Исследование и практическая реализация EM-алгоритма для задачи кластеризации (GMM)

Курсовой проект

Студент: Санчук Сергей Александрович
Руководитель: Буславский Александр Андреевич

Белорусский государственный университет
ФПМИ, Кафедра ДМА

Минск, 2025

Введение

Цель работы: Теоретическое исследование EM-алгоритма и разработка программной реализации модели GMM для кластеризации данных сложной структуры.

Основные задачи:

- Провести теоретический анализ модели смеси гауссовых распределений (GMM).
- Вывести формулы EM-алгоритма через вариационную нижнюю границу (ELBO).
- Реализовать алгоритм на Python с поддержкой полной ковариационной матрицы.
- Сравнить эффективность GMM и K-means на синтетических и реальных данных.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Математическая модель GMM

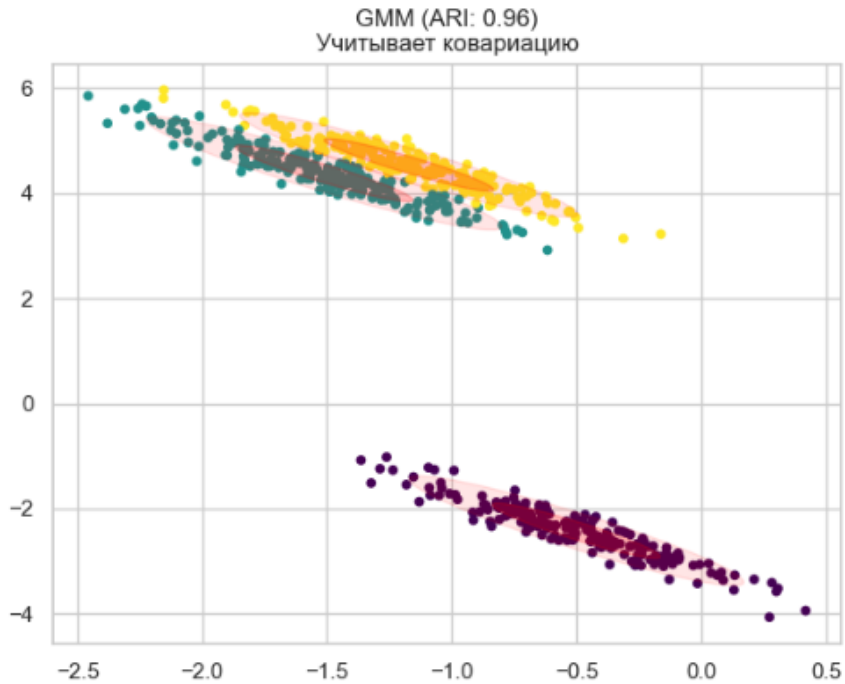
Плотность вероятности наблюдаемой переменной x:

Формула смеси

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Параметры θ:

- π_k — вес компоненты $(\sum \pi_k = 1)$;
- μ_k — центр кластера;
- Σ_k — ковариационная матрица.



Метод максимального правдоподобия (MLE)

Необходимо максимизировать логарифм функции правдоподобия:

$$\ln p(X|\theta) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right) \to \max_{\theta}$$

Проблема: Сумма находится под знаком логарифма ($\ln \sum$).

- Аналитическое решение невозможно.
- Необходим итерационный метод (EM-алгоритм).

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Вариационная нижняя граница (ELBO)

Для формализации задачи вводятся латентные переменные Z и вариационное распределение $q(Z)$.
Используя декомпозицию правдоподобия и неотрицательность KL-дивергенции ($KL \geq 0$), получаем нижнюю оценку:

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

Раскроем структуру ELBO (как разность Энергии и Энтропии):

Структура ELBO

$$\mathcal{L}(q, \theta) = \underbrace{\mathbb{E}_q[\ln p(X, Z|\theta)]}_{\text{Energy (Ожидаемое правдоподобие)}} - \underbrace{\mathbb{E}_q[\ln q(Z)]}_{\text{Entropy (Энтропия)}}$$

Суть EM-алгоритма: Мы итеративно максимизируем этот функционал, приближая нижнюю границу к истинному правдоподобию.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Общая схема итерации EM

Алгоритм состоит из попеременной оптимизации границы $\mathcal{L}(q, \theta)$:

E-step (Optimization w.r.t. q)

Фиксируем $\theta^{(t)}$. Находим распределение q, максимизирующее ELBO (что эквивалентно минимизации KL-дивергенции):

$$q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)}) = \arg\min_q KL(q||p)$$

Решение: $q^{(t+1)}(Z) = p(Z|X, \theta^{(t)})$ (апостериорное распределение).

M-step (Optimization w.r.t. θ)

Фиксируем $q^{(t+1)}$. Находим параметры θ, максимизирующие Энергию:

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{q^{(t+1)}}[\ln p(X, Z|\theta)]$$

Скрытые переменные в задаче GMM

Что выбрать в качестве скрытых переменных?

- $z_{nk} \in \{0, 1\}$ — индикатор: принадлежит ли объект x_n компоненте k.
- $\gamma_{nk} = p(z_{nk} = 1|x_n)$ — вероятность принадлежности.

1. На E-шаге нужно минимизировать $KL(q(Z|X)||p(Z|X, \theta))$.
2. Минимум $KL = 0$ достигается, когда $q(Z|X) = p(Z|X, \theta)$.

По формуле Байеса истинное апостериорное распределение:

$$p(z_{nk} = 1|x_n\mu_k, \Sigma_k) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} = \gamma_{nk}$$

Вывод: Полагая $q(z_{nk} = 1) = \gamma_{nk}$, мы делаем $KL = 0$ и $ELBO = \ln p(X)$.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

M-шаг для смеси гауссиан

Максимизируем ожидаемое правдоподобие:

$$Q(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \ln \left(\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right) \to \max_{\pi, \mu, \Sigma}$$

Результаты оптимизации:

1. **Веса π_k:** Используя метод множителей Лагранжа (ограничение $\sum \pi_k = 1$), получаем:

$$\pi_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}$$

2. **Параметры Гауссиан (μ_k, Σ_k):** Приравнивая производные к нулю:

$$\mu_k^{new} = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}}, \quad \Sigma_k^{new} = \frac{\sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_n \gamma_{nk}}$$

Связь с алгоритмом K-means

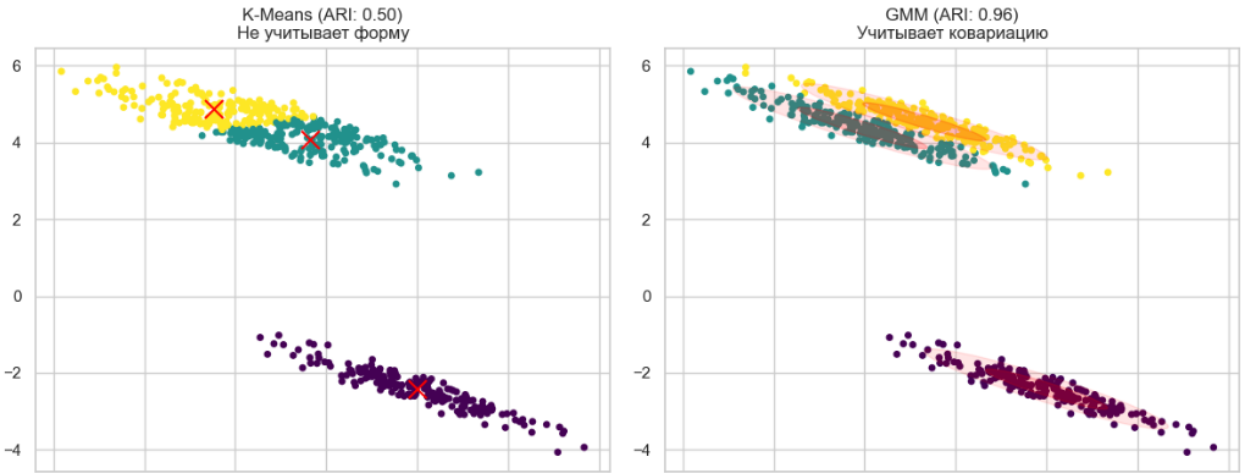
Рассмотрим GMM с ограничениями: $\Sigma_k = \epsilon I, \quad \pi_k = 1/K$.
Подставим плотность нормального распределения в формулу γ_{nk}:

$$\gamma_{nk} = \frac{\exp\left(-\frac{1}{2\epsilon}\|x_n - \mu_k\|^2\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2\epsilon}\|x_n - \mu_j\|^2\right)}$$

Предельный переход $\epsilon \to 0$**:** В сумме доминирует слагаемое с минимальным расстоянием $\|x_n - \mu\|^2$.

$$\lim_{\epsilon \to 0} \gamma_{nk} = \begin{cases} 1, & \text{если } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0, & \text{иначе} \end{cases}$$

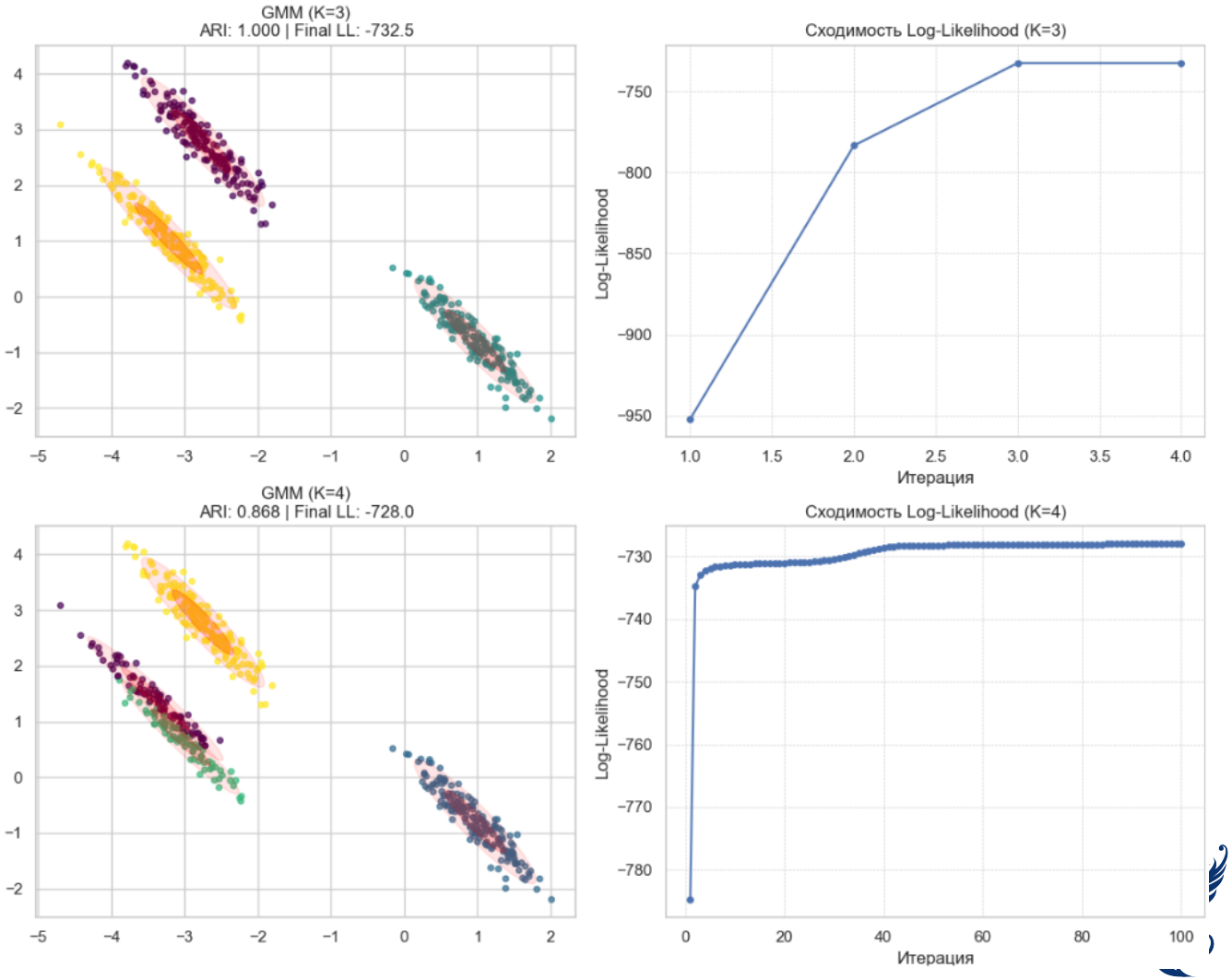
Мягкая вероятность становится жесткой меткой (Hard assignment).



БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 1: Анализ сходимости

На графике представлена зависимость Log-Likelihood от номера итерации. Алгоритм монотонно увеличивает правдоподобие.

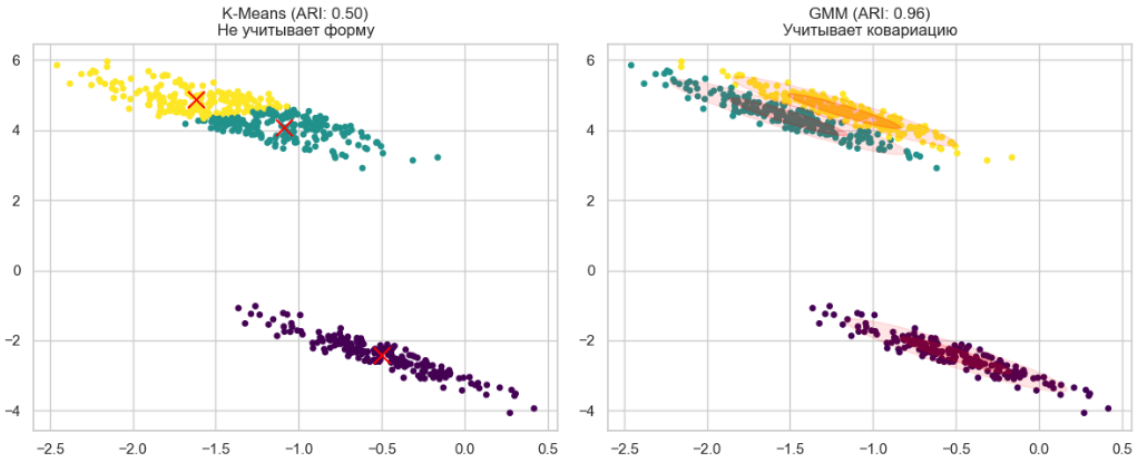


БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 2: Анизотропные данные

Сравнение на вытянутых кластерах.

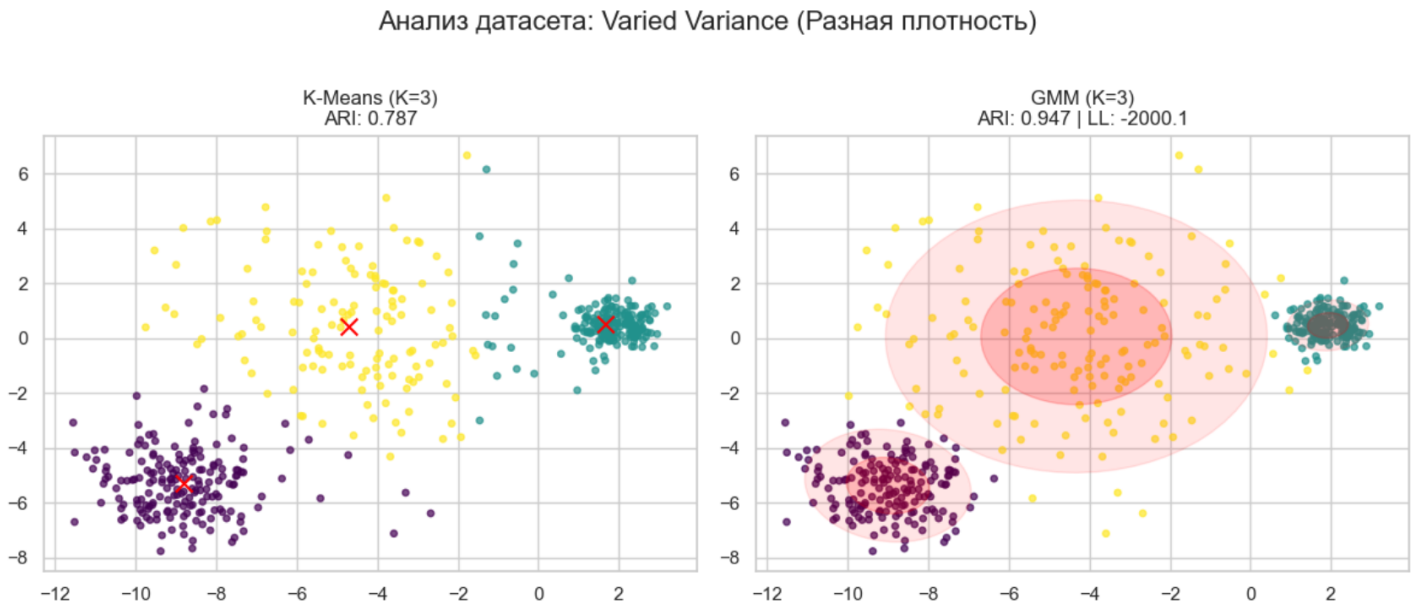
- **K-means:** «Разрезает» кластеры, так как ищет сферы.
- **GMM:** Обучает полную матрицу Σ, подстраиваясь под наклон данных.



БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 3: Различная дисперсия

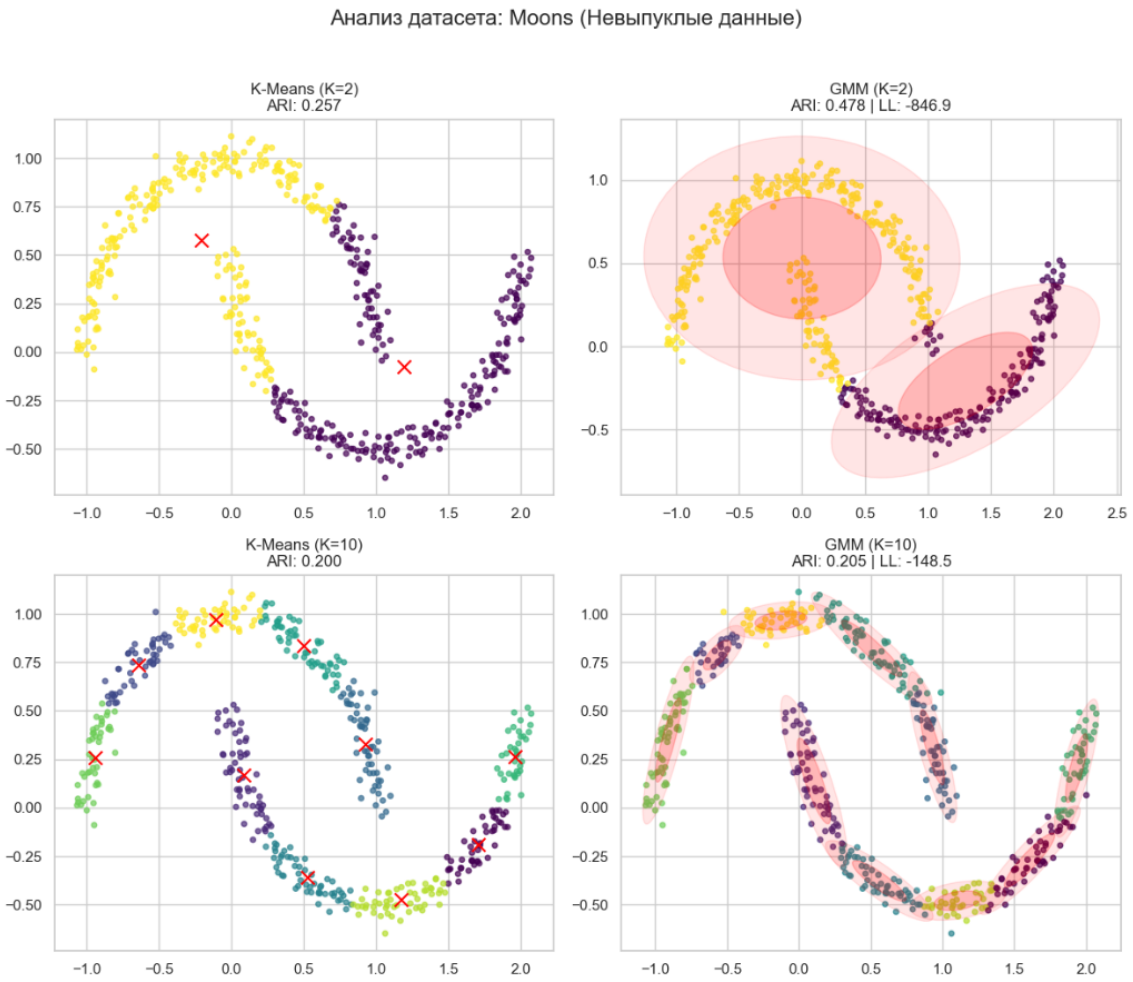
Ситуация: плотный кластер рядом с разреженным.



GMM корректно определяет границы, учитывая разный «размер» (дисперсию) кластеров.

Эксперимент 4: Невыпуклые данные

Аппроксимация датасета «Moons» ($K = 10$).

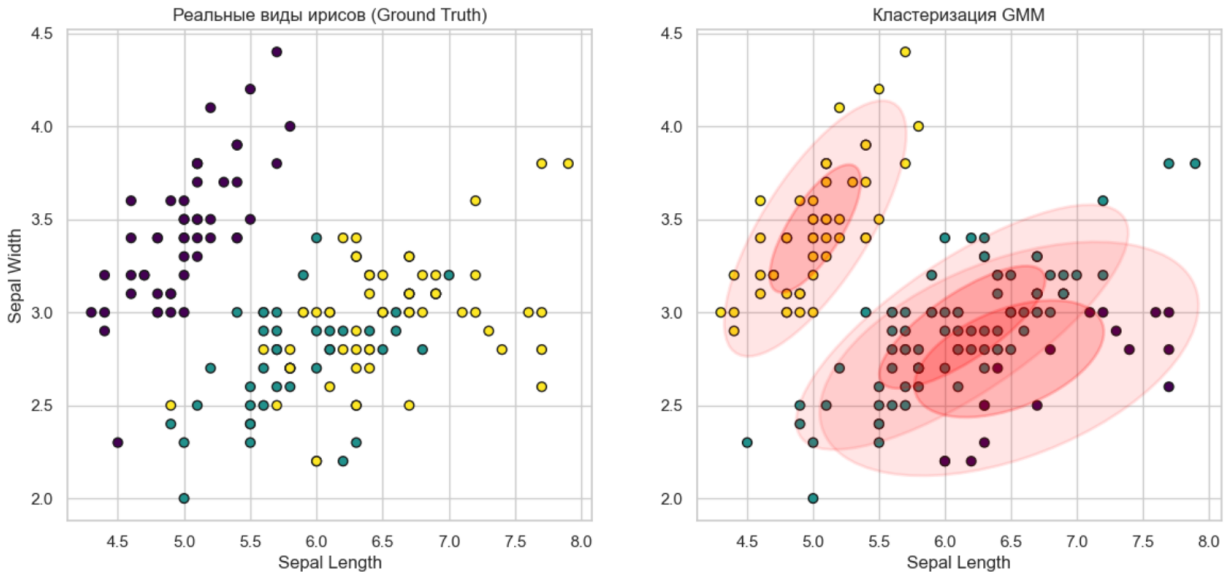


GMM может использоваться как универсальный аппроксиматор произвольной плотности вероятности.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 5: Ирисы Фишера

Визуализация кластеризации реальных данных.



Для пересекающихся классов (*Versicolor* и *Virginica*) алгоритм возвращает мягкие вероятности ($\gamma \approx 0.5$).

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Заключение

Результаты работы:

1. Изучен математический аппарат EM-алгоритма.
2. Разработана эффективная реализация на Python (Vectorization, Log-Sum-Exp trick).
3. Экспериментально подтверждено преимущество GMM перед K-means на данных сложной структуры (анизотропия, разная плотность).

Спасибо за внимание!

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ