План лекции

1. Смесь Гауссовских распределений
2. EM-алгоритм
3. Информационные методы

Смешанные модели

- Пусть $\mathbb{W}$ — подмножество Евклидового пространства $\mathbb{R}^d$ и $|\mathbb{W}| = N < \infty$
- Рассмотрим модель, в которой распределение $\mathrm{P}(\mathbb{W})$ множества $\mathbb{W}$ задается смесью распределений: $\mathrm{P}(\mathbb{W}) = \sum_{i=1}^{k} p_i \mathrm{P}(\mathbb{X}_i)$, где p_i — вероятности кластеров и $\mathrm{P}(\mathbb{X}_i)$ — распределение кластера для всех $i \in 1..k$
- Распределения кластеров $\mathrm{P}(\mathbb{X}_i)$ могут быть разного вида: Бернулли, Пуассона, Нормальное и т.д.
- Для решения задачи кластеризации необходимо построить оценки параметров распределений смеси на основе входных данных $\mathbf{w} = (w_1, \ldots, w_N)$. Для этого обычно применяют техники, основанные на скрытых параметрах, такие как алгоритмы Expectation Maximization (EM) и Gibbs sampling

Смесь Гауссовских распределений

- Смесь Гауссовских распределений, *Gaussian Mixture Model* (GMM)
  $f(\mathcal{X}^k, \mathbf{w}) = \sum_{i=1}^{k} p_i \mathcal{G}(\mathbf{w}|\mathbf{x}_i, \Gamma_i)$, где $\mathcal{G}(\mathbf{w}|\mathbf{x}_i, \Gamma_i)$ — плотность Гауссовского распределения со средним $\mathbf{x}_i$ и ковариационной матрицей Γ_i, $i \in 1..k$
- $\Gamma_i = \beta_i \mathbf{U}_i \mathbf{D}_i \mathbf{U}_i^{\mathrm{T}}$, $i \in 1..k$, где β_i — константа объема ковариационной матрицы, $\mathbf{U}_i$ — матрица собственных векторов, определяющая ориентацию кластера, $\mathbf{D}_i = \mathrm{diag}(\lambda_{1,i}, \ldots, \lambda_{d,i})$ — матрица собственных чисел, определяющая форму ковариационной матрицы, $\lambda_{d,i} \leq \lambda_{d-1,i} \leq \ldots \leq \lambda_{1,i} = 1$

Виды ковариационных матриц

	Γ_i	Форма	Ориентация	Объем
1	$\beta \mathbf{I}$	Сферическая	N/A	Одинаковый
2	$\beta_i \mathbf{I}$	Сферическая	N/A	Разный
3	Γ	Одинаковая	Одинаковая	Одинаковый
4	$\beta_i \Gamma$	Одинаковая	Одинаковая	Разный
5	$\beta \mathbf{U}_i \mathbf{D} \mathbf{U}_i^{\mathrm{T}}$	Одинаковая	Разная	Одинаковый
6	$\beta_i \mathbf{U}_i \mathbf{D} \mathbf{U}_i^{\mathrm{T}}$	Одинаковая	Разная	Разный
7	$\beta_i \mathbf{U} \mathbf{D}_i \mathbf{U}^{\mathrm{T}}$	Разная	Одинаковая	Разный
8	Γ_i	Разная	Разная	Разный

Таблица : Виды кластеров в зависимости от ковариационной матрицы.

Применение GMM

- Финансовые модели
- Определение тем текстовых документов
- Распознавание рукописных символов
- Распознавание речи, вместе с HMM
- Машинный перевод, модель IBM 2
- Сегментация изображения
- Радиально-базисные функции (RBF)
- ...

Пример GMM

Пусть $d = 2$, $k = 2$, $\mathbf{x}_1 = (0, 0)$, $\Gamma_1 = \begin{pmatrix} 0.7 & 0.5 \\ 0.5 & 0.7 \end{pmatrix}$,

$\mathbf{x}_2 = (3, 3)$, $\Gamma_2 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$

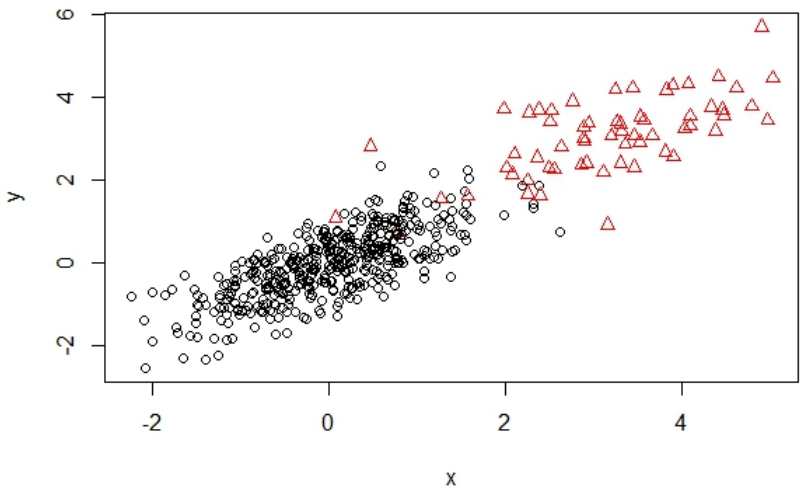

Рис. : Моделирование GMM

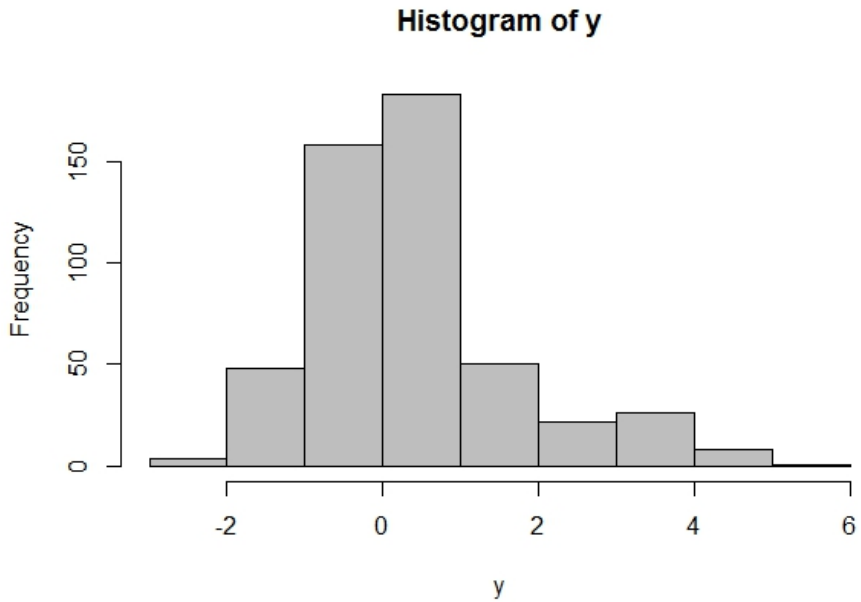

Рис. : Гистограмма одной компоненты входных данных

Метод максимума правдоподобия

- Построим оценку параметров на основе ММП
- Функция правдоподобия:

$$L(\{p_i, \mathbf{x}_i, \Gamma_i\}_{i \in 1..k}) = \prod_{\mathbf{w} \in \mathbb{W}} \left(\sum_{i=1}^{k} p_i \mathcal{G}(\mathbf{w} | \mathbf{x}_i, \Gamma_i) \right)$$

- Прологарифмируем:

$$F(\mathcal{X}^k) = -l(\{p_i, \mathbf{x}_i, \Gamma_i\}_{i \in 1..k}) = -\sum_{\mathbf{w} \in \mathbb{W}} \ln \left(\sum_{i=1}^{k} p_i \mathcal{G}(\mathbf{w} | \mathbf{x}_i, \Gamma_i) \right)$$

Конечная формула

- Обозначим $\bar{\mathbf{x}}_i,\ i \in 1..k$, как $(1 + d + d^2)$-вектор, состоящий из p_i, d-вектоа $\mathbf{x}_i$ и элементов $d \times d$ матрицы Γ_i, $\bar{\mathbf{X}}$ — $(1 + d + d^2) \times k$ матрица, состоящая из векторов $\bar{\mathbf{x}}_i,\ i \in 1..k$.
- $\bar{q}(\bar{\mathbf{x}}, \mathbf{w}) = -\ln p + (\mathbf{w} - \mathbf{x})^{\mathrm{T}} \Gamma^{-1} (\mathbf{w} - \mathbf{x})$
- Тогда, правило кластеризации:
  $\mathbb{X}_i(\bar{\mathbf{X}}) = \{\mathbf{w} \in \mathbb{W} :\ \bar{q}(\bar{\mathbf{x}}_i, \mathbf{w}) < \bar{q}(\bar{\mathbf{x}}_j, \mathbf{w}),\ j = 1, 2, \ldots, i-1,$
  $\bar{q}(\bar{\mathbf{x}}_i, \mathbf{w}) \leq \bar{q}(\bar{\mathbf{x}}_j, \mathbf{w}),\ j = i+1, i+2, \ldots, k\},\ i \in 1..k,$
- $F(\bar{\mathbf{X}}) = \sum\limits_{i=1}^{k} \sum_{\mathbf{w} \in \mathbb{X}_i(\bar{\mathbf{X}})} \bar{q}(\bar{\mathbf{x}}_i, \mathbf{w}) \to \min_{\bar{\mathbf{X}}}$

Алгоритм Expectation Maximization (*EM*)

- Для максимизации логарифма функции правдоподобия применяется *EM* алгоритм
- Алгоритм Expectation Maximization предложен Dempster, Laird и Rubin в 1977 году
- *EM* алгоритм — итеративная оптимизационная процедура, т.ч. на каждом ее шаге функция правдоподобия не уменьшается, что гарантирует сходимость к локальному максимуму функции
- $\mathbf{Z}$ — $N \times k$ матрица скрытых переменных, т.ч. для каждого $\mathbf{w}_t \in \mathbb{W}$, $t \in 1..N$ *t-ый* столбец — вектор скрытых переменных $\mathbf{z}_n = \left(z_{1,t}, \ldots, z_{k,t}\right)$, представляющий вероятности того, что $\mathbf{w}_t$ принадлежит каждому из k кластеров

Алгоритм Expectation Maximization (*EM*): инициализация

- *Вход:* $\mathbb{W}$ — множество для кластеризации размера N, k — число кластеров, $\widehat{\mathcal{X}}_k(0)$ — начальное разбиение (опционально)
- *Выход:* Разбиение $\mathcal{X}^k$ множества $\mathbb{W}$ на k кластеров
- *Инициализация:* Инициализация параметров смеси. Матрица $\mathbf{Z}_0$ может быть инициализирована случайным присвоением 1 одному элементу в каждом столбце. В случае, если $\widehat{\mathcal{X}}_k(0)$ дано, то модель определяется этим разбиением

Алгоритм Expectation Maximization (*EM*): M-шаг

a. выборочный размер кластера

$$\widehat{S}_i = \sum_{t=1}^{N} z_{i,t}.$$

b. выборочная вероятность кластера

$$\widehat{p}_i = \frac{\widehat{S}_i}{N}.$$

c. выборочное среднее кластера

$$\widehat{\mathbf{x}}_i = \frac{1}{\widehat{S}_i} \sum_{t=1}^{N} z_{i,t} \mathbf{w}_t.$$

d. Выборочная ковариационная матрица $\widehat{\Gamma}_i$ вычисляется для каждого кластера.

Алгоритм Expectation Maximization (*EM*): E-шаг и критерий остановки

- *E-шаг:* Вычисление постериорных вероятностей по Байесовскому правилу:

$$z_{i,t} = \frac{\widehat{p}_i \mathcal{G}\left(\mathbf{w}_t | \widehat{\mathbf{x}}_i, \widehat{\Gamma}_i\right)}{\sum_{i=1}^{k} \widehat{p}_i \mathcal{G}\left(\mathbf{w}_t | \widehat{\mathbf{x}}_i, \widehat{\Gamma}_i\right)}$$

- *Критерий остановки:* Остановка, если старая и новая модели достаточно близки, иначе — M-шаг

Кластеризация при помощи *EM*-алгоритма

- Критерий кластеризации (минус функция правдоподобия):
  $cl(\bar{\mathbf{X}}) = \sum_{t=1}^{N} \sum_{i=1}^{k} z_{i,t} \bar{q}(\bar{\mathbf{x}}_i, \mathbf{w}_t)$
- На каждом *EM*-шаге критерий кластеризации уменьшается, и процесс сходится к локальному минимуму за конечное число итераций
- В случае, когда ковариационные матрицы одинаковы $\Gamma_i = \sigma^2 \mathbf{I},\ i \in 1..k$), алгоритм кластеризации *EM* является аналогом алгоритма k-Средних. Таким образом, *EM* алгоритм — обобщение алгоритма k-Средних
- Существуют рандомизированные модификации *EM* алгоритм для кластеризации, имеющие преимущества в скорости и устойчивости к помехам в измерениях

Пример кластеризации при помощи *EM*-алгоритма

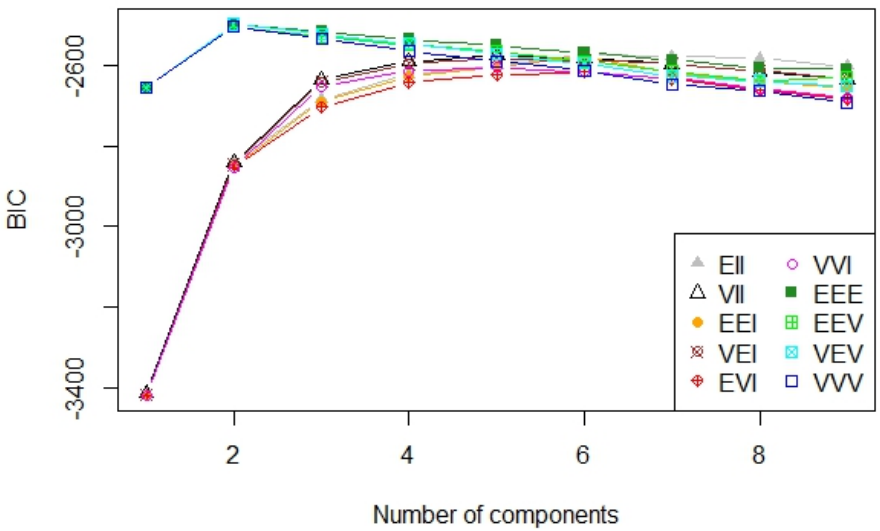

Рис. : Оценка ковариационной матрицы

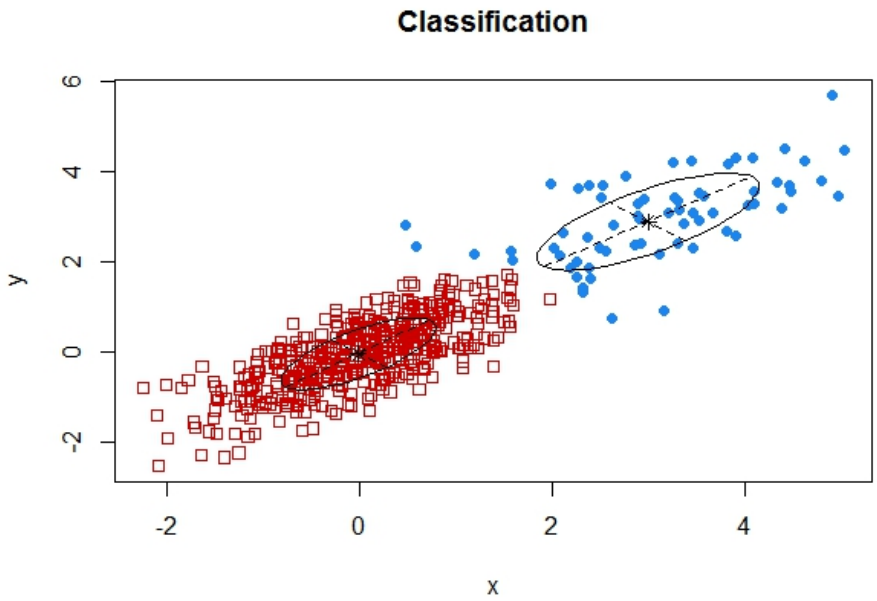

Рис. : Результат работы *EM*-алгоритма

Плюсы и минусы GMM

Плюсы GMM:

- Распространенность модели
- Гибкость структуры кластера, определяемой ковариационной матрицей
- Строгий статистический вывод для полной модели
- Алгоритм k-Средних часто является быстрым

Минусы GMM:

- Выбор модели сложен. Данные могут не являться Гауссовскими.
- Сложность оценки ковариационной матрицы
- Проблема локального минимума
- Плохо работает в случае невыпуклых кластеров

Информационные методы

- Основное предположение: каждый элемент $\mathbb{W}$ принадлежит каждому кластеру с определенной вероятностью. Основной целью кластеризации является нахождение оптимальных вероятностей. Такой подход называется *нечеткая кластеризация (fuzzy clustering)*
- С точки зрения теории информации кластеризация — методика сжатия данных с потерями
- Для двух дискретных случайных величин X и Y, принимающих значения $\{x_i\}$, $i = 1, 2, ...$ и $\{y_j\}$, $j = 1, 2, ...$, соответственно, рассмотрим взаимную информацию:
$I(X, Y) = \sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} = \sum_{i,j} p(x_i, y_j) \log_2 \frac{p(y_j|x_i)}{p(y_j)}$

Информационные методы (кластеризация)

- Кластеризация сжимает исходные данные с помощью отбрасывания менее значимой информации. Для измерения значимости берется мера различия $q_i(\mathbf{w}', \mathbf{w}'')$, $i \in 1..k$
- Сжатие с потерями реализуется минимизацией взаимной информации: $I(\mathcal{X}^k, \mathbb{W}) = \sum_{\mathbf{w},i} P(\mathbb{X}_j|\mathbf{w})P(\mathbf{w}) \log_2 \frac{P(\mathbb{X}_i|\mathbf{w})}{P(\mathbb{X}_i)}$
- Минимизация ограничена фиксированной мерой различия: $F(\mathcal{X}^k, \mathbb{W}) = \sum_{\mathbf{w},i} P(\mathbb{X}_i|\mathbf{w})P(\mathbf{w})q_i(\mathbf{x}_i, \mathbf{w})$,
  где $\mathbf{x}_i$ — представление (центорид) кластера $\mathbb{X}_i$

Информационные методы (свободная энергия)

- Формальное решение задачи достигается с помощью распределения Больцмана
  $P(\mathbb{X}_i|\mathbf{w}) = \frac{P(\mathbb{X}_i)}{Z(\mathbf{w},C)} \exp\left(-\frac{q_i(\mathbf{x}_i,\mathbf{w})}{C}\right)$, где
  $Z(\mathbf{w},C) = \sum_i P(\mathbb{X}_i) \exp\left(-\frac{q_i(\mathbf{x}_i,\mathbf{w})}{C}\right)$ — нормализационная константа и C — множитель Лагранжа
- При обычной кластеризации распределение центоридов кластеров: $P(\mathbf{X}) = \frac{e^{-\frac{F}{C}}}{\sum_Y e^{-\frac{F}{C}}}$, где $F = -C \sum_{\mathbf{w}} \ln\left(\sum_i \exp\left(-\frac{q_i(\mathbf{x}_i,\mathbf{w})}{C}\right)\right)$ — *свободная энергия* разбиения на кластеры
- Оптимальные параметры кластеризации находятся минимизацией свободной энергии

Информационные методы (частный случай)

- Пусть $q_i(\mathbf{x}_i, \mathbf{w}) = (\mathbf{w} - \mathbf{x}_i)\Gamma_i^{-1}(\mathbf{w} - \mathbf{x}_i)$
- Тогда $F = -C \sum_{\mathbf{w}} \ln \left(\sum_i \exp \left(-\frac{(\mathbf{w}-\mathbf{x}_i)\Gamma_i^{-1}(\mathbf{w}-\mathbf{x}_i)}{C} \right) \right)$
- При фиксированном Γ_i из уравнений $\frac{\partial F}{\partial \mathbf{x}_i} = 0,\ i \in 1..k$ получаем результат $\mathbf{x}_i = \frac{\sum_{\mathbf{w}} \mathbf{w} P(\mathbb{X}_j|\mathbf{w})}{\sum_{\mathbf{w}} P(\mathbb{X}_j|\mathbf{w})},\ i \in 1..k$
- Данный результат непосредственно связан с результатом, полученным по ММП с помощью *EM* алгоритма

Information bottleneck

- Information bottleneck method — техника кластеризации, основанная на том, что каждый кластер отражает относительную информацию внутри данных
- $p(w_i|\mathbf{w}) = \frac{w_i}{\sum_{i=1}^{d} w_i}$
- Качество кластеризации определяется с помощью взаимной информации $I(S;Y)/I(X;Y)$
- В качестве мер различия рассматривают расстояние Кульбака-Леейблера (Kullback–Leibler divergence) $D_{KL}(p(y|x)||p(y|s))$, расстояние Йенсена-Шеннона (Jensen-Shannon divergence) $DJS(x,s) = (p(x) + p(s)) * D_{JS}(p(y|x), p(y,s))$

Конец

Спасибо за внимание!