столбец массива – этот столбец будет размера $n \times 1$ (то есть вероятность того, что $z = 1$) – потому что второй столбец определяется однозначно как (1− первый столбец). По формуле условной вероятности:

$$p(z \mid x, \theta^{old}) = \frac{p(z, x \mid \theta^{old})}{p(x \mid \theta^{old})}$$

и

$$p(x \mid z, \theta^{old}) = \frac{p(x, z \mid \theta^{old})}{p(z)}$$

Отметим, что маленькой буквой p обозначается распределение. Мы можем его пока не знать. В конкретных случаях p может быть функцией плотности для непрерывной случайной величины или функцией вероятности для дискретных элементов.

Тогда

$$\mathbb{P}(z_i = 1 \mid x_i, \theta^{old}) = \frac{p(z_i = 1, x_i \mid \theta^{old})}{f(x_i \mid \theta^{old})}$$

Заметим, что для данной задачи мы предполагаем, что x пришли из нормальных распределений, следовательно в знаменателе p заменяется на f - функцию плотности. В числителе пока остается p, потому что конкретный вид распределения пока неизвестен. Конкретные значения z неизвестны (мы сами их придумали и не наблюдаем), поэтому числитель надо преобразовать, перейдя к чему-то известному. Для этого снова воспользуемся формулой условной вероятности, написанной выше: $p(x \mid z, \theta^{old}) = \dfrac{p(x, z \mid \theta^{old})}{p(z)}$.
Из нее получим, что $p(x, z = 1 \mid \theta^{old}) = p(x \mid z = 1, \theta^{old})\mathbb{P}(z = 1)$, где оба множителя

мы знаем: $p(x \mid z, \theta^{old})$ - это функция плотности, а $\mathbb{P}(z = 1) = p_1$ было определено выше. Получаем числитель. Чтобы посчитать знаменатель, воспользуемся знанием о том, что x приходят из 2х распределений и распишем по формуле полной вероятности. Получим:

$$\mathbb{P}(z_i = 1 \mid x_i, \theta^{old}) = \frac{f(x_i \mid z_i = 1, \theta^{old})p_1}{p_1 f(x_i \mid z_i = 1, \theta^{old}) + (1 - p_1)f(x_i \mid z_i = 2, \theta^{old})}$$

3. **E-2 шаг:** Постройте функцию $Q(\theta, \theta^{old}) = \mathbb{E}_{Z|X,\theta}(\ell(x, z \mid \theta) \mid x, \theta^{old})$. По формуле условной вероятности:

$$p(x, z \mid \theta) = p(x \mid \theta, z) \cdot p(z)$$

Легко видеть, что

$$Q(\theta, \theta^{old}) = \sum_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old})[\ln f(x_i \mid \theta) + \ln p_1] + (1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old}))[\ln f(x_i \mid \theta) + \ln(1 - p_1)]$$

Подробно об этом есть в Обосновании EM-алгоритма. Если коротко объяснить формулу: матожидание берется от логарифма совместного правдоподобия x, z, его получаем по формуле условной вероятности, поэтому возникают $[\ln f(x_i \mid \theta) + \ln p_1]$ и $[\ln f(x_i \mid \theta) + \ln(1 - p_1)]$. Матожидание берется по распределению латентных переменных и θ^{old}, поэтому возникают $\mathbb{P}(z_i = 1 \mid x, \theta^{old})$ и $(1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old}))$.

4. **M-шаг:** Выведем формулы для максимизации Q.

Для этого надо найти производные Q по параметрам θ $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_1)$, приравнять производные к 0, найти точки экстремума и доказать, что это точки максимума (для этого рассмотреть вторые производные Q в этих точках) Чтобы не загружать конспект, опустим проверку вторых производных.

$$Q'_{\mu_1} = \sum_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old}) \frac{(x_i - \mu_1)}{\sigma_1^2}$$

$$\mu_1^{new} = \frac{\sum\limits_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old}) x_i}{\sum\limits_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old})}$$

Аналогично,

$$\mu_2^{new} = \frac{\sum\limits_{i=1}^{n} (1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old})) x_i}{\sum\limits_{i=1}^{n} (1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old}))}$$

Далее:

$$Q'_{\sigma_1^2} = \sum_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old}) \left(-\frac{1}{2\sigma_1^2} + \frac{1}{2} \frac{(x_i - \mu)^2}{\sigma_1^4} \right)$$

$$\sigma_1^{2,new} = \frac{\sum\limits_{i=1}^{n} (x_i - \mu)^2 \mathbb{P}(z_i = 1 \mid x, \theta^{old})}{\sum\limits_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old})}$$

Аналогично,

$$\sigma_2^{2,new} = \frac{\sum\limits_{i=1}^{n} (x_i - \mu)^2 (1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old}))}{\sum\limits_{i=1}^{n} (1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old}))}$$

Далее:

$$Q'_{p_1} = \sum_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old}) \frac{1}{p_1} - (1 - \mathbb{P}(z_i = 1 \mid x, \theta^{old})) \frac{1}{1 - p_1}$$

$$p_1^{new} = \frac{\sum\limits_{i=1}^{n} \mathbb{P}(z_i = 1 \mid x, \theta^{old})}{n}$$

Интуиция за полученными формулами:

- μ_1 входит только в функцию плотности нормального распределения, поэтому оценка μ_1 будет очень похожа на оценку маскимального правдоподобия для μ_1, только скорректированная на сумму вероятностей латентных переменных. Для μ_2 аналогично.
- Оценка σ_1^2 тоже выглядит, как оценка максимального правдоподобия, скорректированная относительно z. Для σ_2^2 аналогично.
- Для p_1 оценка получилась вполне логичной - средняя вероятность того, что латентная переменная принимает значение 1.

$\mu_1^{new}, \mu_2^{new}, \sigma_1^{2,new}, \sigma_2^{2,new}, p_1^{new}$ - новые значения параметров после одного шага алгоритма.

После этого $\theta^{old} := [\mu_1^{new}, \mu_2^{new}, \sigma_1^{2,new}, \sigma_2^{2,new}, p_1^{new}]$ и шаг повторяется. (Вновь считаются $\mathbb{P}(z_i = 1 \mid x_i, \theta^{old})$, пересчитываются $\mu_1^{new}, \mu_2^{new}, \sigma_1^{2,new}, \sigma_2^{2,new}, p_1^{new}$)

Условиями остановки алгоритма можно установить, например, максимальное количество шагов, или минимальную разницу Q на 2х последовательных шагах алгоритма, или их комбинацию, или что-то еще.[4]

[4]Пример реализации и работы EM-алгоритма см. в кодспекте семинара