

ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург

18 февраля 2022 г.

Random facts:

- 18 февраля 1268 г. состоялась Раковорская битва, в которой псковский князь Довмонт (пришедший туда из Литвы, обеспечивший независимость Пскова и причисленный к лику святых) разбил объединённые силы рыцарей Ливонского ордена, Дерптского епископства и Датской Эстляндии
- 18 февраля 1861 г. прошла инаугурация Джефферсона Финиса Дэвиса, первого и единственного президента Конфедеративных Штатов Америки
- 18 февраля 1911 г. состоялся первый в истории официальный почтовый авиарейс: Анри Пике доставил 6.5 тысяч писем и 250 открыток из Аллахабада в Нейни, на расстояние 13 км
- 18 февраля 1978 г. на острове Оаху 15 человек вышли на старт комбинации трёх гонок: Waikiki Roughwater Swim (3.9км), Around-Oahu Bike Race (185км) и Honolulu Marathon (42.195км); 12 добрались до финиша, и так появился Ironman
- 18 февраля 1930 г. Клайд Томбо открыл Плутон; с 2006 года он считается карликовой планетой и крупнейшим объектом пояса Койпера

ЕМ В ОБЩЕМ ВИДЕ

- Часто нужно оценивать $p(\mathbf{Z} | \mathbf{X})$ для латентных переменных \mathbf{Z} и данных \mathbf{X} .
- Но это слишком сложно! Один вариант — сэмплировать из $p(\mathbf{Z} | \mathbf{X})$.
- Другой вариант — лапласовские приближения, но тоже нечасто работают.
- Давайте решать в общем виде.

- Вспомним сначала формальное обоснование алгоритма EM.
- Мы решаем задачу максимизации правдоподобия по данным $\mathbf{X} = \{x_1, \dots, x_N\}$.

$$L(\theta | \mathbf{X}) = p(\mathbf{X} | \theta) = \prod p(x_i | \theta)$$

или, что то же самое, максимизации $\ell(\theta | \mathbf{X}) = \log L(\theta | \mathbf{X})$.

- EM может помочь, если этот максимум трудно найти аналитически.

- 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 $\mathbf{Z} = (\mathbf{X}, Y)$ с совместной плотностью

$$p(z | \theta) = p(x, y | \theta) = p(y | x, \theta)p(x | \theta).$$

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

- Заметим, что настоящее правдоподобие $L(\theta) = E_Y [p(\mathbf{X}, Y | \theta) | \mathbf{X}, \theta]$.
- E-шаг алгоритма EM вычисляет условное ожидание (логарифма) полного правдоподобия при условии \mathbf{X} и текущих оценок параметров θ_n :

$$Q(\theta, \theta_n) = E [\log p(\mathbf{X}, Y | \theta) | \mathbf{X}, \theta_n].$$

- 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 \mathbf{X} и текущих оценок параметров θ :

$$Q(\theta, \theta_n) = E[\log p(\mathbf{X}, Y | \theta) | \mathbf{X}, \theta_n].$$

- Условное ожидание – это

$$E[\log p(\mathbf{X}, Y | \theta) | \mathbf{X}, \theta_n] = \int_y \log p(\mathbf{X}, y | \theta) p(y | \mathbf{X}, \theta_n) dy,$$

где $p(y | \mathbf{X}, \theta_n)$ – маргинальное распределение скрытых компонентов данных.

- EM лучше всего применять, когда это выражение легко подсчитать, может быть, даже аналитически.
- Вместо $p(y | \mathbf{X}, \theta_n)$ можно подставить $p(y, \mathbf{X} | \theta_n) = p(y | \mathbf{X}, \theta_n)p(\mathbf{X} | \theta_n)$, 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.

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

$$\begin{aligned}
 \ell(\theta) - \ell(\theta_n) &= \\
 &= \log \left(\int_y p(\mathbf{X} | y, \theta) p(y | \theta) dy \right) - \log p(\mathbf{X} | \theta_n) = \\
 &= \log \left(\int_y p(y | \mathbf{X}, \theta_n) \frac{p(\mathbf{X} | y, \theta) p(y | \theta)}{p(y | \mathbf{X}, \theta_n)} dy \right) - \log p(\mathbf{X} | \theta_n) \geq \\
 &\geq \int_y p(y | \mathbf{X}, \theta_n) \log \left(\frac{p(\mathbf{X} | y, \theta) p(y | \theta)}{p(y | \mathbf{X}, \theta_n)} \right) dy - \log p(\mathbf{X} | \theta_n) = \\
 &= \int_y p(y | \mathbf{X}, \theta_n) \log \left(\frac{p(\mathbf{X} | y, \theta) p(y | \theta)}{p(\mathbf{X} | \theta_n) p(y | \mathbf{X}, \theta_n)} \right) dy.
 \end{aligned}$$

- Получили

$$\begin{aligned}\ell(\theta) &\geq l(\theta, \theta_n) = \\ &= \ell(\theta_n) + \int_y p(y | \mathbf{X}, \theta_n) \log \left(\frac{p(\mathbf{X} | y, \theta)p(y | \theta)}{p(\mathbf{X} | \theta_n)p(y | X, \theta_n)} \right) dy.\end{aligned}$$

- Мы нашли нижнюю оценку на $\ell(\theta)$ везде, касание происходит в точке θ_n .

ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ

- Вариационный вывод: функционалы, производные по функциям... в общем, можно оптимизировать функционалы.
- Для нас это значит, что можно оптимизировать приближение q из какого-то класса к заданному p .
- Пусть есть $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ и $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$.
- Мы знаем $p(\mathbf{X}, \mathbf{Z})$ из модели, хотим найти приближение для $p(\mathbf{Z} | \mathbf{X})$ и $p(\mathbf{X})$.

- Как и в EM:

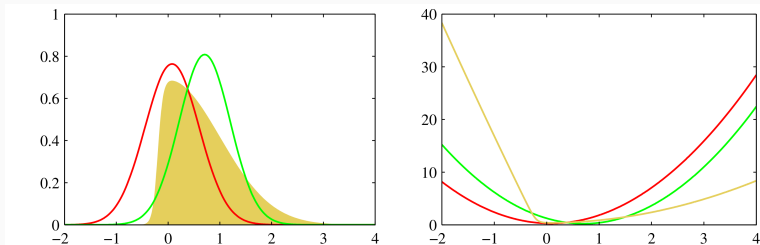
$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \text{KL}(q\|p), \text{ где}$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z},$$

$$\text{KL}(q\|p) = - \int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} | \mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

- $\mathcal{L}(q)$ — это вариационная нижняя оценка, её можно теперь максимизировать, и KL будет автоматически минимизироваться.

- Пример – сравним с лапласовским:



- Если $q(Z)$ произвольное, то мы просто получим $q(Z) = p(\mathbf{Z} | \mathbf{X})$; но это вряд ли получится.
- Будем ограничивать.

- Главный частный случай — пусть $\mathbf{Z} = \mathbf{Z}_1 \sqcup \dots \sqcup \mathbf{Z}_M$, и

$$q(\mathbf{Z}) = \prod_{i=1}^M q_i(\mathbf{Z}_i).$$

- Но больше никаких предположений! В этом прелесть — оптимизируем сразу функции!
- Это соответствует теории среднего поля в физике (mean field theory).

- Тогда:

$$\begin{aligned}
 \mathcal{L}(q) &= \int \prod_i q_i \left(\ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right) d\mathbf{Z} \\
 &= \int q_j \left[\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \right] d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + \text{const} \\
 &= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + \text{const},
 \end{aligned}$$

где $\ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j} [\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$.

- Как максимизировать теперь $\mathcal{L}(q)$ по q_j ?

- Надо заметить, что мы получили там KL-дивергенцию между $q_j(\mathbf{Z}_j)$ и $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$.
- Т.е. оптимальное решение получится при

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E} [\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

- Константа здесь просто для нормализации.
- Оказывается, достаточно взять ожидание от логарифма совместного распределения.
- Но явных формул не получается, потому что ожидание считается по остальным q_i^* , $i \neq j$.
- И всё-таки обычно что-то можно сделать; давайте рассмотрим примеры.

- Первый пример — приблизим двумерный гауссиан одномерными:

$$p(\mathbf{z}) = N(\mathbf{z} \mid \mu, \Lambda^{-1}),$$
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

- И мы хотим приблизить $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$.
- Повычисляем...

- ...получится, что

$$\ln q_1^*(z_1) = -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_{11}\Lambda_{11} - z_1\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) + \text{const.}$$

- Чудесным образом получился гауссиан! Сам собой, без предположений.
- Найдём его среднее и дисперсию...

- ...получится

$$q_1^*(z_1) = N(z_1 \mid m_1, \Lambda_{11}^{-1}), \text{ где}$$

$$m_1 = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (\mathbb{E}[z_2] - \mu_2).$$

- Аналогично,

$$q_2^*(z_2) = N(z_2 \mid m_2, \Lambda_{22}^{-1}), \text{ где}$$

$$m_2 = \mu_2 - \Lambda_{22}^{-1} \Lambda_{21} (\mathbb{E}[z_1] - \mu_1).$$

- Какое решение у этой системы?

- Да просто $E[z_1] = m_1 = \mu_1$, $E[z_2] = m_2 = \mu_2$.
- А если бы мы минимизировали $KL(p\|q)$, получилось бы

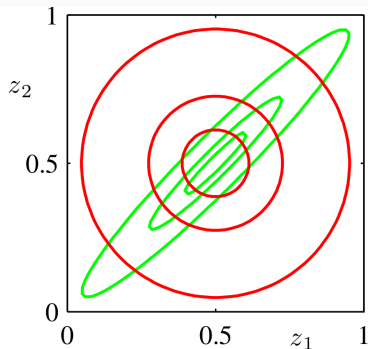
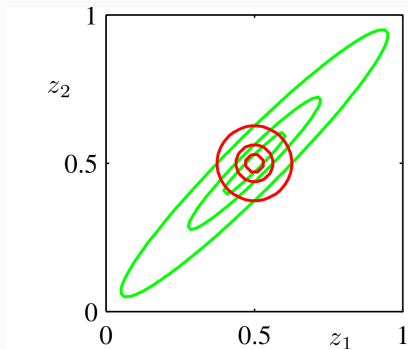
$$KL(p\|q) = - \int p(\mathbf{Z}) \left[\sum_i \ln q_i(\mathbf{Z}_i) \right] d\mathbf{Z} + \text{const},$$

и можно оптимизировать по отдельности:

$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j).$$

- Т.е. просто маргинализация.
- Почему бы так и не сделать? В чём разница?

- Разные дисперсии ответа:

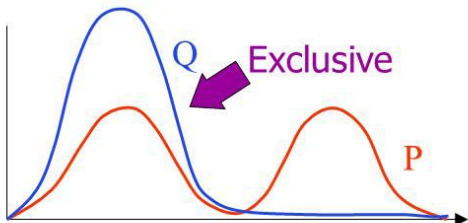


РАЗНЫЕ KL-ДИВЕРГЕНЦИИ

- Минимизация $KL(p||q)$ накрывает всю p , а $KL(q||p)$ строит всю q в регионе больших p :

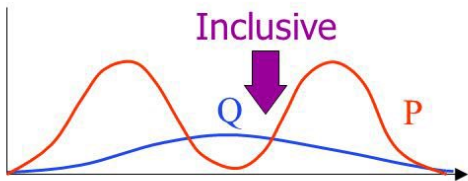
Minimising
 $KL(Q||P)$

$$= \sum_H Q(H) \ln \frac{Q(H)}{P(H|V)}$$



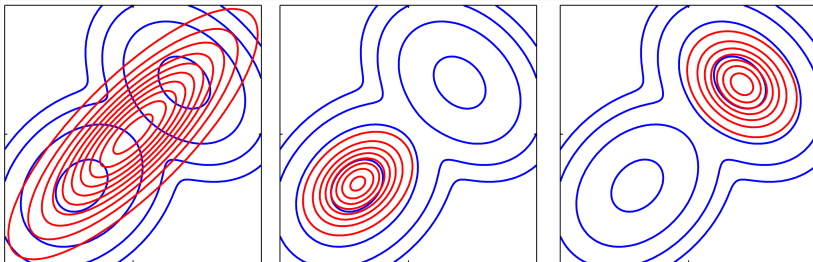
Minimising
 $KL(P||Q)$

$$= \sum_H P(H|V) \ln \frac{P(H|V)}{Q(H)}$$



РАЗНЫЕ KL-ДИВЕРГЕНЦИИ

- Например, для двумерного гауссиана:



- В машинном обучении гораздо интереснее, конечно, пик найти.

Спасибо за внимание!