

МФТИ

**Отчет по практическим заданиям
Задание 1 и Задание 2**

Филак Александр

Долгопрудный
2025

1 ОТЧЁТ ПО ПРАКТИЧЕСКИМ ЗАДАНИЯМ 1 И 2

1.1. Введение

В рамках двух практических заданий были изучены методы подготовки данных для машинного обучения на основе крупномасштабных облаков точек. Задание 1 было посвящено работе с датасетом Semantic3D, представляющим собой набор outdoor-сцен, а задание 2 — датасету S3DIS, содержащему indoor-сцены (помещения). Оба задания направлены на формирование единого массива данных в формате NumPy, объединяющего пространственные координаты, цветовые характеристики и метки семантических классов для последующего использования в моделях машинного обучения.

Основная цель работ — освоить полный цикл подготовки данных: от загрузки сырых файлов до формирования нормализованных массивов, готовых для подачи в нейронные сети. Особое внимание уделено особенностям каждого датасета, проблемам предобработки и способам их решения.

1.2. Теоретическая часть

1.2.1. Датасет Semantic3D

Semantic3D — это крупномасштабный датасет для семантической сегментации облаков точек, содержащий более 4 миллиардов точек, размеченных по 8 классам. Данные получены с помощью наземного лидара в различных outdoor-сценах: городские площади, улицы, исторические здания. Датасет характеризуется высокой плотностью точек, наличием шума и неравномерным распределением классов.

1.2.2. Датасет S3DIS

S3DIS (Stanford 3D Indoor Spaces Dataset) — датасет для семантической сегментации внутренних помещений. Он содержит 6 крупных областей (зданий), разделённых на 271 помещение, с размеченными по 13 классам объектами (стены, пол, потолок, мебель и т.д.). Данные собраны с помощью сканера Matterport и имеют более структурированный характер по сравнению с Semantic3D.

1.2.3. Предобработка облаков точек

Ключевые этапы предобработки включают:

- Нормализацию координат для устранения смещений и масштабных различий
- Преобразование цветов из диапазона $[0, 255]$ в $[0, 1]$
- Формирование единой таблицы данных с признаками и метками
- Визуализацию распределения классов для анализа сбалансированности

1.3. Задание 1: Подготовка данных Semantic3D

1.3.1. Постановка задачи

Цель задания — загрузить данные из файла датасета Semantic3D (например, `'bildstein_station1xyzintensityrgb.label'`), сформировать единый массив NumPy, содержащий координаты (X, Y, Z), интенсивность, цвета (R, G, B) и метку класса, выполнить нормализацию данных и сохранить результат в форматах `.npy`, `.txt` и `.h5`.

1.3.2. Ход выполнения

Полный код реализации и результаты выполнения представлены в приложенном Jupyter Notebook (PDF-файл). Ниже описаны ключевые этапы:

1. **Загрузка данных:** Использована функция `np.loadtxt()` для чтения текстового файла с разделителем-пробелом.
2. **Извлечение признаков:** Из массива выделены столбцы с координатами, интенсивностью, цветами и меткой класса.
3. **Нормализация:**
 - Координаты центрированы и масштабированы
 - Цвета преобразованы из диапазона $[0, 255]$ в $[0, 1]$
 - Интенсивность нормализована
4. **Формирование таблицы:** Все признаки объединены в единый массив размером $N \times 8$, где N — количество точек.
5. **Сохранение:** Данные сохранены в файлы `'semantic3d_dataset.npy'`, `'semantic3d_dataset.txt'` и `'semantic3d_dataset.h5'`.
6. **Визуализация:** Построена гистограмма распределения меток классов.

1.3.3. Ответы на контрольные вопросы

1. **Semantic3D** — это датасет для семантической сегментации outdoor-сцен, содержащий облака точек, полученные с помощью наземного лидара. Он включает 8 классов: здания, дороги, растительность и др.
2. **Отличие от 2D-датасетов:** Semantic3D представляет трёхмерные данные с пространственной информацией, в отличие от ImageNet, который содержит 2D-изображения. От S3DIS отличается outdoor-сценами vs indoor.
3. **Состав датасета:** Разделён на train (15 сцен) и test (15 сцен). Train используется для обучения, test — для оценки.

4. **Технология получения:** Данные получены с помощью лидара, что приводит к высокой плотности точек, но также к шуму и неравномерному покрытию.
5. **8 классов:** building, road, vegetation, trunk, pole, car, fencing, other. Например, building — здания, road — дороги.
6. **Проблемы разметки:** Шум, артефакты, неоднозначность границ между классами.
7. **Технические проблемы:** Большой объём данных требует значительных вычислительных ресурсов для обработки и хранения.
8. **Неравномерная плотность:** В лидарных данных плотность точек уменьшается с расстоянием, что может привести к плохой сегментации удалённых объектов.
9. **Несбалансированность классов:** Некоторые классы (например, vegetation) представлены значительно больше, чем другие (pole), что может смещать модель в сторону частых классов.
10. **Этапы предобработки:** Нормализация координат, даунсэмплинг, фильтрация шума, балансировка классов.

1.4. Задание 2: Подготовка данных S3DIS

1.4.1. Постановка задачи

Цель задания — загрузить данные из датасета S3DIS, сформировать массив для каждого помещения всех областей, содержащий координаты, цвета и метки классов, выполнить нормализацию и сохранить результат.

1.4.2. Ход выполнения

Реализация представлена в Jupyter Notebook (PDF-файл). Основные этапы:

1. **Загрузка данных:** Рекурсивный обход каталогов датасета и чтение .txt файлов.
2. **Обработка каждого помещения:** Для каждого файла выделены координаты, цвета и метки.
3. **Нормализация:** Координаты нормализованы относительно границ помещения, цвета приведены к диапазону $[0, 1]$.
4. **Объединение данных:** Сформирован единый массив для всех помещений выбранной области.

5. **Сохранение:** Данные сохранены в форматах .pru и .h5 для последующего использования.
6. **Визуализация:** Построены гистограммы распределения классов по помещениям.

1.4.3. Ответы на контрольные вопросы

1. **S3DIS** — датасет для семантической сегментации внутренних помещений, содержащий 271 комнату с размеченными объектами.
2. **Типы данных:** Облака точек с координатами, цветами и метками классов.
3. **6 областей:** Area 1–6, каждая содержит несколько зданий.
4. **Формат файла:** Текстовый файл с колонками: X Y Z R G B label.
5. **Задачи ML:** Семантическая сегментация, классификация объектов, кластеризация.
6. **13 классов:** ceiling, floor, wall, beam, column, window, door, table, chair, sofa, bookcase, board, clutter.
7. **Признаки для обучения:** Координаты, цвета, нормали, локальные дескрипторы.
8. **Нормализация координат:** Необходима для устранения смещений и приведения данных к единому масштабу, что улучшает сходимость моделей.
9. **Сохранение в NumPy:** Форматы .pru/.prz обеспечивают быструю загрузку и эффективное хранение данных.
10. **Разделение признаков и меток:** Последний столбец массива — метка класса, остальные — признаки.
11. **Отличие функций сохранения:** `pr.save()` сохраняет один массив, `pr.savez()` — несколько, `pr.savez_compressed()` — сжатые данные.
12. **Потенциальные проблемы:** Большой объём данных, несбалансированность классов, пропуски в данных.

1.5. Выводы

1. **Задание 1:** Освоены методы работы с крупномасштабными outdoor-облаками точек. Особое внимание уделено обработке лидарных данных, характеризующихся неравномерной плотностью и шумом.
2. **Задание 2:** Освоена подготовка данных для indoor-сцен, где важную роль играет структурная информация о помещениях и чёткое разделение на классы объектов интерьера.

1.6. Заключение

В ходе выполнения заданий были успешно подготовлены данные из датасетов Semantic3D и S3DIS для использования в задачах машинного обучения. Реализован полный конвейер обработки: от загрузки сырых данных до формирования нормализованных массивов NumPy. Особое внимание было уделено ответам на контрольные вопросы, которые позволили глубже понять особенности каждого датасета и связанные с ними проблемы. Полученные массивы данных готовы для использования в обучении моделей семантической сегментации облаков точек.