Модели смесей и EM алгоритм

Владимир Анатольевич Судаков

2025

Как использовать латентные переменные

Если мы определяем совместное распределение по наблюдаемым и латентным переменным, то соответствующее распределение исключительно наблюдаемых переменных получается с помощью маргинализации.

Это позволяет выражать относительно сложные маргинальные распределения по наблюдаемым переменным в терминах более удобных совместных распределений по расширенному пространству наблюдаемых и латентных переменных.

Таким образом, введение латентных переменных позволяет создавать сложные распределения из более простых компонентов.

Дискретные латентные переменные можно интерпретировать как разделение точек по конкретным компонентам смеси.

Кластеризация К-средних

Рассмотрим задачу выделения групп, или кластеров, точек данных в многомерном пространстве. Предположим, что у нас есть набор данных $\{x_1, \dots, x_N\}$. Цель состоит в том, чтобы разбить набор данных на K кластеры для заданного значения K. Интуитивно понятно, что кластер — это группа точек данных, расстояния между которыми малы по сравнению с расстояниями до точек вне кластера.

Мы можем формализовать это понятие, введя набор D-мерных векторов $\boldsymbol{\mu}_k$. Каждый такой вектор является прототипом, связанным с k кластером -го размера, и, по сути, представляет собой центры кластеров. Цель состоит в том, чтобы найти распределение точек данных по кластерам и набор векторов $\boldsymbol{\mu}_k$, такой, чтобы сумма квадратов расстояний каждой точки данных до ближайшего к ней вектора $\boldsymbol{\mu}_k$ была минимальной.

Затем мы можем определить целевую функцию, иногда называемую мерой искажения (distortion measure), заданную как

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, ||\mathbf{x}_n - \boldsymbol{\mu}_k||_{,}^2$$

где $r_{nk} \in \{0,1\}$— бинарные индикаторные переменные, описывающие, к какому кластеру K относится точка данных x_n. Целевая функция представляет собой сумму квадратов расстояний каждой точки данных до назначенного ей центра кластера $\boldsymbol{\mu}_k$. Таким образом, цель состоит в том, чтобы найти значения для $\{r_{nk}\}$и , $\boldsymbol{\mu}_k$ минимизирующие J.

Алгоритм K-средних

На первом этапе мы минимизируем Jзначение r_{nk}, сохраняя $\boldsymbol{\mu}_k$значение . На втором этапе мы минимизируем Jзначение $\boldsymbol{\mu}_k$, сохраняя r_{nk}значение . Эти два этапа повторяются до достижения сходимости. Мы увидим, что эти два этапа обновления r_{nk}и обновления $\boldsymbol{\mu}_k$соответствуют шагам E (ожидание) и M (максимизация) алгоритма EM соответственно.

Рассмотрим определение r_{nk}. Поскольку Jявляется линейной функцией r_{nk}, оптимизация даёт решение в замкнутой форме. Члены n независимы, поэтому мы оптимизируем каждый из них отдельно, выбирая $r_{nk} = 1$ для любого значения , которое k даёт минимальное значение $||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2$. Другими словами, мы просто присваиваем каждую точку данных ближайшему к ней центру кластера, или, более формально,

$$r_{nk} = \begin{cases} 1 & \text{if } k = \text{argmin}_j ||\mathbf{x}_n - \boldsymbol{\mu}_j||_,^2 \\ 0 & \text{otherwise} \end{cases}$$

Затем рассмотрим оптимизацию $\boldsymbol{\mu}_k$, сохраняя r_{nk}фиксированное значение.

Алгоритм K-средних

Целевая функция является квадратичной функцией $\boldsymbol{\mu}_k$, и её можно минимизировать, приравняв её производную по $\boldsymbol{\mu}_k$ к нулю, что даёт:

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \mathbf{0} \Leftrightarrow$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, ||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2 = \mathbf{0} \Leftrightarrow$$

$$2 \sum_{n=1}^{N} r_{nk} \, (\mathbf{x}_n - \boldsymbol{\mu}_k) = \mathbf{0} \Leftrightarrow$$

$$\sum_{n=1}^{N} r_{nk} \, \mathbf{x}_n - \sum_{n=1}^{N} r_{nk} \, \boldsymbol{\mu}_k = \mathbf{0} \Leftrightarrow$$

$$\sum_{n=1}^{N} r_{nk} \, \mathbf{x}_n = \sum_{n=1}^{N} r_{nk} \, \boldsymbol{\mu}_k \Leftrightarrow$$

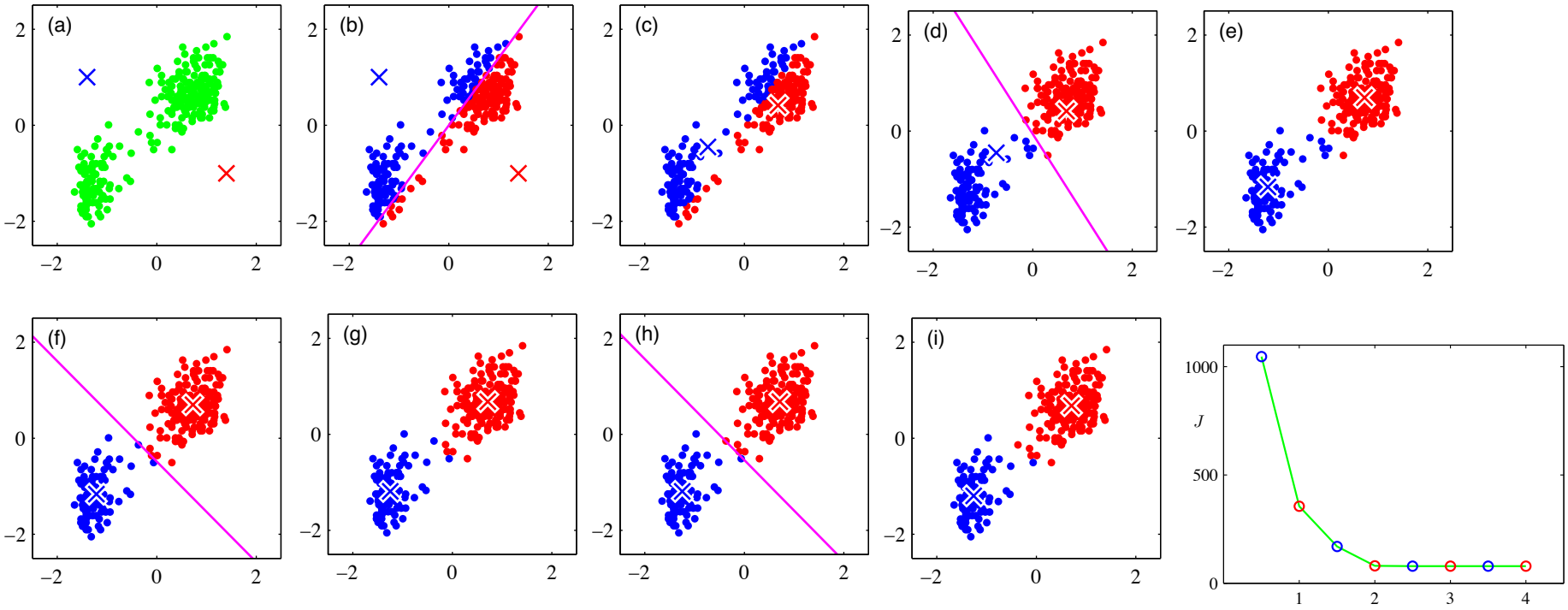
$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk} \, \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

Знаменатель равен количеству точек, отнесённых к кластеру k, и, следовательно, $\boldsymbol{\mu}_k$ равен среднему значению всех точек данных, $\mathbf{x}_n$ отнесённых к кластеру k.

Поскольку каждая фаза уменьшает значение целевой функции, сходимость алгоритма гарантирована. Однако следует учитывать, что он может сходиться к локальному, а не глобальному минимуму.

Применение алгоритма К-средних

к набору данных Old Faithful (Старый служака)



Начальные значения центров кластеров выбираются по случайному подмножеству K точек данных. Также стоит отметить, что Kсам алгоритм K-средних часто используется для инициализации параметров модели гауссовой смеси перед применением алгоритма EM.

Реализация алгоритма K-средних, обсуждаемого здесь, может быть относительно медленной, поскольку для E-шага необходимо вычислить евклидово расстояние между каждым вектором-прототипом и каждой точкой данных. Было предложено несколько схем ускорения K средних, некоторые из которых основаны на предварительном вычислении структуры данных (например, kd-дерева) таким образом, чтобы соседние точки находились в одном поддереве. Другие подходы используют неравенство треугольника для расстояний, тем самым избегая ненужных вычислений расстояний.

K-медоиды

Алгоритм K-средних обычно основан на квадрате евклидова расстояния для измерения расстояния между точкой данных и вектором-прототипом. Это ограничивает тип рассматриваемых переменных (например, не подходит для случаев, когда некоторые или все переменные представляют собой категориальные метки), но также делает определение кластерного среднего неустойчивым к выбросам. Алгоритм K-средних можно обобщить, введя более общую меру несходства $\mathcal{V}(\mathbf{x}, \mathbf{x}')$и затем минимизировав следующую меру искажения:

$$\tilde{J} = \sum_{n=1}^{N} \sum_{m=1}^{M} r_{nk} \, \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

Для общего выбора меры различия шаг M потенциально сложнее, чем для K-means, поэтому обычно ограничивают каждый прототип кластера одним из векторов данных, назначенных этому кластеру. Таким образом, шаг M включает в себя для каждого кластера kдискретный поиск по Nkточкам, назначенным этому кластеру, что требует $O(N^2 k)$вычислений.

Обратите внимание, что Kалгоритм α-средних однозначно относит каждую точку данных к одному и только одному кластеру. Однако некоторые точки данных могут находиться примерно посередине между центрами кластеров. В этом случае неясно, является ли жёсткое отнесение к ближайшему кластеру наиболее подходящим. Используя вероятностный подход, мы получаем *мягкие* отнесения точек данных к кластерам, отражающие уровень неопределённости относительно наиболее подходящего отнесения. Такая вероятностная формулировка даёт множество преимуществ.

Метрика Силуэт

Значение силуэта – это мера того, насколько объект схож со своим кластером (сплоченность) по сравнению с другими кластерами (разделенность). Значение силуэта варьируется от −1 до +1, где высокое значение указывает на то, что объект хорошо соответствует своему кластеру и плохо соответствует соседним кластерам. Если большинство объектов имеют высокое значение, то конфигурация кластеризации подходит. Если многие точки имеют низкое или отрицательное значение, то конфигурация кластеризации может содержать слишком много или слишком мало кластеров. Кластеризация со средней шириной силуэта более 0,7 считается «сильной», значение более 0,5 – «разумной», а более 0,25 – «слабой». Однако с увеличением размерности данных становится сложно достичь таких высоких значений из-за «проклятия размерности» , поскольку расстояния становятся более близкими. Значение силуэта специализировано для оценки качества кластера, когда кластеры имеют выпуклую форму, и может быть неэффективным, если кластеры данных имеют неправильную форму или разный размер. Значение силуэта можно рассчитать с помощью любой метрики расстояния, например, евклидова расстояния или манхэттенского расстояния.

Для точки данных $i \in C_i$(точки данных iв кластере C_i) пусть

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d\,(i, j)$$

будет средним расстоянием между iи всеми другими точками данных в том же кластере, где $|C_i|$— количество точек, принадлежащих кластеру C_i, а $d(i, j)$— расстояние между точками данных iи jв кластере C_i(мы делим на , $|C_i| - 1$поскольку расстояние $d(i, i)$не включается в сумму). $a(i)$может быть интерпретировано как мера того, насколько хорошо iточка приписана к своему кластеру (чем меньше значение, тем лучше приписывание).

Затем мы определяем среднее различие точки i с некоторым кластером C_j как среднее расстояние от i до всех точек в C_j(где $C_j \neq C_i$).

Для каждой точки данных $i \in C_i$ мы теперь определяем

$$b(i) = \min_{j \neq i} \frac{1}{|C_j|} \sum_{j \in C_j} d\,(i, j)$$

наименьшее (отсюда и $\min$ *оператор* в формуле) среднее расстояние до iвсех точек любого другого кластера (т.е. любого кластера, iчленом которого он не является). Кластер с наименьшим средним различием называется «соседним кластером», поскольку он является следующим наиболее подходящим кластером для точки i.

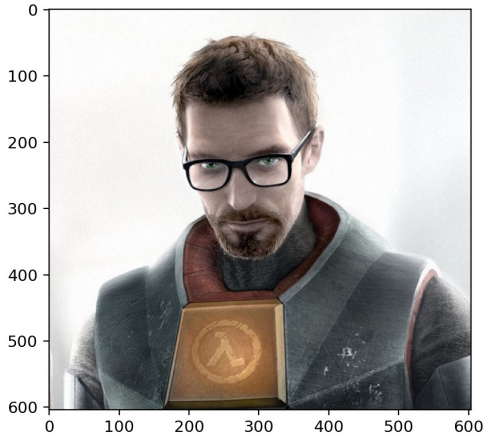
Теперь мы определим *силуэт* (значение) одной точки данных

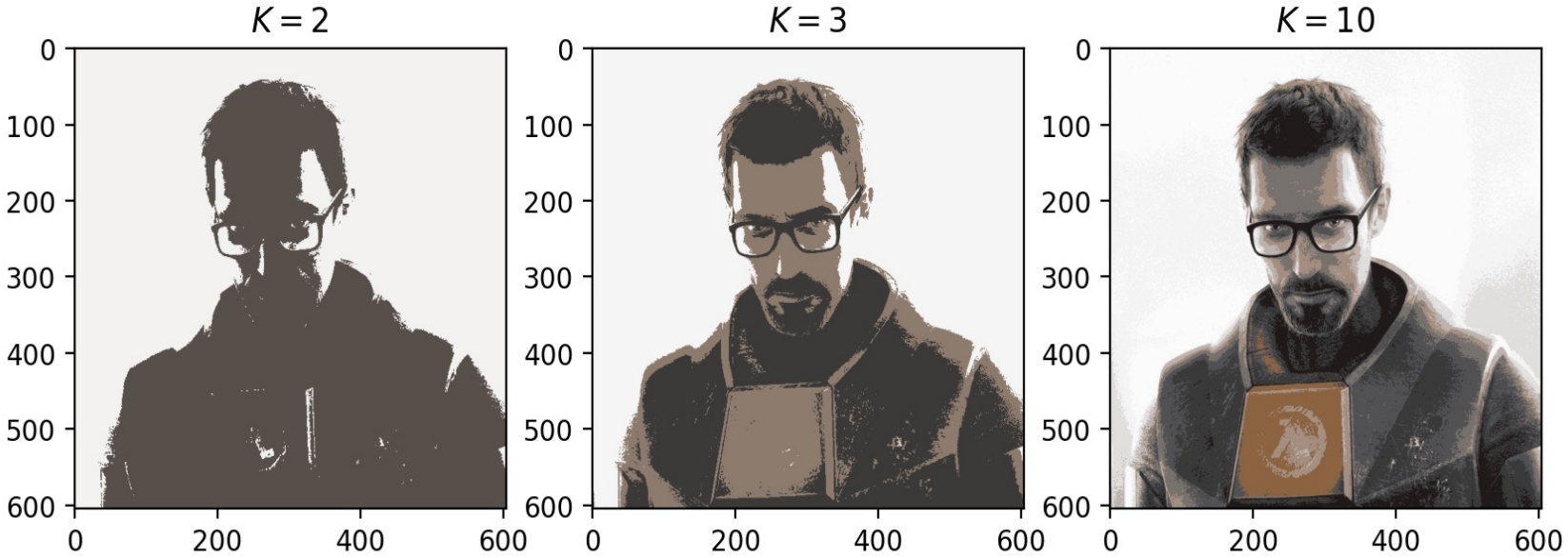
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Метод локтя


Сумма квадратов внутрикластерных расстояний (WCSS)
В этой точке WCSS перестает существенно уменьшаться
1
2
3
4
5
6
7
Количество кластеров

Сегментация и сжатие изображений





Смесь гауссианов

Распределение гауссовой смеси можно записать как линейную суперпозицию гауссианов в форме:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Затем мы вводим K-мерную двоичную случайную величину, $\mathbf{z}$ имеющую представление типа 1 из K. Маргинальное распределение по $\mathbf{z}$ определяется через коэффициенты смешивания π_k, такие, что

$$p(z_k = 1) = \pi_k$$

Поскольку $\mathbf{z}$ используется представление 1 $из$ K, мы также можем записать это распределение в форме

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Смесь гауссианов

Условное распределение для $\mathbf{x}$данного конкретного значения $\mathbf{z}$является гауссовым,

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

или

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}\,(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Совместное распределение задается как $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$, а предельное распределение $\mathbf{x}$затем получается путем суммирования совместного распределения по всем возможным состояниям , $\mathbf{z}$чтобы получить,

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p\,(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} \pi_k^{z_k}\, \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} = \sum_{k=1}^{K} \pi_k\, \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Так как суммирование по $\mathbf{z}$ фактически состоит из K членов и k-й член соответствует z_kравенству 1. Более того, для k-го члена произведение сократится до $\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Смесь гауссианов

Эта формулировка гауссовой смеси, включающая явную скрытую переменную, позволяет работать с совместным распределением $p(\mathbf{x}, \mathbf{z})$ вместо маргинального распределения $p(\mathbf{x})$, что приводит к значительным упрощениям за счет введения алгоритма максимизации ожидания (EM).

Другая важная величина — условная вероятность $p(\mathbf{z}|\mathbf{x})$. Мы будем использовать $\gamma(z_k)$ для обозначения $p(z_k = 1|\mathbf{x})$, значение которой можно найти с помощью теоремы Байеса.

$$\gamma(z_k) = p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p\,(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

что также можно рассматривать как *ответственность (responsibility)*, которую компонент k берет на себя за *объяснение* наблюдения $\mathbf{x}$.

Максимальное правдоподобие

Предположим, что нам дан набор данных наблюдений $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, и мы хотим смоделировать эти данные, используя смесь гауссовских распределений. Мы можем представить набор данных в виде матрицы $\mathbf{X}$ размера $N \times D$.

Аналогично, соответствующие скрытые переменные могут быть обозначены матрицей $\mathbf{Z}$ размера $N \times K$. Если предположить, что точки данных являются независимыми, то логарифм функции правдоподобия для модели смеси гауссовых распределений определяется как

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

Максимизация логарифмической функции правдоподобия для модели смеси гауссовых функций оказывается более сложной задачей, чем для случая одной гауссовой функции. Сложность возникает из-за наличия суммирования по k, которое появляется внутри логарифма, так что логарифмическая функция больше не действует непосредственно на гауссову функцию. Если мы попытаемся приравнять производные логарифмической функции правдоподобия к нулю, мы не получим решение в замкнутой форме.

Один из подходов заключается в применении оптимизации на основе градиента.

EM для гауссовых смесей

Установим производные логарифмического правдоподобия относительно средних значений гауссовых компонент равными нулю. Получим:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{Z}) = 0 \Leftrightarrow \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln\left\{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right\} = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = 0$$

$$\overset{(e^z)' = e^z z'}{\Leftrightarrow} -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\Leftrightarrow -\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\overset{\times \boldsymbol{\Sigma}_k^{-1}}{\Leftrightarrow} -\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\mu}_k - \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n = 0$$

$$\Leftrightarrow \boldsymbol{\mu}_k = \frac{1}{\sum_{n=1}^{N} \gamma(z_{nk})} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

Обратите внимание, что среднее значение $\boldsymbol{\mu}_k$ для k-го гауссовского компонента получается путем взятия взвешенного среднего значения всех точек в наборе данных, в котором весовой коэффициент для точки данных $\mathbf{x}_n$ задается апостериорной вероятностью $\gamma(z_{nk})$ того, что компонент k был ответственен за генерацию $\mathbf{x}_n$.

EM для гауссовых смесей

Далее, устанавливая производную логарифмического правдоподобия по отношению к $\boldsymbol{\Sigma}_k$ нулю и используя результат для решения задачи максимального правдоподобия для ковариационной матрицы одного гауссовского распределения, получаем,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{Z}) = 0 \& \Leftrightarrow \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln \left\{ \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = 0$$

$$\overset{(e^z)' = e^z z'}{\Leftrightarrow} -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} (\boldsymbol{\Sigma}_k - (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T) = 0$$

$$\Leftrightarrow -\sum_{n=1}^{N} \gamma\,(z_{nk})(\boldsymbol{\Sigma}_k - (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T) = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \gamma\,(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T - \sum_{n=1}^{N} \gamma\,(z_{nk})\boldsymbol{\Sigma}_k = 0$$

$$\Leftrightarrow \boldsymbol{\Sigma}_k = \frac{1}{\sum_{n=1}^{N} \gamma\,(z_{nk})} \sum_{n=1}^{N} \gamma\,(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

EM для гауссовых смесей

Приравнивание производной логарифмического правдоподобия по π_k к нулю требует соблюдения ограничения (сумма коэффициентов смешивания должна быть равна единице). Этого можно добиться, используя множитель Лагранжа и максимизируя его, что даёт:

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

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

Обратите внимание, что эти результаты не представляют собой решение в аналитической форме для параметров модели смеси, поскольку функции $\gamma(z_{nk})$ зависят от этих параметров сложным образом. Однако эти результаты предлагают простую итеративную схему поиска решения, которая оказывается примером алгоритма EM для частного случая модели смеси Гаусса.

EM-алгоритм для гауссовых смесей

Сначала выбираем начальные значения для средних значений, ковариаций и коэффициентов смешивания. Затем мы чередуем два этапа обновления: этап E и этап M. На этапе E мы используем текущие значения параметров для оценки апостериорных вероятностей или ответственности. Затем, на этапе M, мы используем эти вероятности для повторной оценки средних значений, ковариаций и коэффициентов смешивания. Обратите внимание, что на этапе M мы сначала оцениваем новые средние значения, а затем используем их для нахождения ковариаций.

1. Инициализируйте средние значения $\boldsymbol{\mu}_k$, ковариации $\boldsymbol{\Sigma}_k$ и коэффициенты смешивания π_k и оцените логарифм правдоподобия.

2. **E-шаг:** Оцените обязанности, используя текущие значения параметров

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3. **M-шаг:** переоценка параметров с использованием текущих обязанностей

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma\,(z_{nk}) \mathbf{x}_n$$

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

$$\pi_k^{new} = \frac{N_k}{N}$$

$$N_k = \sum_{n=1}^{N} \gamma\,(z_{nk})$$

4. Оценить логарифм правдоподобия $\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$

5. Повторяйте до тех пор, пока не будет достигнута сходимость либо параметров, либо логарифмического правдоподобия.

EM на датасете Old Faithful