Вариационные приближения

Сергей Николенко

Академия MADE — Mail.Ru
15 мая 2021 г.

Random facts:

- 15 мая 1157 г. «великий любитель женщин, сладкой пищи и пития» князь Юрий Долгорукий был отравлен на пиру у киевского боярина Петрилы; а 15 мая 1682 г. его потомок князь Юрий Долгоруков был убит взбунтовавшимися стрельцами
- 15 мая 1703 г. милостью короля Георга II первым премьер-министром Великобритании стал лидер партии вигов, лорд-канцлер Роберт Уолпол
- 15 мая 1942 г. на аэродроме Кольцово (Свердловск) Григорий Бахчиванджи испытал первый советский ракетный самолёт БИ-1; а 15 мая 1954 г. из ворот завода в Рентоне (Вашингтон) вышел самолёт-прототип, позже названный Boeing 707
- 15 мая 2001 г. произошёл «CSX 8888 incident»: никем не управляемый поезд #8888 с 42 вагонами (в том числе с опасными химикатами) в течение двух часов двигался со скоростью около 80 км/ч; в 2010 г. об этом вышел фильм *Unstoppable*
- 15 мая 2009 г. президент Таджикистана Эмомали Рахмон запретил чиновникам появляться на одних с ним портретах

ЕМ в общем виде

Приближения

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

Обоснование алгоритма EM

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

$$L(\theta \mid \mathsf{X}) = p(\mathsf{X} \mid \theta) = \prod p(x_i \mid \theta)$$

  или, что то же самое, максимизации $\ell(\theta \mid \mathsf{X}) = \log L(\theta \mid \mathsf{X})$.
- EM может помочь, если этот максимум трудно найти аналитически.

Обоснование алгоритма EM

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

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

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

Обоснование алгоритма EM

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

$$Q(\theta, \theta_n) = E\left[\log p(\mathrm{X}, \mathcal{Y} \mid \theta) \mid \mathrm{X}, \theta_n\right].$$

- 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 X и текущих оценок параметров θ:

$$Q(\theta, \theta_n) = E\left[\log p(\mathsf{X}, \mathcal{Y} \mid \theta) \mid \mathsf{X}, \theta_n\right].$$

- Условное ожидание – это

$$E\left[\log p(\mathsf{X}, \mathcal{Y} \mid \theta) \mid \mathsf{X}, \theta_n\right] = \int_y \log p(\mathsf{X}, y \mid \theta) p(y \mid \mathsf{X}, \theta_n) \mathrm{d}y,$$

  где $p(y \mid \mathsf{X}, \theta_n)$ – маргинальное распределение скрытых компонентов данных.
- EM лучше всего применять, когда это выражение легко подсчитать, может быть, даже аналитически.
- Вместо $p(y \mid \mathsf{X}, \theta_n)$ можно подставить $p(y, \mathsf{X} \mid \theta_n) = p(y \mid \mathsf{X}, \theta_n) p(\mathsf{X} \mid \theta_n)$, 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.

Обоснование алгоритма EM

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

$$\ell(\theta) - \ell(\theta_n) =$$

$$= \log\left(\int_y p(\mathsf{X} \mid y, \theta) p(y \mid \theta) \mathrm{d}y\right) - \log p(\mathsf{X} \mid \theta_n) =$$

$$= \log\left(\int_y p(y \mid \mathsf{X}, \theta_n) \frac{p(\mathsf{X} \mid y, \theta) p(y \mid \theta)}{p(y \mid \mathsf{X}, \theta_n)} \mathrm{d}y\right) - \log p(\mathsf{X} \mid \theta_n) G_{\mathrm{enh}}$$

$$G_{\mathrm{enh}} \int_y p(y \mid \mathsf{X}, \theta_n) \log\left(\frac{p(\mathsf{X} \mid y, \theta) p(y \mid \theta)}{p(y \mid \mathsf{X}, \theta_n)}\right) \mathrm{d}y - \log p(\mathsf{X} \mid \theta_n) =$$

$$= \int_y p(y \mid \mathsf{X}, \theta_n) \log\left(\frac{p(\mathsf{X} \mid y, \theta) p(y \mid \theta)}{p(\mathsf{X} \mid \theta_n) p(y \mid \mathsf{X}, \theta_n)}\right) \mathrm{d}y.$$

Обоснование алгоритма EM

- Получили

$$\ell(\theta) G_{\mathrm{enh}} l(\theta, \theta_n) =$$
$$= \ell(\theta_n) + \int_y p(y \mid \mathsf{X}, \theta_n) \log \left(\frac{p(\mathsf{X} \mid y, \theta) p(y \mid \theta)}{p(\mathsf{X} \mid \theta_n) p(y \mid \mathcal{X}, \theta_n)} \right) \mathrm{d}y.$$

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

Вариационные приближения

Вариационный вывод

- 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 $\mathsf{X} = \{\mathsf{x}_1, \ldots, \mathsf{x}_N\}$ и $\mathsf{Z} = \{\mathsf{z}_1, \ldots, \mathsf{z}_N\}$.
- Мы знаем $p(\mathsf{X}, \mathsf{Z})$ из модели, хотим найти приближение для $p(\mathsf{Z} \mid \mathsf{X})$ и $p(\mathsf{X})$.

Вариационный вывод

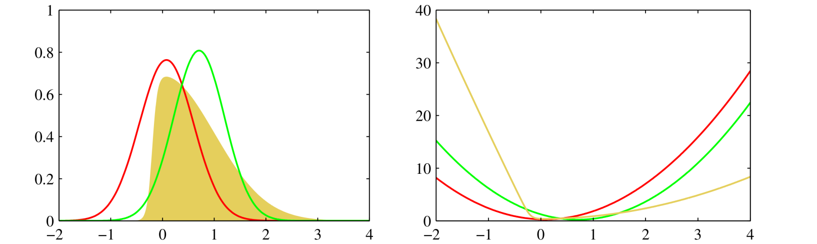
- Как и в EM:

$$\ln p(\mathsf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \text{ где}$$
$$\mathcal{L}(q) = \int q(\mathsf{Z}) \ln \frac{p(\mathsf{X}, \mathsf{Z})}{q(\mathsf{Z})} \mathrm{d}\mathsf{Z},$$
$$\mathrm{KL}(q\|p) = -\int q(\mathsf{Z}) \ln \frac{p(\mathsf{Z} \mid \mathsf{X})}{q(\mathsf{Z})} \mathrm{d}\mathsf{Z}.$$

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

Вариационный вывод

- Пример – сравним с лапласовским:



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

Факторизуемые распределения

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

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

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

Факторизуемые распределения

- Тогда:

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

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

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

Факторизуемые распределения

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

$$\ln q_j^*(\mathsf{Z}_j) = \mathsf{E}\left[\ln p(\mathsf{X}, \mathsf{Z})\right] + \mathrm{const}.$$

- 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.

Двумерный гауссиан

- Первый пример — приблизим двумерный гауссиан одномерными:
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \Lambda^{-1}),$$
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$
- И мы хотим приблизить $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$.
- Повычисляем...

Двумерный гауссиан

- …получится, что

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

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

Двумерный гауссиан

- …получится
$$q_1^*(z_1) = \mathcal{N}(z_1 \mid m_1, \Lambda_{11}^{-1}), \text{ где}$$
$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathsf{E}[z_2] - \mu_2).$$
- Аналогично,
$$q_2^*(z_2) = \mathcal{N}(z_2 \mid m_2, \Lambda_{11}^{-1}), \text{ где}$$
$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathsf{E}[z_1] - \mu_1).$$
- Какое решение у этой системы?

Двумерный гауссиан

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

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

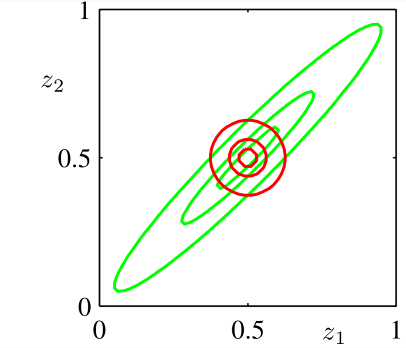
  и можно оптимизировать по отдельности:

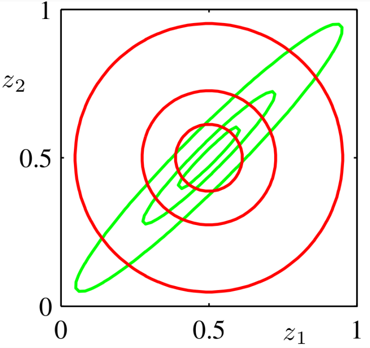
$$q_j^*(\mathsf{Z}_j) = \int p(\mathsf{Z}) \prod_{i \neq j} \mathrm{d}\mathsf{Z}_i = p(\mathsf{Z}_j).$$

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

Разные KL-дивергенции

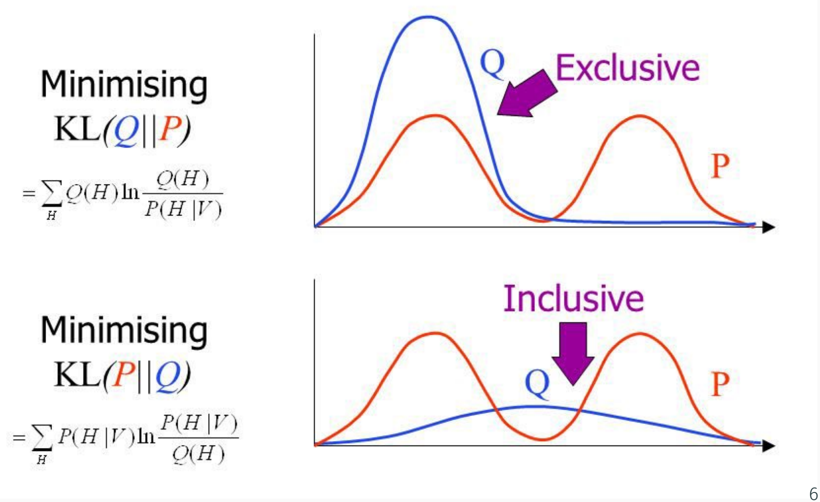
- Разные дисперсии ответа:





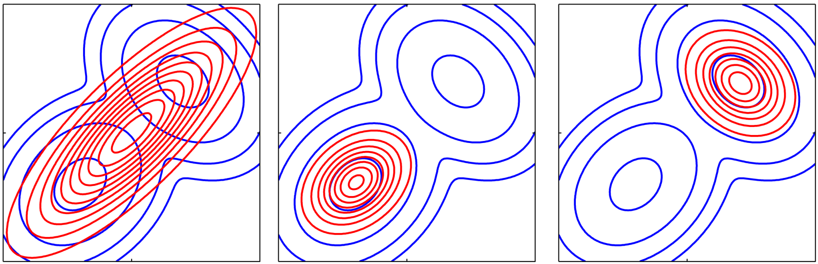
Разные KL-дивергенции

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



Разные KL-дивергенции

- Например, для двумерного гауссиана:



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

Вариационное приближение для гауссиана

Одномерный гауссиан

- И ещё пример: давайте найдём параметры одномерного гауссиана по точкам $\mathsf{X} = \{x_1, \dots, x_N\}$. Правдоподобие:

$$p(\mathsf{X} \mid \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} e^{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2}.$$

- Вводим сопряжённые априорные распределения:

$$p(\mu \mid \tau) = \mathcal{N}(\mu \mid \mu_0, (\lambda_0\tau)^{-1}),$$
$$p(\tau) = \mathrm{Gamma}(\tau \mid a_0, b_0).$$

- Мы это только что подсчитали точно, но давайте приблизим теперь апостериорное распределение как

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau).$$

- На самом деле так не раскладывается!
- Это то, что мы делали для $q(\mathsf{Z}) = \prod_{i=1}^{M} q_i(\mathsf{Z}_i)$. Посчитаем...

Одномерный гауссиан

- ...$q_\mu(\mu)$ – гауссиан с параметрами

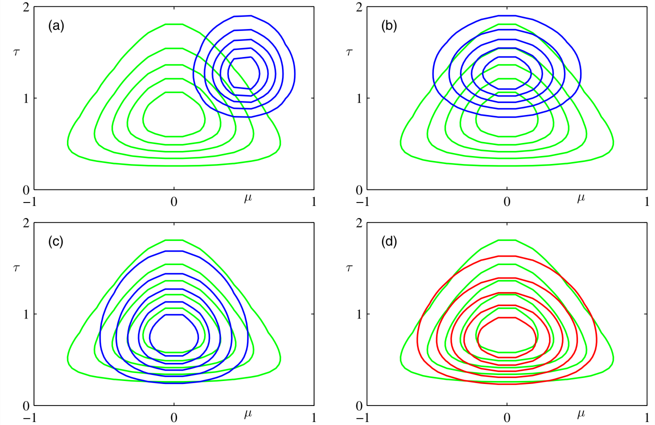
$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N}, \quad \lambda_N = (\lambda_0 + N)\mathsf{E}[\tau].$$

- А $q_\tau(\tau)$ – гамма-распределение с параметрами

$$a_N = a_0 + \frac{N+1}{2}, \quad b_N = b_0 + \frac{1}{2}\mathsf{E}_\mu\left[\sum_n (x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

- Всё получилось как надо, но без предположений о форме q_τ и q_μ.

- Вот такой вывод в пространстве (μ, τ):



- А для $\mu_0 = a_0 = b_0 = \lambda_0 = 0$ (non-informative priors) можно и точно посчитать...

Одномерный гауссиан

- Получатся моменты для μ

$$\mathsf{E}[\mu] = \bar{x}, \quad \mathsf{E}[\mu^2] = \bar{x}^2 + \frac{1}{N\mathsf{E}[\tau]}.$$

- Это можно подставить и найти $\mathsf{E}[\tau]$:

$$\frac{1}{\mathsf{E}[\tau]} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2.$$

Вариационное приближение для смеси гауссианов

Смесь гауссианов

- Смесь гауссианов: $\mathsf{X} = \{\mathsf{x}_1, \dots, \mathsf{x}_n\}$, $\mathsf{Z} = \{\mathsf{z}_1, \dots, \mathsf{z}_n\}$,

$$p(\mathsf{Z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}},$$

$$p(\mathsf{X} \mid \mathsf{Z}, \boldsymbol{\mu}, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(\mathsf{x}_n \mid \boldsymbol{\mu}_k, \Lambda_k^{-1}\right).$$

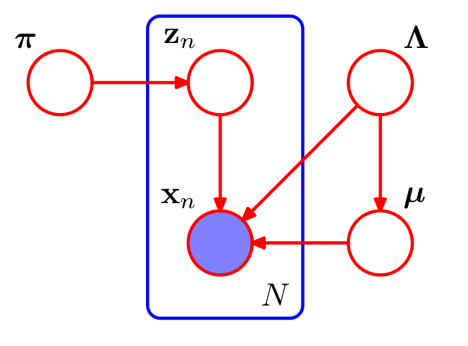
- Выберем сопряжённые априорные распределения:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1},$$

$$\begin{aligned} p(\boldsymbol{\mu}, \Lambda) &= p(\boldsymbol{\mu} \mid \Lambda) p(\Lambda) \\ &= \prod_{k=1}^{K} \mathcal{N}\left(\mu_k \mid \mathsf{m}_0, (\beta_0 \Lambda_k)^{-1}\right) \mathcal{W}(\Lambda_k \mid \mathsf{W}_0, \nu_0). \end{aligned}$$

Смесь гауссианов

- Вот такая графическая модель:



- Распределение Дирихле пусть будет симметричное для простоты; часто ещё $\mathbf{m}_0 = \mathbf{0}$.
- Заметьте разницу между латентными переменными и параметрами модели.

Вариационное приближение

- Теперь вариационное приближение. Сначала сама факторизация:

$$p(\mathsf{X}, \mathsf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = p(\mathsf{X} \mid \mathsf{Z}, \boldsymbol{\mu}, \Lambda) p(\mathsf{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu} \mid \Lambda) p(\Lambda).$$

- Мы наблюдаем только X, остальное всё надо как-то оценить.
- Интересно, что единственное предположение про наше вариационное приближение выглядит так:

$$q(\mathsf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = q(\mathsf{Z}) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda).$$

- И всё! Дальше всё само собой получится. Но не сразу...

Вариационное приближение

- Сначала $q^*(\mathsf{Z})$:

$$\begin{aligned}\ln q^*(\mathsf{Z}) &= \mathsf{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\Lambda}\left[\ln p(\mathsf{X},\mathsf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\Lambda)\right] + \mathrm{const} \\ &= \mathsf{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\Lambda}\left[\ln p(\mathsf{Z}\mid\boldsymbol{\pi})\right] + \mathsf{E}_{\boldsymbol{\mu},\Lambda}\left[\ln p(\mathsf{X}\mid\mathsf{Z},\boldsymbol{\mu},\Lambda)\right] + \mathrm{const} \\ &= \ldots = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk} + \mathrm{const},\end{aligned}$$

$$\text{где }\ \ln\rho_{nk} = \mathsf{E}[\ln\pi_k] + \frac{1}{2}\mathsf{E}[\ln|\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \\ - \frac{1}{2}\mathsf{E}_{\boldsymbol{\mu}_k,\Lambda_k}\left[(\mathsf{x}_n-\boldsymbol{\mu}_k)^\top\Lambda_k(\mathsf{x}_n-\boldsymbol{\mu}_k)\right].$$

- Нормируем:

$$q^*(\mathsf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}},\ \text{ где } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}.$$

Вариационное приближение

- Теперь $\mathrm{E}[z_{nk}] = r_{nk}$, т.е. r_{nk} – то, насколько точка $\mathbf{x}_n$ принадлежит кластеру k.
- Можно определить статистики с их учётом, как обычно:

$$N_k = \sum_{n=1}^{N} r_{nk},$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n,$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top.$$

- То же самое происходило и в EM-алгоритме.

Вариационное приближение

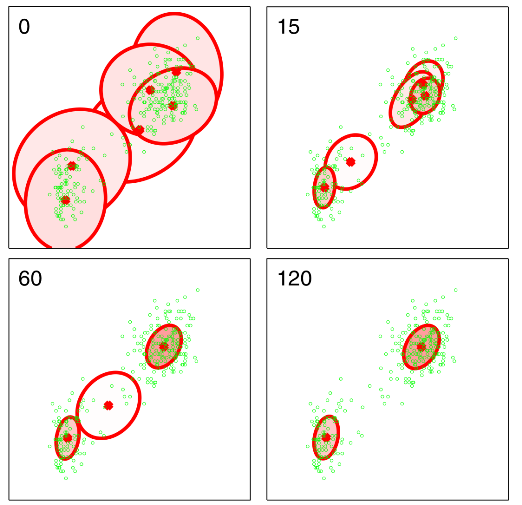
- Теперь $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$:

$$\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = \ln p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k, \Lambda_k) + \mathrm{E}_{Z}[\ln p(Z \mid \boldsymbol{\pi})] + \sum_{k=1}^{K} \sum_{n=1}^{N} \mathrm{E}[z_{nk}] \ln \mathcal{N}(\mathrm{x}_n \mid \boldsymbol{\mu}_k \Lambda_k^{-1}) + \mathrm{const.}$$

- Вот уже получилось, что $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda)$ раскладывается в $q^*(\boldsymbol{\pi}) q^*(\boldsymbol{\mu}, \Lambda)$, опять же без предположений.
- Более того, $q^*(\boldsymbol{\mu}, \Lambda) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \Lambda_k)$.
- И теперь можно по отдельности посчитать (упражнение), получится типичный М-шаг.
- Причём распределения останутся той же формы (т.к. были сопряжённые).

Вариационное приближение

- Теперь даже model selection автоматически получается, просто у некоторых компонент $N_k \approx 0$:



- Никакого оверфиттинга или коллапса компонент.

Спасибо!

Спасибо за внимание!