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 (NumPy) с поддержкой полной ковариационной матрицы.
- Сравнить эффективность GMM и K-means на синтетических и реальных данных.

Математическая модель GMM

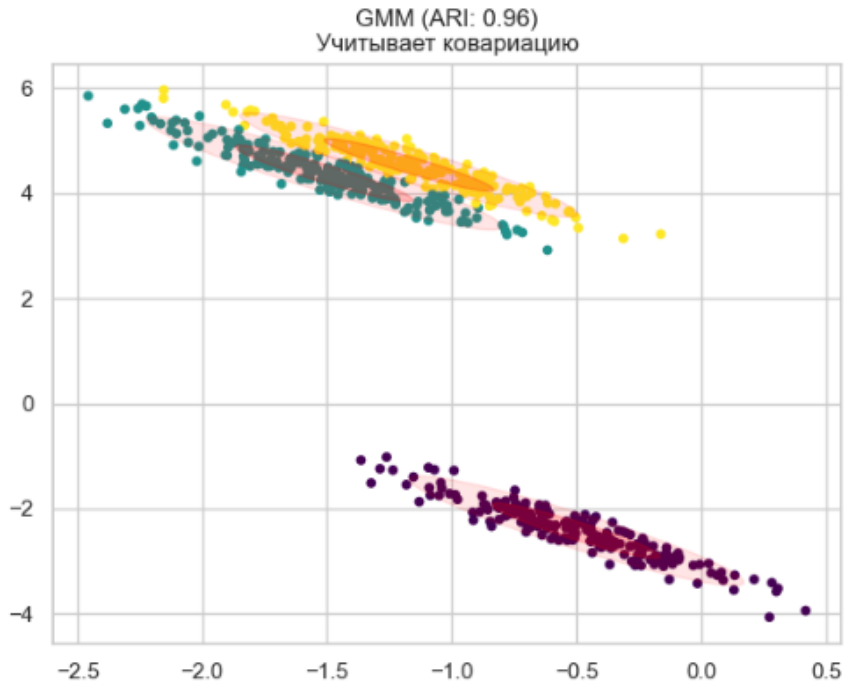
Плотность вероятности наблюдаемой переменной x:

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.



Скрытые переменные

Для формализации задачи вводятся латентные переменные Z.

- Пусть z — K-мерный бинарный вектор, указывающий, из какой компоненты взят объект.
- Априорное распределение: $p(z_k = 1) = \pi_k$.
- Условное распределение: $p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$.

Байесовский вывод: Нам неизвестны значения Z. Задача состоит в оценке апостериорной вероятности $p(Z|X)$.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

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

Декомпозиция правдоподобия

Вводим вариационное распределение $q(Z)$. Правдоподобие раскладывается на две составляющие:

Фундаментальное тождество

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

Где:

- $\mathcal{L}(q, \theta)$ — Вариационная нижняя граница (**ELBO**).
- $KL(q||p)$ — Дивергенция Кульбака-Лейблера.

Идея: Вместо прямой максимизации $\ln p(X|\theta)$, мы итеративно максимизируем нижнюю границу $\mathcal{L}(q, \theta)$.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Формальное определение ELBO и KL

Подробный вид составляющих уравнения декомпозиции:

Вариационная нижняя граница (ELBO)

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

Дивергенция Кульбака-Лейблера (KL)

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

Свойство: Согласно неравенству Гиббса, $KL(q||p) \geq 0$. Следовательно, $\mathcal{L}(q, \theta)$ действительно является **нижней оценкой** правдоподобия.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Физический смысл переменной γ

Перед переходом к E-шагу введем ключевое понятие:

Ответственность (Responsibility) γ_{nk}

Это апостериорная вероятность того, что n-е наблюдение порождено k-й компонентой смеси:

$$\gamma_{nk} \equiv p(z_k = 1 | x_n)$$

Интерпретация:

- Она показывает степень нашей уверенности в том, к какому кластеру принадлежит точка.
- В алгоритме K-means принадлежность жесткая: $\gamma \in \{0, 1\}$.
- В GMM принадлежность мягкая: $\gamma \in [0, 1]$, при этом $\sum_k \gamma_{nk} = 1$.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

E-шаг (Expectation)

Задача: Найти такое распределение $q(Z)$, которое минимизирует KL-дивергенцию (тем самым приравнивая ELBO к правдоподобию). Это достигается, когда $q(Z) = p(Z|X, \theta)$. Вычисляем «ответственности»:

Расчет γ_{nk}

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

На этом шаге параметры θ фиксированы.

M-шаг (Maximization)

Задача: Максимизировать ожидаемое правдоподобие (ELBO) по параметрам θ при фиксированных γ_{nk}.
Формулы пересчета параметров:

- **Веса:** $\pi_k^{new} = \frac{N_k}{N}$, где $N_k = \sum_n \gamma_{nk}$.
- **Средние:**

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n$$

- **Ковариации:**

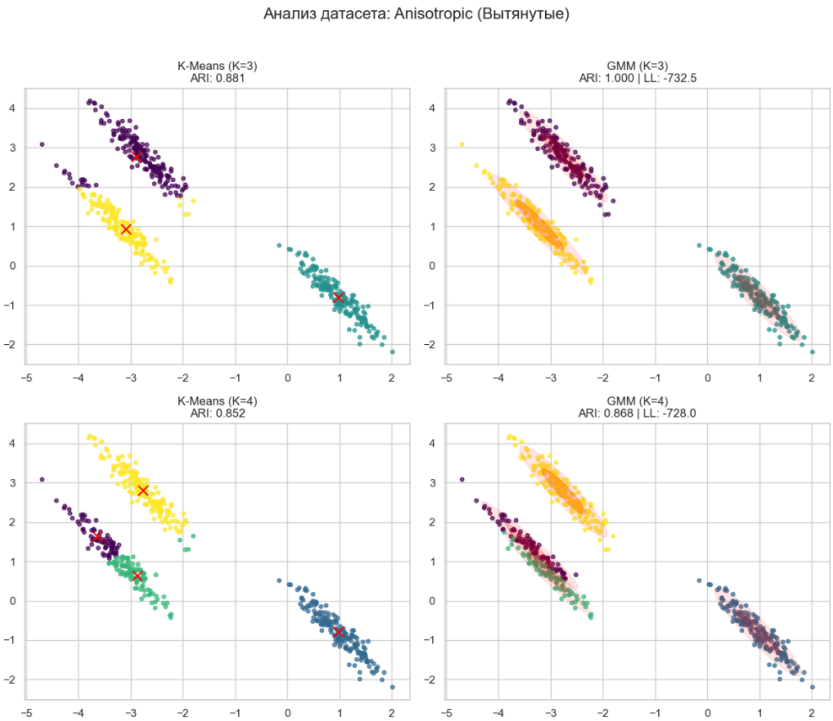
$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Связь с алгоритмом K-means

K-means является частным (предельным) случаем GMM при наложении жестких ограничений:

1. **Фиксированная ковариация:** $\Sigma_k = \epsilon I$.
2. **Фиксированные веса:** $\pi_k = 1/K$.
3. **Предел** $\epsilon \to 0$**:** Мягкая вероятность γ_{nk} превращается в жесткую метку.

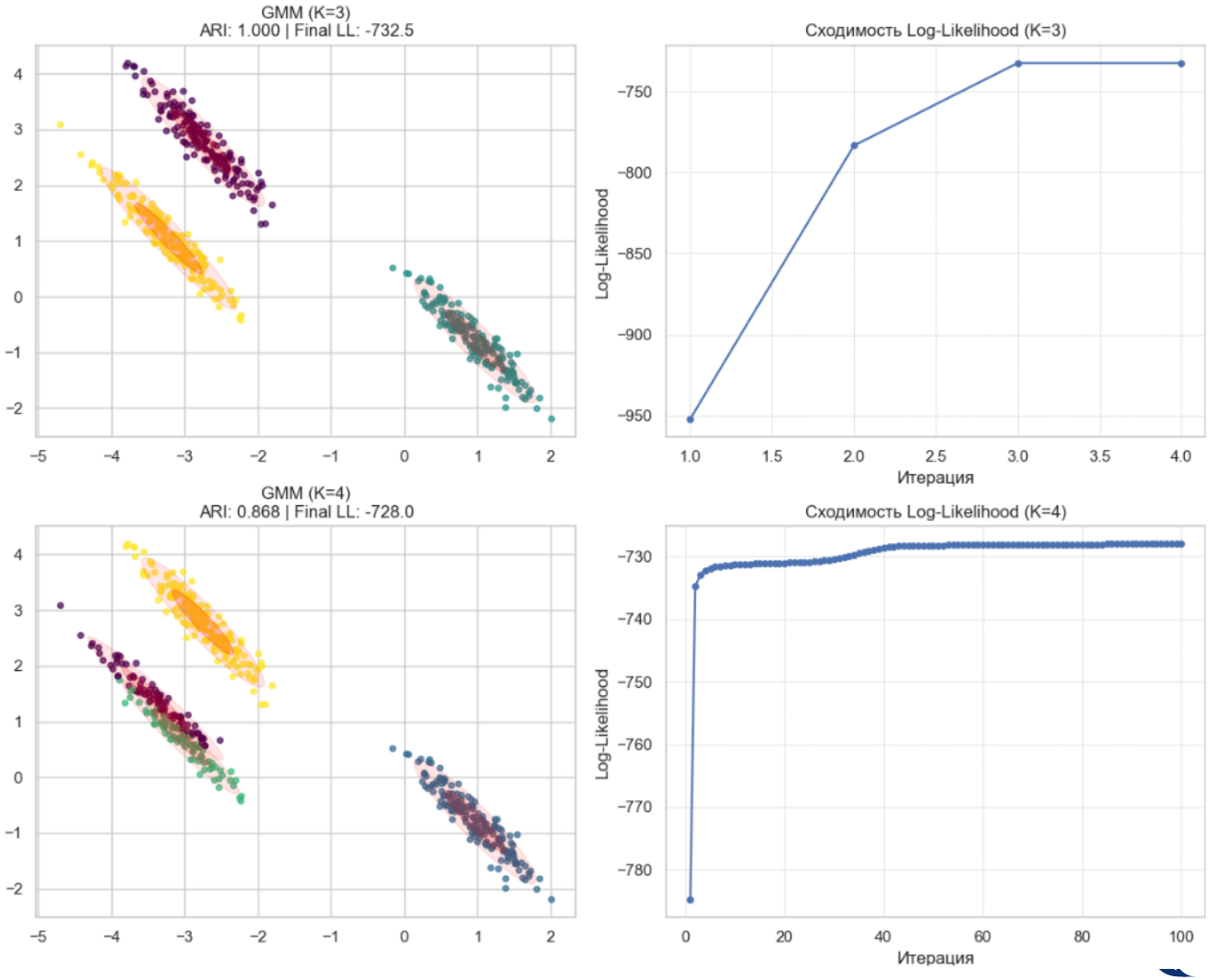


Слева: K-means. Справа: GMM.

БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 1: Анализ сходимости

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

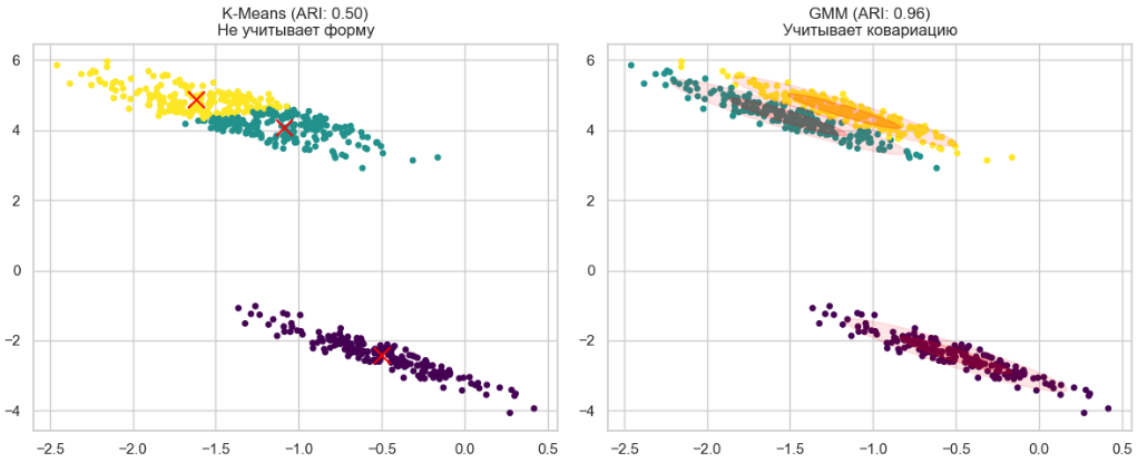


БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 2: Анизотропные данные

Сравнение на вытянутых кластерах.

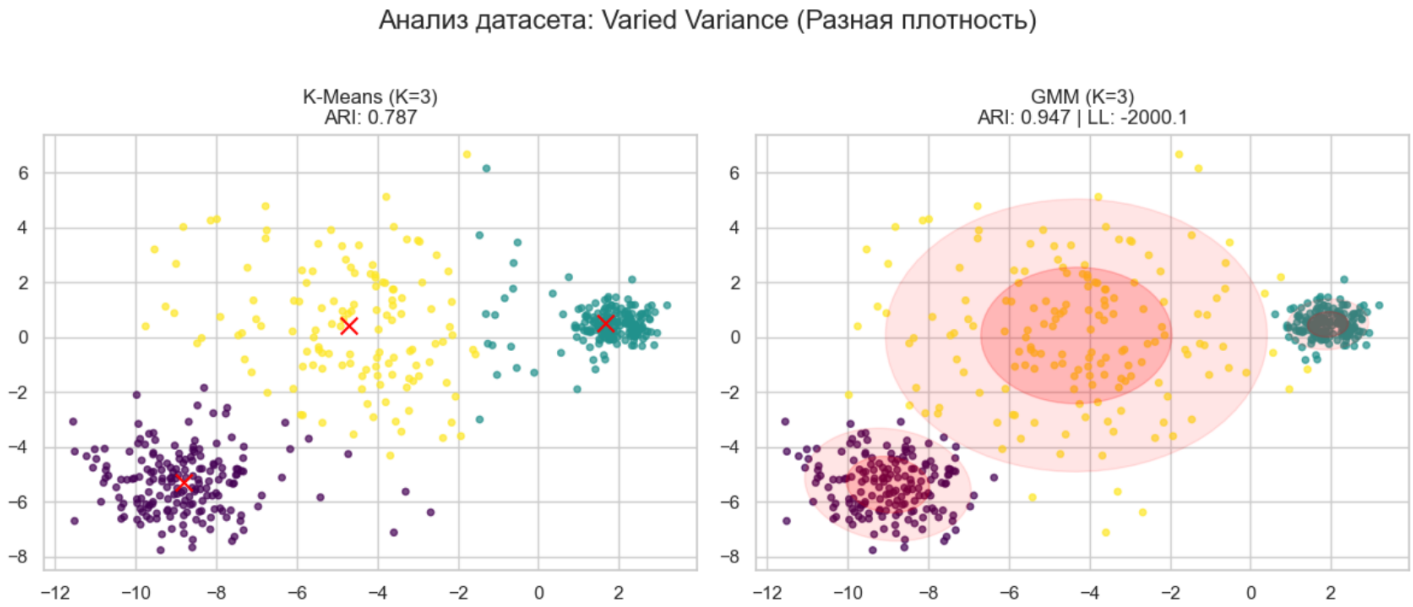
- **K-means:** «Разрезает» кластеры, так как ищет сферы.
- **GMM:** Обучает полную матрицу Σ, подстраиваясь под наклон данных.



БЕЛАРУСКІ ДЗЯРЖАЎНЫ ЎНІВЕРСІТЭТ

Эксперимент 3: Различная дисперсия

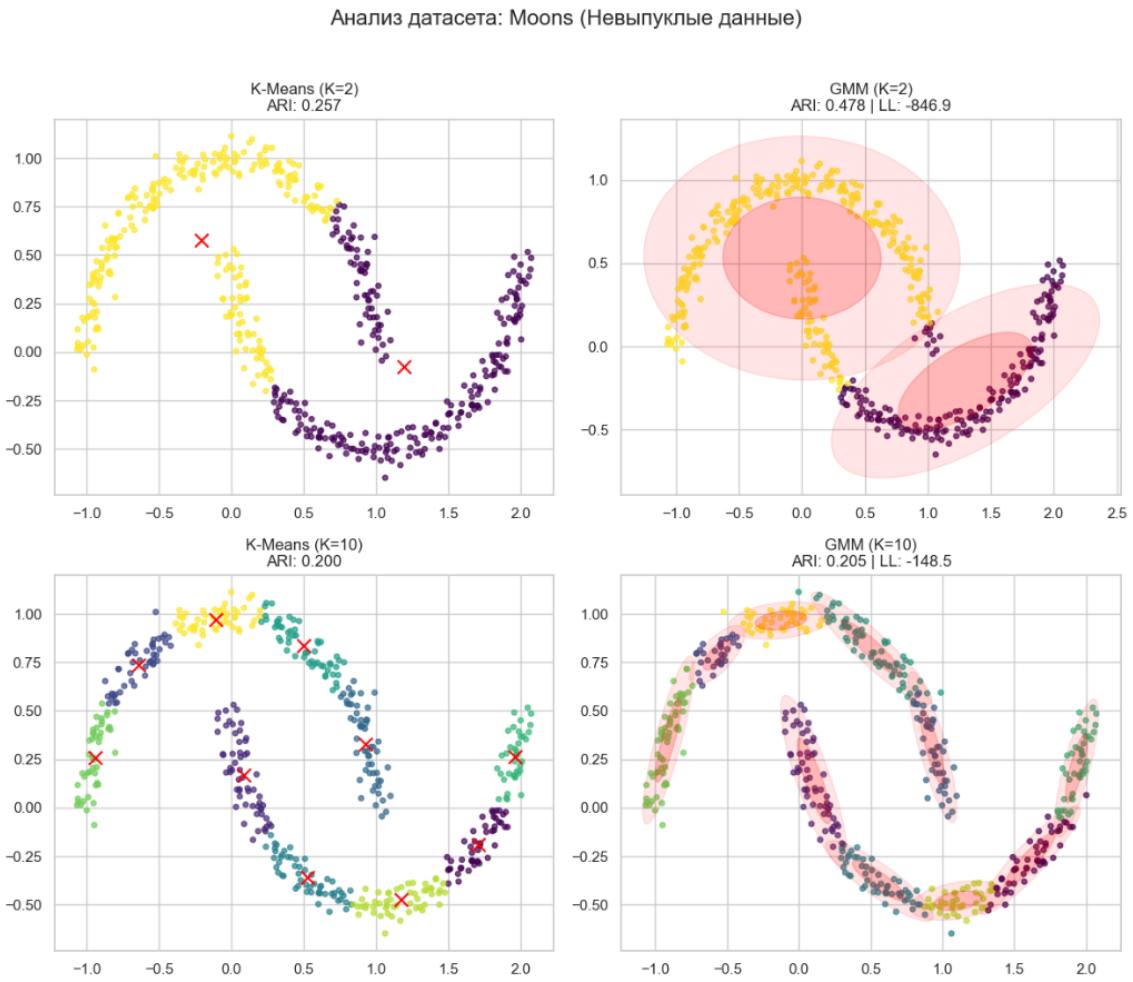
Ситуация: плотный кластер рядом с разреженным.



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

Эксперимент 4: Невыпуклые данные

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

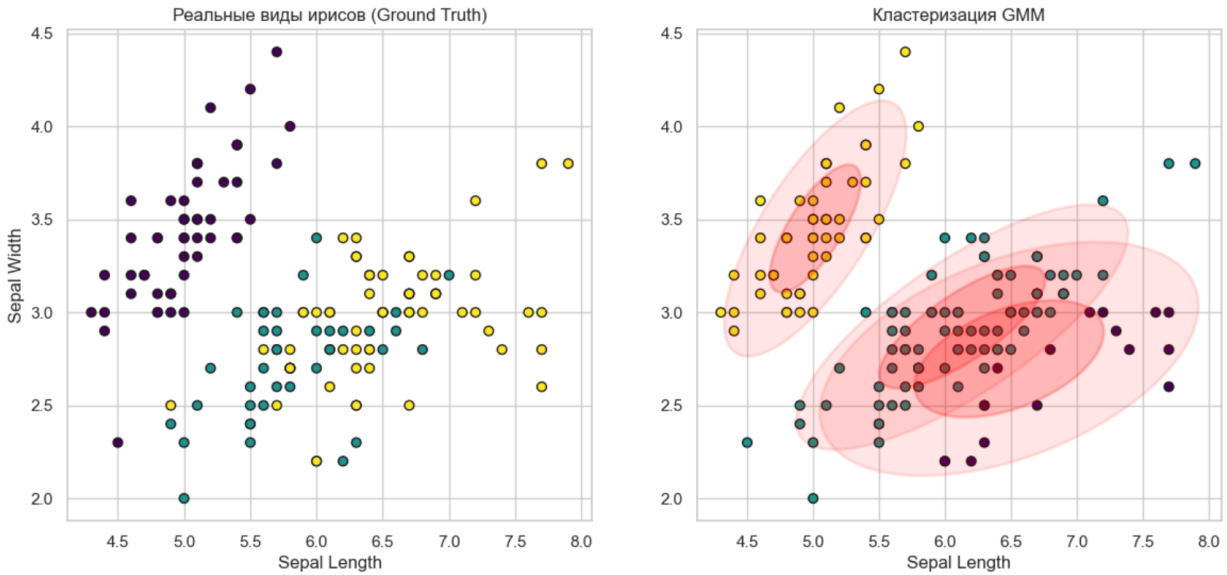


GMM может использоваться как универсальный аппроксиматор произвольной плотности вероятности.

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