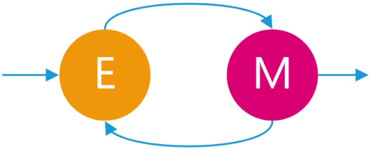
EM алгоритм

Виктор Китов

victorkitov.github.io

Курс поддержан фондом 'Интеллект'



Победитель конкурса VK среди курсов по IT

Содержание

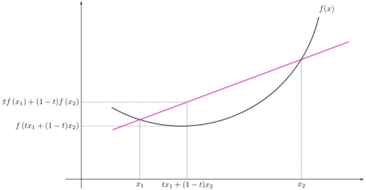
Строго выпуклые функции

- Множество X выпукло, если $\forall x, y \in X$, $\forall \alpha \in (0,1)$:

$$\alpha x + (1-\alpha)y \in X$$

- Функция $f(x)$ строго выпукла на выпуклом X, если $\forall \alpha \in (0,1)$, $\forall x_1 \neq x_2 \in X$:

$$f\left(\alpha x_1 + (1-\alpha)\, x_2\right) < \alpha f(x_1) + (1-\alpha) f(x_2)$$



- Что можно сказать о минимумах выпуклых/строго выпуклых ф-ций и достаточном условии минимума?

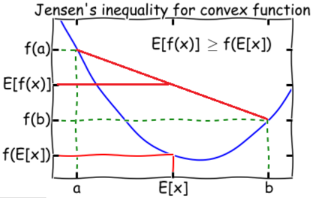
Признаки и свойства[1]

- $f(x)$ строго выпукла <=> она всегда выше касательной

$$f(y) > f(x) + \nabla f(x)^T (y - x) \quad \forall y \neq x \in X$$

- Если $\nabla^2 f(x) \succ 0 \quad \forall x \in X$, то $f(x)$ - строго выпукла на X.
- Если $f(x)$ - строго выпукла, то выполнено нер-во Йенсена

$$\mathbb{E}[f(X)] > f(\mathbb{E}X) \quad \forall X \overset{\text{п.в.}}{\neq} const$$



[1] Докажите утверждения. Верны ли они в обратную сторону?

Доказательство неравенства Йенсена

Для строго выпуклой $f(x)$:

$$f(x) > f(y) + \nabla f(y)^T (x - y)$$

в частности, для не константной X подставим $x = X$ и $y = \mathbb{E}X$

$$f(X) > f(\mathbb{E}X) + \nabla f(\mathbb{E}X)^T (X - \mathbb{E}X)$$

$$\mathbb{E}: \quad \mathbb{E}f(X) > f(\mathbb{E}X) + \nabla f(\mathbb{E}X)^T (\mathbb{E}X - \mathbb{E}X) = f(\mathbb{E}X)$$

Для $X \overset{\text{п.в.}}{=} \mathbb{E}X$:

$$f(X) = f(\mathbb{E}X)$$

$$\mathbb{E}: \quad \mathbb{E}f(X) = \mathbb{E}f(\mathbb{E}X) = f(\mathbb{E}X)$$

Содержание

Вероятностная модель

Рассмотрим вероятностную модель с наблюдаемыми переменными x и ненаблюдаемыми (латентными) переменными z.

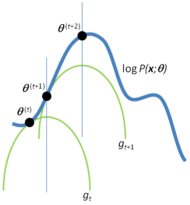
- обозначим $X = [x_1, x_2, ...x_N]$, и $Z = [z_1, z_2, ...z_M]$.

Для нахождения $\widehat{\theta}$ решим:

$$L(\theta) = \ln p_\theta(X) = \ln \sum_Z p_\theta(X, Z) \to \max_\theta$$

- Решение $p_\theta(X, Z) \to \max_\theta$ не применимо, т.к. не знаем Z.
- Оценим распределение $Z \sim q(Z)$, зная X, и решим $\mathbb{E}_Z p_\theta(X, Z) \to \max_\theta$.
- Повторять до сходимости (EM алгоритм):
  - E шаг: оценить, как распределено Z при $\widehat{\theta}$
  - M шаг: максимизировать $\ln p_\theta(X, Z)$, усредненное по вариантам Z.

Общая идея EM алгоритма



$L(\theta) \geq G(q(Z), \theta)\ \forall q(Z), \forall\theta$ (G-нижняя граница L $\forall\theta$ и $\forall q(Z)$)

- Инициализировать $\widehat{\theta}_0$ случайно, $t = 0$
- Повторять до сходимости:
  1. Выбрать $q(Z|\widehat{\theta}_t)$ так, что $L(\widehat{\theta}_t) = G(q(Z), \widehat{\theta}_t)$
  2. $\widehat{\theta}_{t+1} = \arg\max_\theta G(q(Z|\widehat{\theta}_t), \theta)$
  3. $t = t + 1$

Комментарии

- Е-шаг: $G(q(Z), \widehat{\theta}_t) = \arg\max_{p(Z)} G(p(Z), \widehat{\theta}_t)$
- М-шаг: $\widehat{\theta}_{t+1} = \arg\max_{\theta} G(q(Z), \theta)$
- EM алгоритм - метод покоординатного подъема нижней границы $G(p(Z), \theta)$.
- $L(\widehat{\theta}_t)$ сходится, т.к.
    1. $L(\widehat{\theta}_t) = G(q(Z), \widehat{\theta}_t) \le G(q(Z), \widehat{\theta}_{t+1}) \le L(\widehat{\theta}_{t+1})$ => $\left\{L(\widehat{\theta}_t)\right\} \uparrow$
    2. $\left\{L(\widehat{\theta}_t)\right\}$ ограничена сверху, т.к. $L(\theta) = \ln p(X|\theta) \le \ln 1)$

Вывод нижней оценки

Пусть $q(Z)$ - некоторое распределение над Z, $q(Z) \geq 0$, $\sum_Z q(Z) = 1$. Тогда

$$
\begin{aligned}
L(\theta) = \ln p_\theta(X) &= \ln \sum_Z p_\theta(X, Z) \\
&= \ln \sum_Z q(Z) \frac{p_\theta(X, Z)}{q(Z)} && (1) \\
&\geq \sum_Z q(Z) \ln \frac{p_\theta(X, Z)}{q(Z)} = G(q(Z), \theta) && (2)
\end{aligned}
$$

Использовали неравенство Йенсена $f(\mathbb{E}U) \geq \mathbb{E}(fU)$ $\forall$ сл.вел. U и вогнутой f.

1. $f(x) = \ln x$ вогнута, т.к. $(\ln x)'' = -\frac{1}{x^2} < 0$
2. сл. вел. U: $p\left(U = \frac{p(X,Z,\theta)}{q(Z)}\right) = q(Z)$ для всевозможных Z.

Е-шаг: делаем нижнюю грань точной при $\widehat{\theta}_t$

- Неравенство Йенсена: $f(\mathbb{E}U) \ge \mathbb{E}(fU)$, при этом $f(\mathbb{E}U) = \mathbb{E}(fU) <=> U \overset{\text{п.в.}}{=} c = const$:
- $L(\widehat{\theta}_t) = G(q(Z), \widehat{\theta}_t)$ при

$$U = \frac{p_{\widehat{\theta}_t}(X, Z)}{q(Z)} = c \quad \forall Z$$

$$cq(Z) = p_{\widehat{\theta}_t}(X, Z)$$

$$c \sum_Z q(Z) = \sum_Z p_{\widehat{\theta}_t}(X, Z)$$

$$c = p_{\widehat{\theta}_t}(X)$$

$$q(Z) = \frac{p_{\widehat{\theta}_t}(X, Z)}{p_{\widehat{\theta}_t}(X)} = p_{\widehat{\theta}_t}(Z|X)$$

М-шаг: усредненный $\log$(правдоподобия)$\to \max$

М-шаг: усредненный $\log$(правдоподобия)$\to \max$

$$
\begin{aligned}
\hat{\theta}_{t+1} &= \arg\max_{\theta}\{\sum_{Z} q(Z)\ln\frac{p_{\theta}(X,Z)}{q(Z)}\} \\
&= \arg\max_{\theta}\{\sum_{Z} q(Z)\ln p_{\theta}(X,Z) - \overbrace{\sum_{Z} q(Z)\ln q(Z)}^{const(\theta)}\} \\
&= \arg\max_{\theta}\{\sum_{Z} q(Z)\ln p_{\theta}(X,Z)\} \\
&= \arg\max_{\theta}\{\mathbb{E}_{Z\sim q(Z)}\ln p_{\theta}(X,Z)\}
\end{aligned}
$$

Замечание: от θ зависит лишь $p_{\theta}(X,Z)$, $q(Z)=p_{\widehat{\theta}_t}(Z|X)$ - не зависит

EM алгоритм

ВХОД:

выборка $X = [x_1, ...x_N]$, критерий сходимости

АЛГОРИТМ:

Инициализировать $t = 0$, θ_0 - случайно

ПОВТОРЯТЬ до сходиомти:

Е-шаг: уточнить распределение над латентными переменными:

$$q(Z) = p(Z|X, \hat{\theta}_t)$$

М-шаг: уточнить параметры θ:

$$\hat{\theta}_{t+1} = \arg\max_\theta \{\sum_Z q(Z) \ln p(X, Z|\theta)\}$$

$$t = t + 1$$

ВЫХОД: $\hat{\theta}_{t+1}$

Комментарии по EM алгоритму

- Возможные критерии сходимости:
  - $\left\|\widehat{\theta}_{t+1}-\widehat{\theta}_t\right\|<\varepsilon$
  - $L(\widehat{\theta}_{t+1})-L(\widehat{\theta}_t)<\varepsilon$
  - #итераций>порога
- EM сходится к локальному оптимуму
  - можно перезапустить несколько раз из разных $\widehat{\theta}_0$ и выбрать лучшее решение
- Обобщеный EM алгоритм (generalized EM, GEM)
  - для сходимости достаточно выбрать $\widehat{\theta}_{t+1}$ так, что

$$G(q(Z),\widehat{\theta}_{t+1})>G(q(Z),\widehat{\theta}_t)$$

    - например, сделать один шаг в оптимизации
    - а не решать $\widehat{\theta}_{t+1}=\arg\max_\theta G(q(Z),\theta)$ точно.

Содержание

ЕМ алгоритм с регуляризацией

- Добавим регуляризацию $R(\theta)$ в задачу

$$L(\theta) = \ln p(X|\theta) - \lambda R(\theta) \to \max_\theta$$

  - $R(\theta)$ штрафует сложность
  - нужно вычитать, т.к. $\ln p(X|\theta)$ максимизируется.
  - Байесовская MAP оценка:
    $\ln p(X, \theta) = \ln p(X|\theta)p(\theta) = \ln p(X|\theta) + \underbrace{\ln p(\theta)}_{\lambda R(\theta)} \to \max_\theta$

- Нижняя грань:
  $L(\theta) - \lambda R(\theta) \geq G(q(Z), \theta) - \lambda R(\theta)\ \forall q(Z), \forall\theta$

EM алгоритм с регуляризацией

- **E-шаг**: не меняется (равенство из неравенства Йенсена)

$$q(Z) = p_{\widehat{\theta}_t}(Z|X)$$

- **M-шаг**:

$$\widehat{\theta} = \arg\max_{\theta} \left\{ \mathbb{E}_{Z\sim q(Z)} \ln p_{\theta}(X, Z) - \lambda R(\theta) \right\}$$

Содержание

Е-шаг для независимых (x_n, z_n)

- Рассмотрим частный случай независимых наблюдений $\{(x_n, z_n)\}_{n=1}^N$, x_n - наблюдаемые, z_n - латентные
  - пример: смесь Гауссиан, z_n-#компоненты, x_n-реализация.
- Е-шаг становится:

$$q(Z) = p(Z|X, \theta) = p(z_1|x_1, \theta)...p(z_N|x_N, \theta) = q_1(z_1)...q_N(z_N)$$

$$q_n(z_n) = p(z_n|x_n, \theta)$$

M-шаг для независимых (x_n, z_n)

Для независимых объектов (x_n, z_n):

$$\sum_{Z} q(Z) \ln p(X, Z|\theta) = \sum_{z_1, \dots z_N} q_1(z_1) \dots q_N(z_N) \ln \left\{ \prod_{n=1}^{N} p(x_n, z_n|\theta) \right\}$$

$$= \sum_{z_1, \dots z_N} q_1(z_1) \cdot \dots \cdot q_N(z_N) \cdot \left(\sum_{n=1}^{N} \ln p(x_n, z_n|\theta) \right)$$

$$= \left(\sum_{z_1} q_1(z_1) \right) \cdot \dots \cdot \left(\sum_{z_N} q_N(z_N) \right) \cdot \left(\sum_{n=1}^{N} \ln p(x_n, z_n|\theta) \right)$$

$$= \sum_{n=1}^{N} q_n(z_n) \ln p(x_n, z_n|\theta) \to \max_{\theta}$$