ЕМ-алгоритм

Д. П. Ветров[1] Д. А. Кропотов[2] А. А. Осокин[1]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

Дифференцирование по вектору

- Пусть $f(\boldsymbol{x})$ — некоторая скалярная функция, зависящая от вектора $\boldsymbol{x} \in \mathbb{R}^n$. Тогда ее производная по вектору по определению есть
$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right) = \nabla f(\boldsymbol{x})$$
- Пусть $\boldsymbol{f}(x) = (f_1(x), \dots, f_m(x))^T$ — некоторая векторная функция от скалярной переменной $x \in \mathbb{R}$. Тогда ее производная по аргументу по определению есть
$$\frac{\partial \boldsymbol{f}(x)}{\partial x} = \left(\frac{\partial f_1(x)}{\partial x}, \dots, \frac{\partial f_m(x)}{\partial x}\right)^T$$
- Пусть $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}))^T$ — некоторая векторная функция, зависящая от вектора $\boldsymbol{x} \in \mathbb{R}^n$. Тогда ее производная по вектору будет матрицей
$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right) \in \mathbb{R}^{n\times m}$$

Дифференцирование матриц

- Пусть $A(x) = (a_{ij}(x)) \in \mathbb{R}^{n\times n}$ — квадратная матрица, зависящая от параметра x. Тогда ее производная по параметру по определению равна

$$\frac{\partial A(x)}{\partial x} = \left(\frac{\partial a_{ij}(x)}{\partial x}\right)$$

- В частности, выписывая выражения покоординатно можно показать (Упр.), что

$$\frac{\partial AB}{\partial x} = \frac{\partial A}{\partial x}B + A\frac{\partial B}{\partial x}$$

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$$

$$\frac{\partial \log\det(A)}{\partial x} = \mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

Дифференцирование по матрице

- Рассмотрим некоторую скалярную функцию, зависящую от матрицы $f(A)$, $A = (a_{ij}) \in \mathbb{R}^{n \times n}$
- При поиске оптимальной матрицы

$$A_* = \arg\max_A f(A)$$

возникает задача дифференцирования функции по матрице

- Производной функции по матрице назовем матрицу производных по соответствующим элементам A

$$\frac{\partial f(A)}{\partial A} = \left(\frac{\partial f(A)}{\partial a_{ij}}\right) \in \mathbb{R}^{n \times n}$$

Полезные формулы

- Производная следа матрицы
$$\frac{\partial\,\mathrm{tr}(AB)}{\partial A} = B^T, \quad \frac{\partial\,\mathrm{tr}(A^TB)}{\partial A} = B$$
- Выведем производную определителя матрицы $\frac{\partial \det(A)}{\partial A}$. Для этого распишем определитель по строке
$$\det(A) = \sum_{i=1}^{n} a_{ij}M_{ij},$$
где $M_{ij} = (-1)^{i+j-1}\det(A^{ij})$ — алгебраическое дополнение, а A^{ij} — матрица, полученная из A путем вычеркивания i-ой строки и j-го столбца. Тогда, учитывая, что M_{ik} не зависит от a_{ij} для любых $k \neq j$, получаем
$$\frac{\partial \det(A)}{\partial a_{ij}} = \frac{\partial \sum_{i=1}^{n} a_{ij}M_{ij}}{\partial a_{ij}} = M_{ij}.$$
Каждый элемент матрицы A^{-1} выражается через алгебраические дополнения матрицы A как $a_{ij}^{-1} = \frac{1}{\det(A)}M_{ji}$, отсюда
$$\frac{\partial \det(A)}{\partial A} = \det(A)A^{-T}$$

Нормальное распределение

- Многомерное нормальное распределение имеет вид

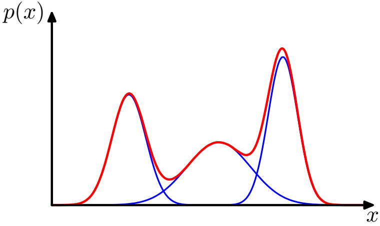
$$X \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

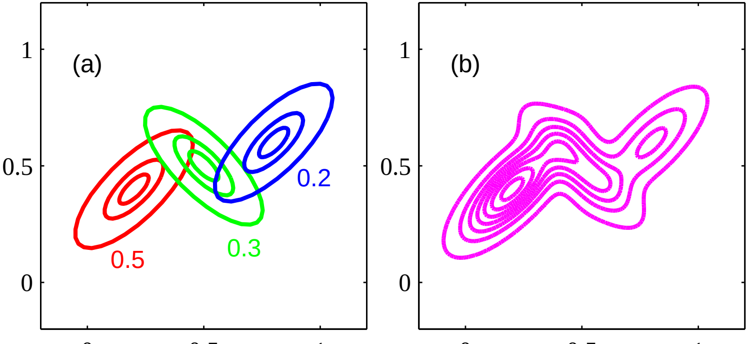
  где $\boldsymbol{\mu} = \mathbb{E}X$, $\Sigma = \mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T$ — вектор математических ожиданий каждой из n компонент и матрица ковариаций соответственно

- Смесь нормальных распределений

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k \geq 0.$$

Смесь нормальных распределений





Постановка задачи

- Дана выборка X.
- Применить М.М.П.:

$$\ln p(X \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right\}.$$

- Получается сложная оптимизационная задача с сингулярностями. (E.g., если компонента превращается в одну точку).

Вывод формул.

Вывод формул на доске.

ЕМ-алгоритм

1. Инициализировать $\boldsymbol{\mu}_k$, Σ_k, π_k.
2. Е-шаг. Вычислить апостериорные вероятности:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_l \pi_l \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l)}.$$

3. М-шаг. Пересчет параметров.

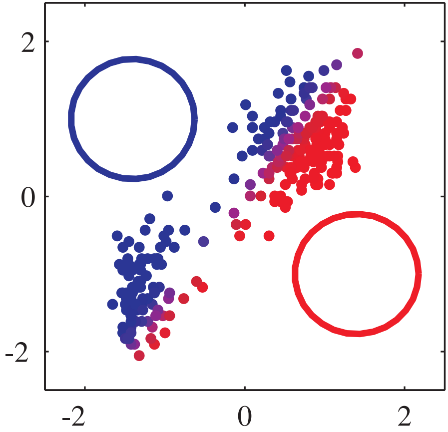
$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n,$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})^T,$$

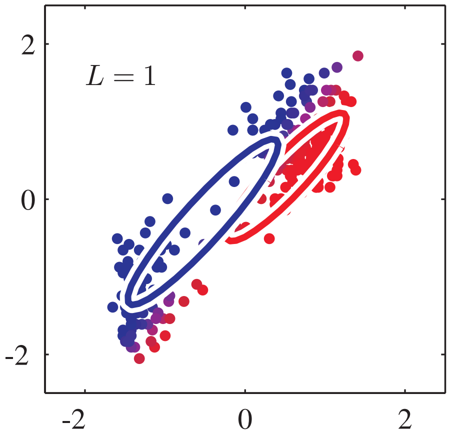
$$\pi_k = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} \gamma_{nk}.$$

4. Вычислить логарифм правдоподобия, и, если не останов, то шаг 2.

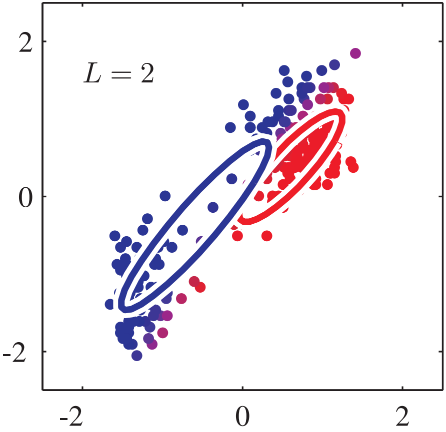
Пример



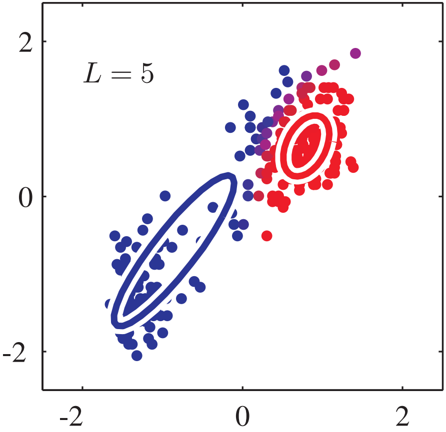
Пример



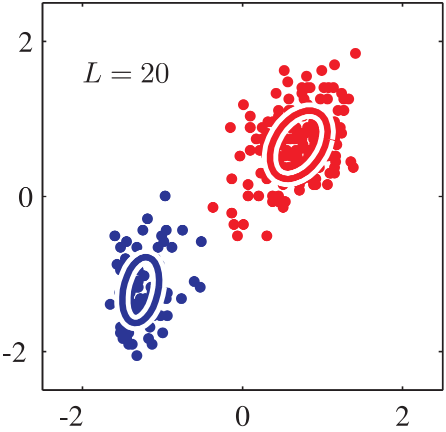
Пример



Пример



Пример



EM-алгоритм

- X – наблюдаемые переменные; Z – не наблюдаемые переменные; θ –параметры
- $p(X \mid \theta)$ – неполное правдоподобие; $p(X, Z \mid \theta)$ – полное правдоподобие
- Соотношение полного и неполного правдоподобия:

$$\ln p(X \mid \theta) = \ln \left\{ \sum_{Z} p(X, Z \mid \theta) \right\}.$$

- Сумма внутри логарифма!
- О скрытых переменных известно распределение $p(Z \mid X, \theta^{old})$
- Вместо полного правдоподобия возьмем его М.О. по апостериорному распределению.

ЕМ-алгоритм

1. Инициализировать θ^{old}.
2. Е-шаг. Вычислить апостериорные вероятности:
$$p(Z \mid X, \theta^{old})$$
3. М-шаг. Пересчет параметров.
$$\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}),$$
где
$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{Z} p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$$
4. Вычислить логарифм правдоподобия, и, если не останов, то шаг 2.
$$\theta^{old} := \theta^{new}.$$

ЕМ для смеси нормальных распределений.

Вывод формул на доске.

EM как покоординатный спуск

$$p(X \mid \theta) = \sum_Z p(X, Z \mid \theta)$$

Заметим, что

$$p(X \mid \theta) = \mathcal{L}(q, \theta) + KL(q(Z) \| p(Z \mid X, \theta)),$$

где

$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z \mid \theta)}{q(Z)} \right\},$$

$$KL(q(Z) \| p(Z \mid X, \theta)) = -\sum_Z q(Z) \ln \left\{ \frac{p(Z \mid X, \theta)}{q(Z)} \right\}.$$

Это верно, поскольку

$$\ln p(X, Z \mid \theta) = \ln p(Z \mid X, \theta) + \ln p(X \mid \theta).$$

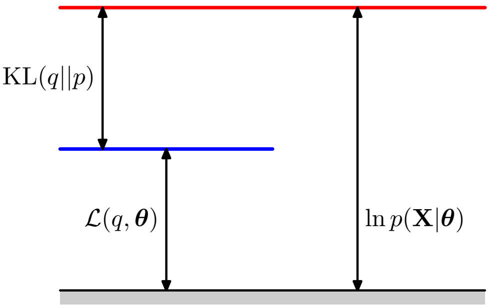
KL-дивергенция

$$KL(q(Z)\|p(Z)) = -\sum_Z q(Z)\ln\left\{\frac{p(Z)}{q(Z)}\right\}.$$

- $KL(q(Z)\|p(Z)) \geq 0$
- $KL(q(Z)\|p(Z)) = 0 \quad \Leftrightarrow q \equiv p.$
- KL несимметрична

Нижняя граница

$$p(X \mid \theta) = \mathcal{L}(q, \theta) + KL(q(Z) \| p(Z \mid X, \theta)),$$

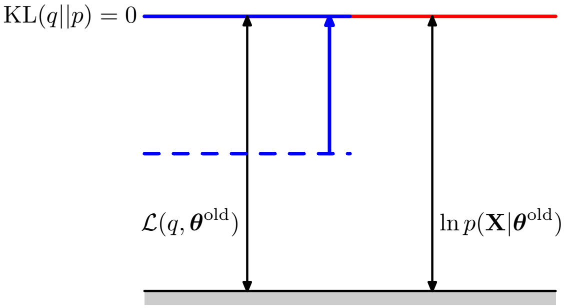


$KL \geq 0$, а значит $\mathcal{L}$ – нижняя граница неполного правдоподобия $p(X \mid \theta)$.

E-шаг

$$p(X \mid \theta) = \mathcal{L}(q, \theta) + KL(q(Z) \| p(Z \mid X, \theta)),$$

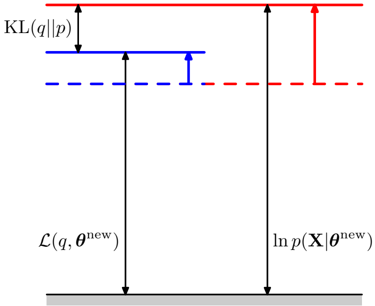
Максимизируем $\mathcal{L}(q, \theta)$ по q за счет минимизации KL.



М-шаг

$$p(X \mid \theta) = \mathcal{L}(q, \theta) + KL(q(Z) \| p(Z \mid X, \theta)),$$

Максимизируем $\mathcal{L}(q, \theta)$ по θ.



Если $q(Z) = p(Z \mid X, \theta^{old})$, то

$$\mathcal{L}(q, \theta) = \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta) - \sum_Z p(Z \mid X, \theta^{old}) \ln p(Z \mid X,$$

$$= \mathcal{Q}(\theta, \theta^{old}) + \text{const}.$$

Варианты Е-шага

- $p(Z \mid X, \theta^{old})$ точно вычислимо
- $p(Z \mid X, \theta^{old}) = \prod_i q(Z_i)$ – вариационный подход (будет на следующих лекциях)
- $p(Z \mid X, \theta^{old}) = \delta\left(\arg\max_Z p(Z \mid X, \theta^{old})\right)$ – жесткое приписывание классам. Например, K средних.