Министерство образования и науки РФ
Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и кибербезопасности
Высшая школа «Компьютерные технологий и информационные системы»

ОТЧЕТ

по дисциплине «Цифровая культура»

Основы машинного обучения

Выполнил:
Студент гр. з5130902/30001 М.Е. Иванов

Проверил
Преподаватель А.С. Свистунова

Санкт-Петербург
2025 г.

Содержание

1 Введение

Данная работа посвящена анализу набора данных о гималайских экспедициях с применением современных методов многомерной статистики и машинного обучения. Целью работы является исследование структуры данных, выявление скрытых закономерностей и построение генеративных моделей, способных воспроизводить статистические свойства исходного распределения.

В ходе работы были решены следующие ключевые задачи:

- Загрузка и предварительный анализ реального набора данных «Himalayan Expeditions» из Kaggle.
- Разработка конвейера предобработки данных с автоматическим обнаружением выбросов.
- Построение и сравнение двух типов генеративных моделей: категориального многомерного нормального распределения и смешанной гауссовой модели (GMM).
- Комплексная визуализация результатов и оценка качества моделей.
- Анализ применимости различных подходов к моделированию реальных данных.

2 Описание данных

2.1 Структура набора данных

В качестве исходных данных использовался набор данных «Himalayan Expeditions», представляющий собой комплексную базу данных о экспедициях в Гималайском регионе. Набор включает:

- **Основной файл**: `exped.csv` — 11,425 записей об экспедициях с 65 атрибутами каждая
- **Временной охват**: 1905-2024 годы (119 лет наблюдений)
- **Числовые переменные**: 14 атрибутов (включая временные, высотные и количественные показатели)
- **Категориальные переменные**: 51 атрибут (включая географические, логистические и результативные показатели)

2.2 Ключевые переменные анализа

Для построения моделей были выбраны следующие переменные:

Переменная	Описание	Тип
`year`	Год проведения экспедиции	Числовая (1905-2024)
`smtdays`	Дни от базового лагеря до вершины	Числовая (0-388)
`season`	Сезон экспедиции	Категориальная (4 значения)
`totmembers`	Общее количество участников	Числовая (0-99)
`highpoint`	Максимальная достигнутая высота	Числовая (0-8850 м)

Таблица 1: Основные переменные, использованные в анализе

2.3 Распределение данных по сезонам

Анализ показал следующее распределение экспедиций по сезонам:

- **Осень (Autumn)**: 5,634 экспедиции (49.3%)
- **Весна (Spring)**: 5,334 экспедиции (46.7%)
- **Зима (Winter)**: 340 экспедиций (3.0%)
- **Лето (Summer)**: 115 экспедиций (1.0%)

3 Методология

3.1 Предобработка данных

3.1.1 Разделение выборки

Данные были разделены на обучающую и тестовую выборки в соотношении 80:20 с применением стратификации по переменной `season`:

- **Обучающая выборка**: 9,140 записей
- **Тестовая выборка**: 2,285 записей

3.1.2 Обнаружение и удаление выбросов

Для выявления аномальных наблюдений применялся алгоритм Isolation Forest со следующими параметрами:

- Доля загрязнения (contamination): 0.1
- Количество деревьев: 100
- Случайное состояние: фиксированное для воспроизводимости

Результаты обнаружения выбросов:

- Выявлено аномалий: 329 из 3,282 записей (10.0%)
- Размер очищенной выборки: 8,811 записей
- Сохранено данных: 77.1% от исходного объема

3.2 Модель многомерного нормального распределения

3.2.1 Теоретические основы

Первая модель основана на предположении о многомерной нормальности данных с разделением по категориям. Для каждой категории k (сезона) вычисляются:

$$\mu_k = E[X|C = k] \text{ — вектор математических ожиданий}$$
$$\Sigma_k = \text{Cov}[X|C = k] \text{ — ковариационная матрица}$$

Плотность вероятности:

$$f(x|k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

3.2.2 Параметры обученной модели

В результате обучения получены следующие параметры:

Весна (Spring):

$$\mu_{spring} = \begin{pmatrix} 2004.15 \\ 26.69 \end{pmatrix}, \quad \Sigma_{spring} = \begin{pmatrix} 182.71 & -20.91 \\ -20.91 & 205.20 \end{pmatrix}$$

Осень (Autumn):

$$\mu_{autumn} = \begin{pmatrix} 2002.81 \\ 15.84 \end{pmatrix}, \quad \Sigma_{autumn} = \begin{pmatrix} 145.43 & -45.41 \\ -45.41 & 115.60 \end{pmatrix}$$

Зима (Winter):

$$\mu_{winter} = \begin{pmatrix} 1994.70 \\ 20.07 \end{pmatrix}, \quad \Sigma_{winter} = \begin{pmatrix} 124.92 & -43.82 \\ -43.82 & 821.36 \end{pmatrix}$$

Лето (Summer):

$$\mu_{summer} = \begin{pmatrix} 1993.57 \\ 20.32 \end{pmatrix}, \quad \Sigma_{summer} = \begin{pmatrix} 415.86 & -160.51 \\ -160.51 & 552.06 \end{pmatrix}$$

3.3 Модель смешанного гауссова распределения (GMM)

3.3.1 Теоретические основы

Альтернативный подход основан на представлении данных в виде смеси гауссовых распределений:

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

где $K = 3$ — количество компонент, π_k — веса компонент, определяющие вероятность принадлежности к каждому кластеру.

3.3.2 Параметры обученной модели

Результаты обучения GMM с тремя компонентами:

Веса компонент: $\pi = [0.299, 0.355, 0.346]$

Средние значения:

$$\mu_1 = \begin{pmatrix} 1991.43 \\ 26.44 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 2006.22 \\ 10.29 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 2010.00 \\ 27.48 \end{pmatrix}$$

3.4 Метрики оценки качества

Качество моделей оценивалось с помощью среднего логарифма правдоподобия:

$$\text{Log-likelihood} = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i)$$

4 Результаты и обсуждение

4.1 Визуальный анализ исходных данных

4.1.1 Обнаружение выбросов

На рисунке 1 представлено распределение данных с выделенными выбросами. Красные точки показывают аномальные наблюдения, которые были исключены из дальнейшего анализа. Выбросы в основном соответствуют экстремальным значениям продолжительности экспедиций.

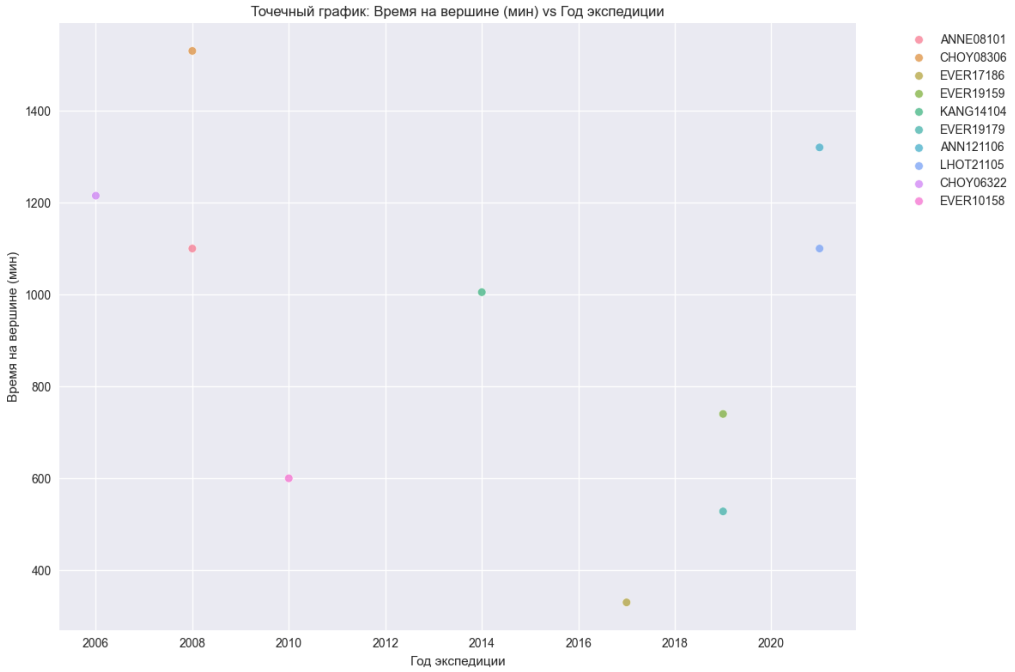

Рис. 1: Точечный график с выделенными выбросами (красные точки)

4.1.2 Распределение переменных

Рисунок 2 демонстрирует гистограммы всех числовых переменных набора данных. Наблюдается:

- Экспоненциальный рост количества экспедиций после 1980 года
- Логнормальное распределение времени до вершины и общей продолжительности
- Концентрация экспедиций на высотах 6000-8000 метров
- Преобладание малых экспедиций (2-8 участников)

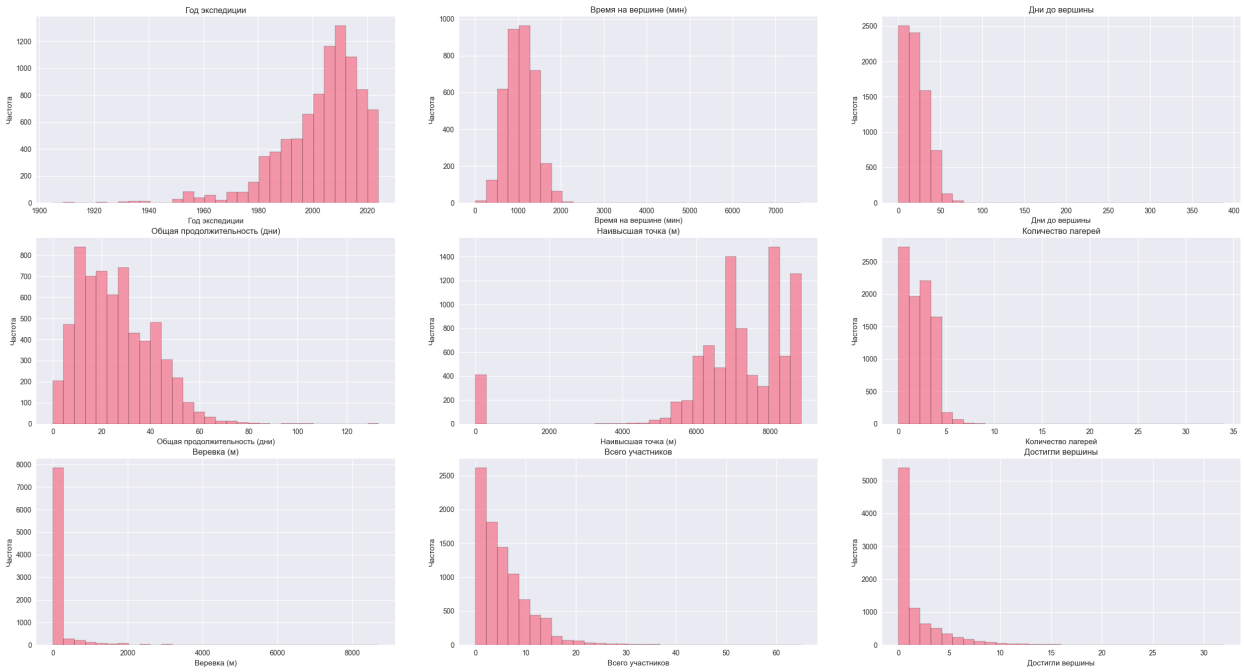

Рис. 2: Гистограммы распределения ключевых числовых переменных

4.2 Сравнение генеративных моделей

4.2.1 Модель многомерного нормального распределения

Рисунок 3 показывает сравнение исходных данных со сгенерированными с помощью категориального многомерного нормального распределения. Модель успешно воспроизводит:

- Временную динамику роста экспедиционной активности
- Различия в продолжительности экспедиций между сезонами
- Корреляционную структуру между годом и днями до вершины

Рис. 3: Сравнение исходных и сгенерированных данных (многомерное нормальное распределение)

4.2.2 Комплексное сравнение методов

На рисунке 4 представлено сопоставление всех подходов. GMM демонстрирует способность выявлять латентные кластеры в данных, не связанные напрямую с сезонностью, что может отражать различные стратегии проведения экспедиций.

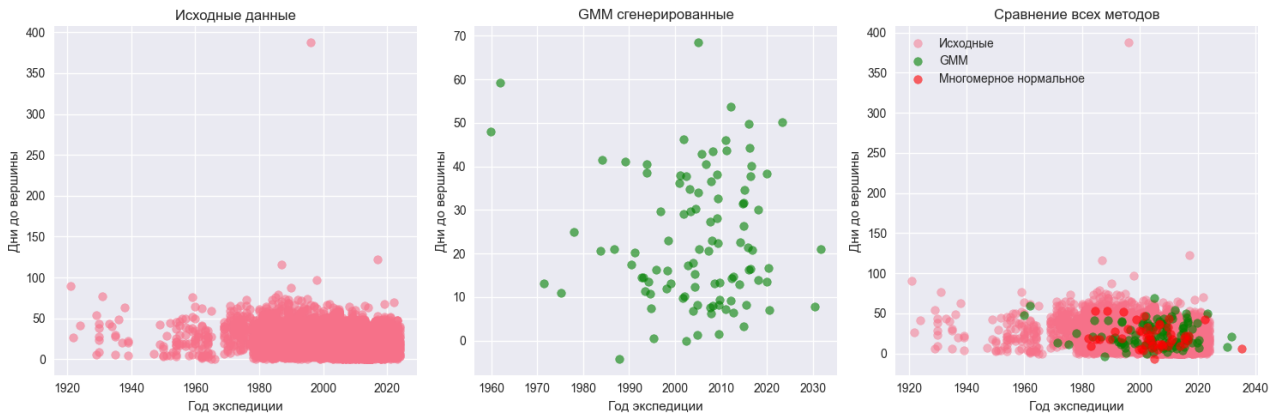

Рис. 4: Комплексное сравнение исходных данных и результатов обеих генеративных моделей

4.3 Количественная оценка качества

4.3.1 Результаты категориальной модели

Средний логарифм правдоподобия для модели многомерного нормального распределения:

- **Обучающая выборка**: -8.740
- **Тестовая выборка**: -8.795

Близость значений на обучающей и тестовой выборках указывает на отсутствие переобучения и хорошую обобщающую способность модели.

4.3.2 Анализ ковариационных структур

Анализ ковариационных матриц выявил интересные закономерности:

1. **Отрицательная корреляция** между годом и днями до вершины в осенних экспедициях (-45.41) указывает на тенденцию сокращения времени восхождения в современных экспедициях.
2. **Высокая вариативность** зимних экспедиций (821.36 для дней до вершины) отражает экстремальные условия и непредсказуемость зимних восхождений.
3. **Стабильность весенних экспедиций** с умеренными значениями дисперсии.

4.3.3 Кластерная структура GMM

Трехкомпонентная модель GMM выявила следующие латентные группы:

- **Кластер 1** (29.9%): Ранние экспедиции с длительными восхождениями
- **Кластер 2** (35.5%): Современные быстрые экспедиции
- **Кластер 3** (34.6%): Новейшие экспедиции с возвратом к длительным восхождениям

5 Заключение

Проведенный анализ данных о гималайских экспедициях с применением двух различных генеративных подходов позволил получить следующие ключевые результаты:

5.1 Основные достижения

1. **Успешная реализация конвейера обработки данных** с автоматическим обнаружением и удалением выбросов, что повысило качество моделирования.
2. **Построение эффективной категориальной модели**, учитывающей сезонную специфику экспедиций с логарифмом правдоподобия -8.795 на тестовой выборке.
3. **Выявление латентной кластерной структуры** с помощью GMM, не связанной напрямую с временными факторами.
4. **Обнаружение эволюции стратегий восхождения** через анализ корреляционных структур различных сезонов.

5.2 Практическая значимость

Разработанные модели могут быть применены для:

- Прогнозирования характеристик будущих экспедиций
- Планирования логистики горных восхождений
- Анализа рисков и безопасности экспедиций
- Исследования влияния климатических изменений на альпинизм

5.3 Направления дальнейших исследований

1. Включение дополнительных переменных (погодные условия, экономические факторы)
2. Применение более сложных нелинейных моделей (Variational Autoencoders, Flow-based models)
3. Анализ временных рядов для прогнозирования трендов
4. Исследование причинно-следственных связей в данных

A Приложение: Технические детали

A.1 Листинг исходного кода

Код работы доступен в репозитории: https://github.com/mitchivanov/_LABS_digital_culture