

Прикладная статистика в машинном обучении

Лекция 11

EM-алгоритм

И. К. Козлов
(Мехмат МГУ)

2022

EM-алгоритм

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

Разделение выборок

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

Хьюстон, у нас проблемы

Проблема дня. Дан мешок с монетами.

В нём **2 типа** монет. Они неотличимы на глаз.

У них разная (*неизвестная нам*) вероятность выпадения орла.

Как понять — **какие монеты какого типа?**



Prob. p_A



Prob. p_B

2 типа монет

Уточним задачу. Возьмём горсть из m монет.

Подбросим каждую из них n раз.

Нужно по таблице восстановить тип каждой взятой монеты.



Н Т Т Т Н Н Т Н Т Н



Н Н Н Н Т Н Н Н Н Н



Н Т Н Н Н Н Н Т Н Н



Н Т Н Т Т Т Н Н Т Т

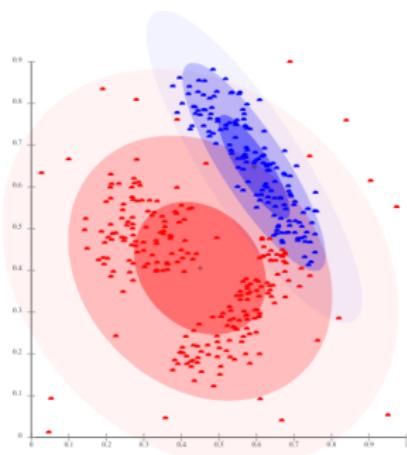


Т Н Н Н Т Н Н Н Т Н

2 типа монет

Более формально — это задача **разделения выборок**.

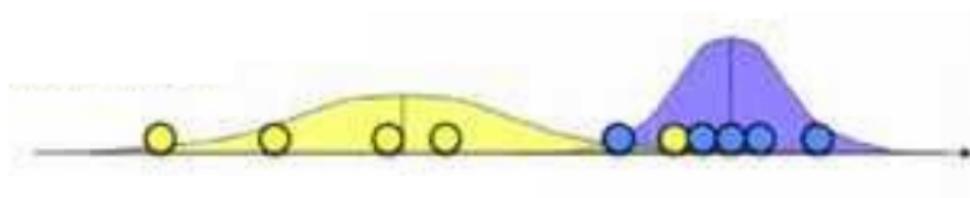
Она тесно связана с задачей **кластеризации**
(разбиения объектов по кластерам).



Разделение гауссиан

Пример. Разделение двух гауссиан.

- Пусть мы знаем типы точек.

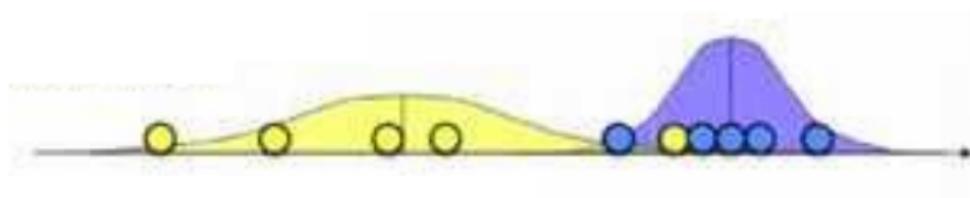


Q: Можем найти параметры гауссиан μ, σ^2 ?

Разделение гауссиан

Пример. Разделение двух гауссиан.

- Пусть мы знаем типы точек.



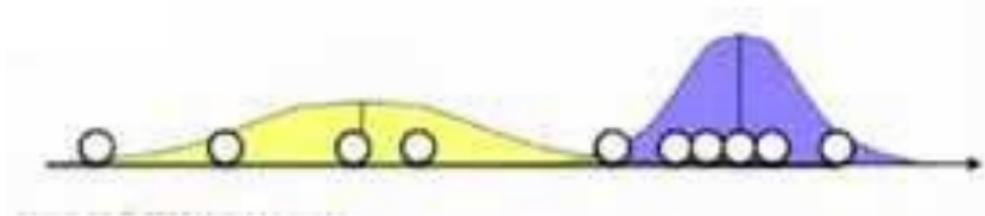
Q: Можем найти параметры гауссиан μ, σ^2 ?

A: Да, например по принципу максимума правдоподобия

$$\hat{\mu} = \mu_{ML}, \quad \hat{\sigma}^2 = \sigma_{ML}^2.$$

Разделение гауссиан

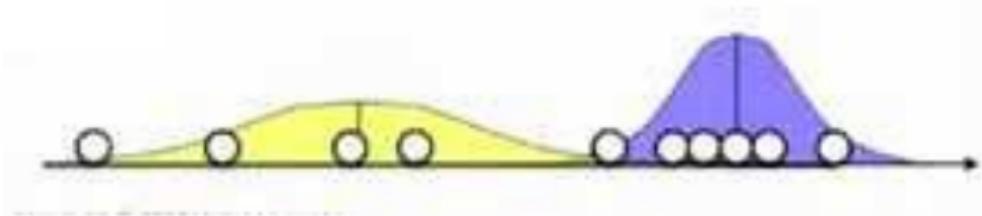
- Пусть мы знаем параметры распределений (но НЕ знаем типы точек).



Q: Как найти вероятности принадлежности точки классу $p(\text{class} | x)$?

Разделение гауссиан

- Пусть мы знаем параметры распределений (но НЕ знаем типы точек).



Q: Как найти вероятности принадлежности точки классу $p(\text{class} | x)$?

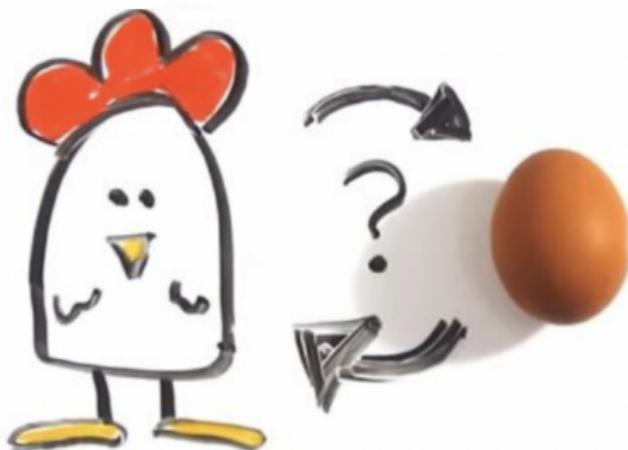
A: По [формуле Байеса](#). Обозначим классы a и b , тогда

$$p(b | x) = \frac{p(x | b)p(b)}{p(x | b)p(b) + p(x | a)p(a)}.$$

Разделение гауссиан

Q: А что делать, если мы НЕ знаем ни параметры, ни типы точек?

Получается проблема вида “курица или яйцо”.



Идея алгоритма

Общая идея EM-алгоритма.

На примере задачи с монетками. Вводим переменную z — тип монетки.

Итеративный процесс:

- ① Обновляем вероятности попадания в класс $p(z | x, \theta)$.
- ② Обновляем параметры $\theta = \arg \max_{\theta} \ln(Likelihood)$.

Повторяем до сходимости.

Латентные переменные

В общем случае будет распределение $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$ с 3 типами переменных:

- $\textcolor{red}{X}$ — наблюдаемые переменные;
- $\textcolor{blue}{Z}$ — латентные (или скрытые) переменные;
- θ — параметры распределений.

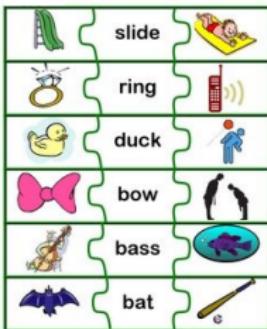
Латентные переменные

Q: Примеры латентных переменных?

Латентные переменные

Q: Примеры латентных переменных?

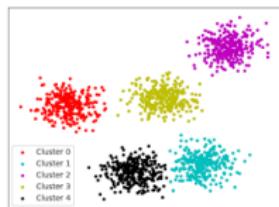
Переводы и смыслы слов, их произношения



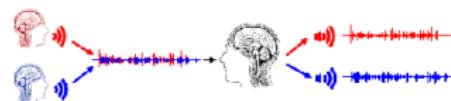
		kun'yomi 訓読み	on'yomi 音読み
"one"	→	ひと hi to	い ち chi
"person"	人	り ri	に ん n



Сегментация изображений



Номер кластера



Разделение спикеров

EM-алгоритм

E = Expectation, M = Maximization.

EM-алгоритм

- **E-шаг.** Вычисляем апостериорное распределение для латентных переменных:

$$q^t(\mathbf{Z}) = p(\mathbf{Z} | \mathbf{X}, \theta^t).$$

Находим матожидание логарифма правдоподобия латентным переменным:

$$Q(\theta) = \mathbb{E}_{\mathbf{Z}} \ln p(\mathbf{X}, \mathbf{Z} | \theta^t) = \int q^t(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \theta) d\mathbf{Z}.$$

- **M-шаг.** Максимизируем найденную функцию

$$\theta_{t+1} = \arg \max_{\theta} Q(\theta).$$

Самая сложная теорема

EM-алгоритм — бесспорно, самая сложная теорема этого курса.

Вся оставшаяся лекция будет посвящена её доказательству.



Рис.: Финальный босс этого курса

KL-дивергенция

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

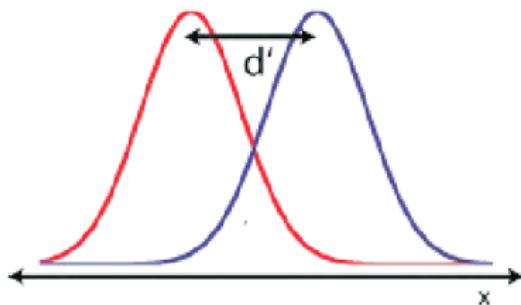
4 Применение EM-алгоритма

- Пример. Смеси гауссиан

Расстояние между распределениями

Для доказательства EM-алгоритма изучим одно важное понятие.

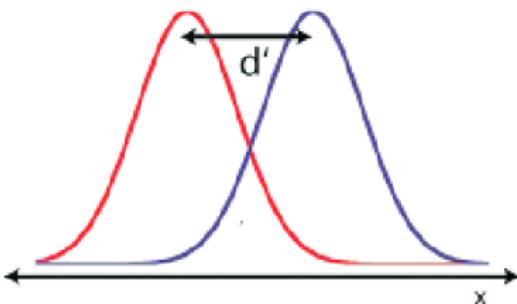
Q: Как измерить расстояние между распределениями?



Расстояние между распределениями

Для доказательства EM-алгоритма изучим одно важное понятие.

Q: Как измерить расстояние между распределениями?



A: Есть разные способы. Обсудим расстояние Кульбака-Лейблера

Другие названия: расхождение Кульбака-Лейблера, KL-дивергенция, различающая информация, относительная энтропия.

KL-дивергенция

Пусть P и Q — распределения с плотностями $p(x)$ и $q(x)$.

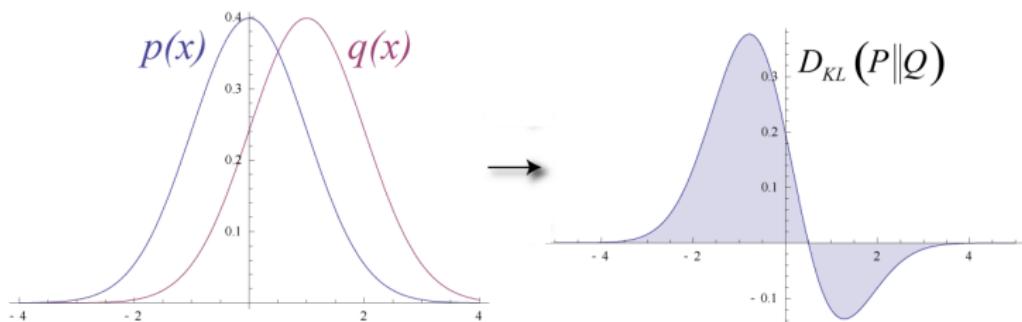
Расхождение Кульбака–Лейблера распределения Q относительно P задаётся формулой

$$\text{KL}(P \parallel Q) = \int p(x) \ln \frac{p(x)}{q(x)} dx.$$

Также обозначает $D_{\text{KL}}(P \parallel Q)$.

KL-дивергенция

KL-дивергенция = берём логарифм отношения $\ln \frac{p}{q}$ и затем интегрируем по p .



Дilemma смещения-дисперсия

Дивергенция Кульбака–Лейблера — “**несимметричное расстояние**”, она

- несимметрична

$$\text{KL}(p \parallel q) \neq \text{KL}(q \parallel p),$$

- неотрицательна

$$\text{KL}(p \parallel q) \geq 0,$$

- и равна нулю \Leftrightarrow распределения совпадают почти всюду:

$$\text{KL}(p \parallel q) = 0 \quad \Leftrightarrow \quad p(x) = q(x) \quad \text{п.в.}.$$

Неравенство Гиббса

Неравенство $KL(p \parallel q) \geq 0$ называется **неравенством Гиббса**.

Докажем его через **неравенство Йенсена**.

Неравенство Гиббса

Неравенство $\text{KL}(p \parallel q) \geq 0$ называется **неравенством Гиббса**.

Докажем его через **неравенство Йенсена**.

Отметим, что для дискретных распределений

$$P = \{p_1, \dots, p_n\}, \quad Q = \{q_1, \dots, q_n\}$$

неравенство можно записать в виде

$$-\sum_{i=1}^n p_i \log p_i \leq -\sum_{i=1}^n p_i \log q_i.$$

Информационная энтропия распределения P не больше **кросс-энтропии** распределения P с любым другим распределением Q .

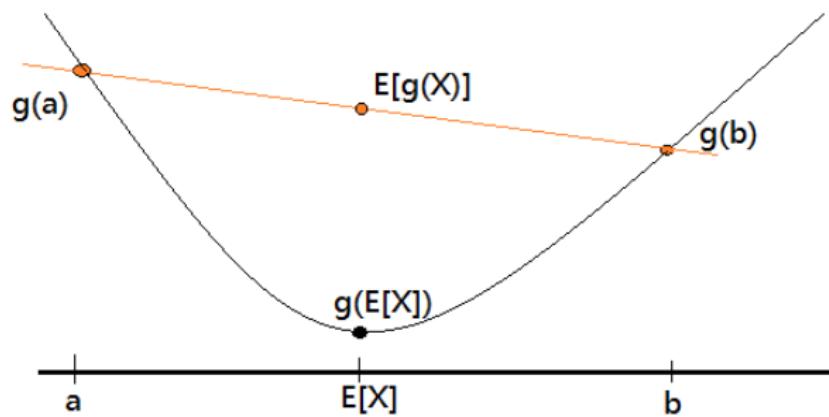
Неравенство Йенсена

Пусть $g(x)$ — выпуклая вниз функция, X — случайная величина.

Неравенство Йенсена:

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Предполагаем, что оба матожидания конечны.



Неравенство Гиббса

Доказательство неравенства Гиббса $\text{KL}(p \parallel q) \geq 0$.

- Для удобства меняем знак

$$-\text{KL}(p \parallel q) = \int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx.$$

Неравенство Гиббса

Доказательство неравенства Гиббса $\text{KL}(p \parallel q) \geq 0$.

- Для удобства меняем знак

$$-\text{KL}(p \parallel q) = \int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx.$$

- Применяем неравенство Йенсена

$$\int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx \leq \log \int p(x) \left(\frac{q(x)}{p(x)} \right) dx.$$

Неравенство Гиббса

Доказательство неравенства Гиббса $\text{KL}(p \parallel q) \geq 0$.

- Для удобства меняем знак

$$-\text{KL}(p \parallel q) = \int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx.$$

- Применяем неравенство Йенсена

$$\int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx \leq \log \int p(x) \left(\frac{q(x)}{p(x)} \right) dx.$$

- Правая часть равна нулю.

$$\log \int p(x) \left(\frac{q(x)}{p(x)} \right) dx = \log \int q(x) dx = \log 1 = 0.$$

Неравенство Гиббса

Доказательство неравенства Гиббса $\text{KL}(p \parallel q) \geq 0$.

- Для удобства меняем знак

$$-\text{KL}(p \parallel q) = \int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx.$$

- Применяем неравенство Йенсена

$$\int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx \leq \log \int p(x) \left(\frac{q(x)}{p(x)} \right) dx.$$

- Правая часть равна нулю.

$$\log \int p(x) \left(\frac{q(x)}{p(x)} \right) dx = \log \int q(x) dx = \log 1 = 0.$$

- При этом равенство \Leftrightarrow под логарифмом 1, т.е. $p(x) = q(x)$ (п.в.).

Перерыв

Перерыв

Доказательство EM-алгоритма

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

- Вернёмся к модели с латентными переменными

$$p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$$

- Q: Если бы не было Z , то как найти параметры θ ?

- Вернёмся к модели с латентными переменными

$$p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$$

- **Q:** Если бы не было Z , то как найти параметры θ ?
- **A:** Конечно, по принципу максимума правдоподобия:

$$p(\textcolor{red}{X} | \theta) \rightarrow \arg \max_{\theta},$$

где правдоподобие

$$p(\textcolor{red}{X} | \theta) = \prod_{i=1}^n p(\textcolor{red}{x}_i | \theta).$$

MLE

- Q: Как из распределения $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$ получить $p(\textcolor{red}{X} | \theta)$?

- **Q:** Как из распределения $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$ получить $p(\textcolor{red}{X} | \theta)$?
- **A:** Вспоминаем — по неизвестным параметрам **маргинализуем**:

$$p(\textcolor{red}{X} | \theta) = \int p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta) d\textcolor{blue}{Z} = \int p(\textcolor{red}{X}, | \textcolor{blue}{Z}, \theta) p(\textcolor{blue}{Z} | \theta) d\textcolor{blue}{Z}.$$

Второе равенство — по формуле условной вероятности.

- **Q:** Как из распределения $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$ получить $p(\textcolor{red}{X} | \theta)$?
- **A:** Вспоминаем — по неизвестным параметрам **маргинализуем**:

$$p(\textcolor{red}{X} | \theta) = \int p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta) d\textcolor{blue}{Z} = \int p(\textcolor{red}{X}, | \textcolor{blue}{Z}, \theta) p(\textcolor{blue}{Z} | \theta) d\textcolor{blue}{Z}.$$

Второе равенство — по формуле условной вероятности.

- Распределение $p(\textcolor{blue}{Z} | \theta)$ обозначим за $q(\textcolor{blue}{Z})$.

Шаг 1

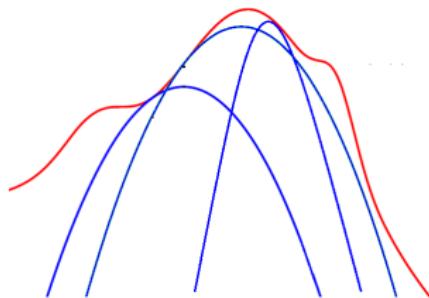
Как всегда, удобнее максимизировать логарифм правдоподобия $\ln p(\mathbf{X} | \theta)$.

Доказательство ЕМ-алгоритма состоит из 2ух шагов:

- Шаг 1. Найдём вариационную нижнюю оценку — функцию $\mathcal{L}(q, \theta)$ т.ч.

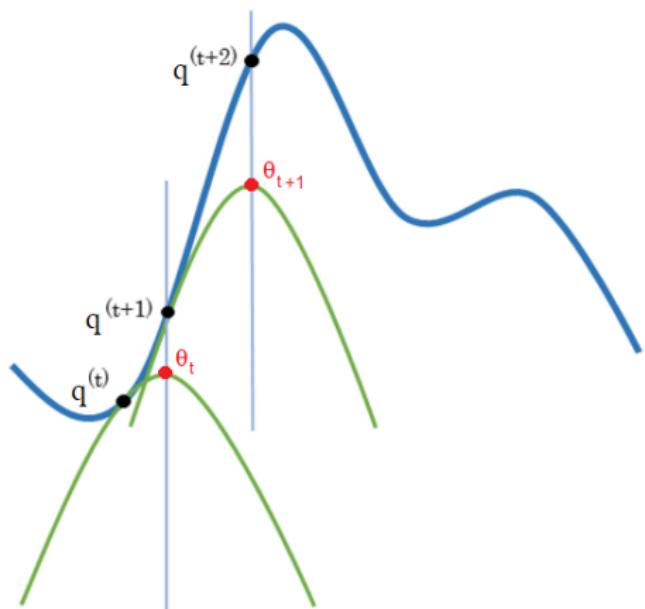
$$\ln p(\mathbf{X} | \theta) \geq \mathcal{L}(q, \theta).$$

По-английски ELBO = Evidence Lower BOund.



Шаг 2

- Шаг 2. Оптимизируем $\mathcal{L}(q, \theta)$, как на прошлом занятии.



ELBO

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

Находим вариационную нижнюю оценку.

Лемма

$$\ln p(\textcolor{red}{X} | \theta) = \mathcal{L}(q, \theta) + \text{KL}(q \| p),$$

- 1ое слагаемое:

$$\mathcal{L}(q, \theta) = \int q(\textcolor{blue}{Z}) \ln \frac{p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)}{q(\textcolor{blue}{Z})} d\textcolor{blue}{Z}.$$

- 2ое слагаемое:

$$\text{KL}(q \| p) = \int q(\textcolor{blue}{Z}) \ln \frac{q(\textcolor{blue}{Z})}{p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta)} d\textcolor{blue}{Z}.$$

Доказательство Леммы — серия несложных преобразований.

- Пролографмируем правило произведения для $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$:

$$\ln p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta) = \ln p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta) + \ln p(\textcolor{red}{X} | \theta).$$

ELBO

Доказательство Леммы — серия несложных преобразований.

- Пролографмируем правило произведения для $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$:

$$\ln p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta) = \ln p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta) + \ln p(\textcolor{red}{X} | \theta).$$

- Выразим отсюда $\ln p(\textcolor{red}{X} | \theta)$ и возьмём матожидание по плотности $q(\textcolor{blue}{Z})$:

$$\ln p(\textcolor{red}{X} | \theta) = \int q(\textcolor{blue}{Z}) \ln \frac{p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)}{p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta)} d\textcolor{blue}{Z}$$

ELBO

- Домножим числитель и знаменатель дроби на $q(\textcolor{blue}{Z})$ и разобьём дробь:

$$\ln p(\textcolor{red}{X} | \theta) = \int q(\textcolor{blue}{Z}) \ln \frac{p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)}{q(\textcolor{blue}{Z})} d\textcolor{blue}{Z} + \int q(\textcolor{blue}{Z}) \ln \frac{q(\textcolor{blue}{Z})}{p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta)} d\textcolor{blue}{Z}$$

Это в точности то разложение, которое нам нужно.

Лемма доказана.

ELBO

Q: ELBO можно записать в виде

$$-\int q(Z) \ln \frac{q(Z)}{p(X, Z | \theta)} dZ$$

Это не KL-дивергенция?

$$\text{KL}(Q \| P) = \int q(x) \log \frac{q(x)}{p(x)} dx.$$

ELBO

Q: ELBO можно записать в виде

$$-\int q(Z) \ln \frac{q(Z)}{p(X, Z | \theta)} dZ$$

Это не KL-дивергенция?

$$\text{KL}(Q \| P) = \int q(x) \log \frac{q(x)}{p(x)} dx.$$

A: Нет, $q(Z)$ — функция от Z , а $p(X, Z | \theta)$ — от (X, Z) .

E и M шаги

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

Maximization-Maximization

Поскольку KL-дивергенция неотрицательна,

$$\ln p(\textcolor{red}{X} | \theta) \geq \int q(\textcolor{blue}{Z}) \ln \frac{p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)}{q(\textcolor{blue}{Z})} d\textcolor{blue}{Z} = \mathcal{L}(q, \theta).$$

Начинаем попаременно оптимизировать $\mathcal{L}(q, \theta)$ по q и по θ .

Е-шаг

Е-шаг

Оптимизируем по q :

$$q^t = \arg \max_q \mathcal{L}(q, \theta^t).$$

Мы знаем, что

$$\ln p(\textcolor{red}{X} | \theta^t) = \mathcal{L}(q, \theta^t) + \text{KL}(q \| p),$$

Q: Чему равно q^t ?

Е-шаг

Е-шаг

Оптимизируем по q :

$$q^t = \arg \max_q \mathcal{L}(q, \theta^t).$$

Мы знаем, что

$$\ln p(\textcolor{red}{X} | \theta^t) = \mathcal{L}(q, \theta^t) + \text{KL}(q \| p),$$

Q: Чему равно q^t ?

A: Левая часть не зависит от q , поэтому максимум ELBO — когда

$$\text{KL}(q \| p) = 0 \quad \Leftrightarrow \quad q(\textcolor{blue}{Z}) = p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta^t).$$

М-шаг

М-шаг

Оптимизируем по θ :

$$\theta^{t+1} = \arg \max_{\theta} \mathcal{L}(q^t, \theta).$$

По определению

$$\mathcal{L}(q, \theta) = \int q(Z) \ln \frac{p(X, Z | \theta)}{q(Z)} dZ = \int q(Z) \ln p(X, Z | \theta) dZ - \int q(Z) \ln q(Z) dZ.$$

Второе слагаемое не зависит от θ , поэтому его можно отбросить.

Получаем требуемое выражение

$$\theta^{t+1} = \arg \max_{\theta} \int q(\textcolor{blue}{Z}) \ln p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta) d\textcolor{blue}{Z}.$$

Victory!

ЕМ-алгоритм доказан.



Перерыв

Перерыв

Применение EM-алгоритма

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

E-шаг

Насколько просто выполнять шаги EM-алгоритма?

E-шаг. Это байесовский вывод

$$q(\textcolor{blue}{Z}) = p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta^t) = \frac{p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta^t)p(\textcolor{blue}{Z} | \theta^t)}{\int p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta^t)p(\textcolor{blue}{Z} | \theta^t)d\textcolor{blue}{Z}}.$$

E-шаг

Насколько просто выполнять шаги EM-алгоритма?

E-шаг. Это байесовский вывод

$$q(\textcolor{blue}{Z}) = p(\textcolor{blue}{Z} | \textcolor{red}{X}, \theta^t) = \frac{p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta^t)p(\textcolor{blue}{Z} | \theta^t)}{\int p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta^t)p(\textcolor{blue}{Z} | \theta^t)d\textcolor{blue}{Z}}.$$

Если у Z — конечное число значений, то вместо интеграла — сумма, и всё считается.

В общем случае — как обычно:



Апостериорное можно аналитически посчитать, если $p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta^t)$ и $p(\textcolor{blue}{Z} | \theta^t)$ сопряжены.

М-шаг

М-шаг.

$$\theta^{t+1} = \arg \max_{\theta} \int q(\textcolor{blue}{Z}) \ln p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta) d\textcolor{blue}{Z}.$$

Максимизация по параметру — хорошая с вычислительной точки зрения задача.

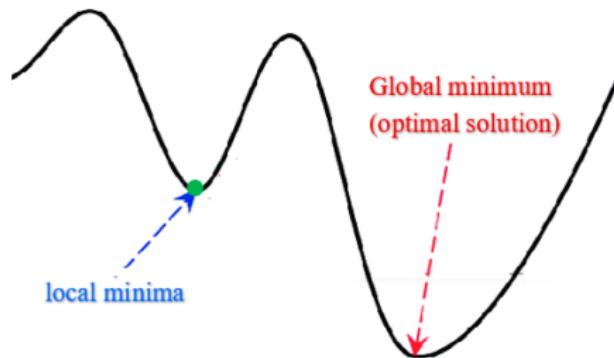
Если вдруг $\ln p(\textcolor{red}{X}, \textcolor{blue}{Z} | \theta)$ — вогнутая функция, то совсем просто:

- Сумма (и интеграл от) вогнутых функций — вогнутая функция.
- Максимум вогнутой функции легко находится градиентным спуском.

Локальные минимумы

EM-алгоритм не гарантирует сходимости к **глобальному минимуму**.
Он может застревать в **локальных минимумах**.

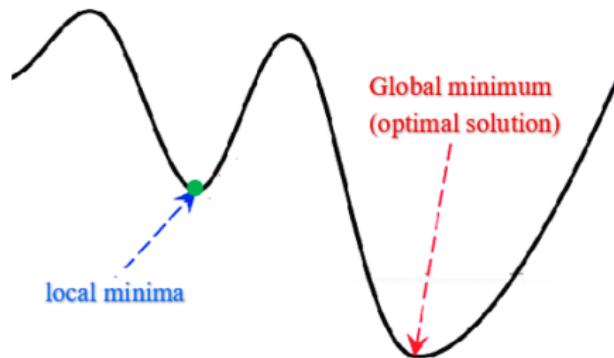
Q: Как с этим бороться?



Локальные минимумы

EM-алгоритм не гарантирует сходимости к **глобальному минимуму**.
Он может застревать в **локальных минимумах**.

Q: Как с этим бороться?



A: Запуск с разных стартовых позиций.

Локальные минимумы

Q: Как выбрать количество компонент в EM-алгоритме для смеси распределений?

A: Непростой вопрос. В малой размерности — “на глаз”.

Эвристика — посчитать с разным числом компонент и сравнить BIC.

Другие способы: <https://scikit-learn.org/stable/modules/mixture.html>

Локальные минимумы

Q: Как долго выполнять EM-алгоритм? (Что мы максимизируем?)

Локальные минимумы

Q: Как долго выполнять EM-алгоритм? (Что мы максимизируем?)

A: Мы хотим максимизировать $\ln p(\textcolor{red}{X} | \mu, \Sigma, \pi)$.

Можно остановить, если на новом шаге логарифм правдоподобия увеличился менее, чем на ε .

Е-шаг

Естественно, на каждом шаге могут возникать проблемы.

Мы рассмотрели простейшую схему ЕМ-алгоритма, без модификаций.

Подробнее — соответственно в Главах 9 и 11



Bishop C.M.

Pattern Recognition and Machine Learning.



Murphy K.P.

Machine Learning: A Probabilistic Perspective.

Пример. Смеси гауссиан

1 EM-алгоритм

- Разделение выборок

2 KL-дивергенция

3 Доказательство EM-алгоритма

- Вариационная нижняя оценка (ELBO)
- E и M шаги

4 Применение EM-алгоритма

- Пример. Смеси гауссиан

Смесь гауссиан

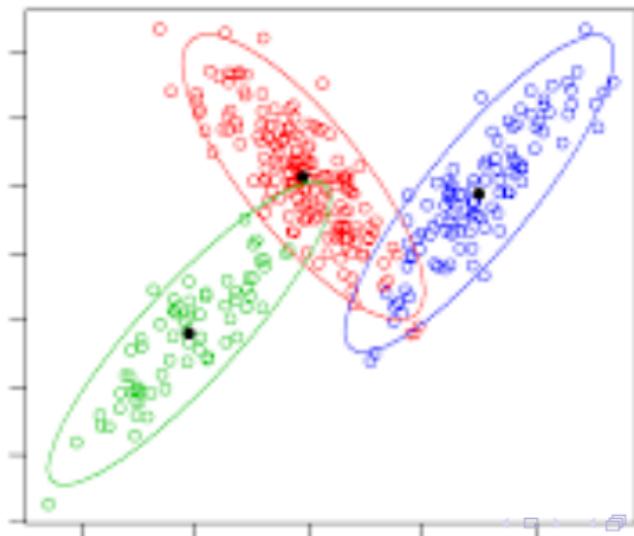
Рассмотрим смесь K гауссиан.

Тип гауссианы зададим K -мерным базисным вектором:

$$z = (z_1, \dots, z_k), \quad z_i \in \{0, 1\} \quad \sum_k z_k = 1$$

Плотность смеси:

$$p(x | z) = \prod_{k=1}^K \mathcal{N}(x | \mu_k, \Sigma_k)^{z_k}.$$



Смесь гауссиан

Выбираем априорное распределение для z .

Чтобы всё считалось, хочется, чтобы оно было сопряжено с правдоподобием

$$p(x | \textcolor{blue}{z}) = \prod_{k=1}^K \mathcal{N}(x | \mu_k, \Sigma_k)^{\textcolor{blue}{z_k}}.$$

Q: Какой вид у правдоподобия как функции от z ?

Смесь гауссиан

Выбираем априорное распределение для z .

Чтобы всё считалось, хочется, чтобы оно было сопряжено с правдоподобием

$$p(x | \textcolor{blue}{z}) = \prod_{k=1}^K \mathcal{N}(x | \mu_k, \Sigma_k)^{\textcolor{blue}{z_k}}.$$

Q: Какой вид у правдоподобия как функции от z ?

A: В качестве априорного возьмём **категориальное распределение**

$$p(\textcolor{blue}{z} | \pi) = \prod_{k=1}^K \pi_k^{\textcolor{blue}{z_k}}, \quad 0 \leq \pi_k \leq 1, \quad \sum_k \pi_k = 1.$$

Проще говоря, вероятность быть в k -той компоненте смеси:

$$p(z_k = 1 | \pi) = \pi_k.$$

Смесь гауссиан

Плотность смеси — линейная комбинация плотностей.

Если

$$p(\textcolor{red}{X} \mid \textcolor{blue}{z}, \theta) = \prod_{k=1}^K \mathcal{N}(\textcolor{red}{X} \mid \mu_k, \Sigma_k)^{\textcolor{blue}{z}_k}, \quad p(\textcolor{blue}{z} \mid \pi) = \prod_{k=1}^K \pi_k^{\textcolor{blue}{z}_k},$$

то итоговую плотность можно записать в виде

$$p(\textcolor{red}{X} \mid \theta) = \sum_{\textcolor{blue}{z}} p(\textcolor{red}{X} \mid \textcolor{blue}{z}, \theta) p(\textcolor{blue}{z} \mid \pi) = \sum_{k=1}^K \pi_k \mathcal{N}(\textcolor{red}{X} \mid \mu_k, \Sigma_k).$$

Е-шаг

Е-шаг

Вычисляем апостериорное распределение по формуле Байеса:

$$q(z_k) = p(z_k = 1 | x) = \frac{p(x | z_k = 1)p(z_k = 1 | \pi)}{\sum_j p(x | z_k = j)p(z_k = j | \pi)}$$

E-шаг

E-шаг

Вычисляем апостериорное распределение по формуле Байеса:

$$q(z_k) = p(z_k = 1 | x) = \frac{p(x | z_k = 1)p(z_k = 1 | \pi)}{\sum_j p(x | z_k = j)p(z_k = j | \pi)}$$

Подставляем плотности — получаем простой ответ:

Вероятности компонент пропорциональны взвешенным плотностям гауссиан:

$$q(z_k) = \frac{\pi_k \mathcal{N}(x | \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x | \mu_j, \Sigma_j)}$$

E-шаг

Матожидание логарифма правдоподобия превращается в сумму:

$$Q(\theta) = \mathbb{E}_{\mathbf{z}} \ln p(\mathbf{X}, \mathbf{z} | \theta) = \sum_{k=1}^K q(z_k) \ln p(\mathbf{X}, z_k | \theta).$$

Е-шаг

Матожидание логарифма правдоподобия превращается в сумму:

$$Q(\theta) = \mathbb{E}_{\mathbf{z}} \ln p(\mathbf{X}, \mathbf{z} | \theta) = \sum_{k=1}^K q(\mathbf{z}_k) \ln p(\mathbf{X}, \mathbf{z}_k | \theta).$$

Вычисляем правдоподобие:

- Это произведение вероятностей объектов:

$$p(\mathbf{X}, \mathbf{z}_k | \theta) = \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_k | \theta)$$

- По правилу произведения:

$$p(\mathbf{x}_n, \mathbf{z}_k | \theta) = p(\mathbf{x}_n | \mathbf{z}_k, \mu_k, \Sigma_k) p(\mathbf{z}_k | \pi) = [\mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \cdot \pi_k]^{z_k}.$$

М-шаг

М-шаг

Честно подставляя найденные выражения, получаем:

$$Q(\theta) = \sum_{n=1}^N \sum_{k=1}^K q(z_k) \left[-\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) + \log \pi_k \right].$$

Q: По каким параметрам нужно максимизировать на этом М-шаге?

M-шаг

M-шаг

Честно подставляя найденные выражения, получаем:

$$Q(\theta) = \sum_{n=1}^N \sum_{k=1}^K q(z_k) \left[-\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (\mathbf{x}_n - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) + \log \pi_k \right].$$

Q: По каким параметрам нужно максимизировать на этом M-шаге?

A: По μ , Σ и π .

Замечание. По π_k и по различным (μ_j, Σ_j) можно максимизировать независимо — они в разных слагаемых.

M-шаг

Нахождение оптимума — **упражнение по матанализу.**

- Параметры μ_j и Σ_j находятся из условий

$$\frac{\partial Q}{\partial \mu_j} = 0, \quad \frac{\partial Q}{\partial \Sigma_j} = 0.$$

M-шаг

Нахождение оптимума — **упражнение по матанализу.**

- Параметры μ_j и Σ_j находятся из условий

$$\frac{\partial Q}{\partial \mu_j} = 0, \quad \frac{\partial Q}{\partial \Sigma_j} = 0.$$

- Параметры π_k связаны соотношением

$$\sum_k \pi_k = 1,$$

поэтому их можно найти **методом множителей Лагранжа** или выразив одно из слагаемых:

$$\frac{\partial Q}{\partial \pi_k} = 0, \quad j = 1, \dots, K - 1,$$

$$\pi_K = 1 - \pi_1 - \dots - \pi_{K-1}.$$

М-шаг

М-шаг

- “Ожидаемая доля точек в k -том кластере”:

$$N_k = \sum_{n=1}^N q(z_{nk}).$$

- Новая вероятность попасть в k -тый кластер:

$$\pi_k = \frac{N_k}{N}.$$

M-шаг

M-шаг

- “Ожидаемая доля точек в k -том кластере”:

$$N_k = \sum_{n=1}^N q(z_{nk}).$$

- Новая вероятность попасть в k -тый кластер:

$$\pi_k = \frac{N_k}{N}.$$

- Новые средние — взвешенное среднее данных точек:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N q(z_{nk}) x_n.$$

- Новые матрицы ковариации:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N q(z_{nk}) (x_n - \mu_k) (x_n - \mu_k)^T.$$

ELBO

ELBO

Для проверки сходимости вычисляем логарифм правдоподобия

$$\ln p(\mathbf{X} | \mu, \Sigma, \pi) = \sum_{n=1}^N \ln \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right]$$

Если $\ln p(\mathbf{X} | \mu, \Sigma, \pi)$ увеличился более, чем на $\varepsilon \Rightarrow$ повторяем Е и М-Шаги.

ELBO

Вот и весь алгоритм для смеси гауссиан. С такими красивыми ответами.

В задаче про **разделение монеток** похожий простенький ответ — см. Главу 9.3

 Bishop C.M.

Pattern Recognition and Machine Learning.



Сегодня мы обогатили свои знания

Итак, мы научились различать монетки друг от друга.

Ещё немного — и мы будем в них купаться)



To Be Continued