

ML Roadmap

16 февраля 2025 г.

Розовым цветом буду обозначать дополнительные штуки, которые не обязательно изучать, но потенциально могут встретиться где-нибудь. **Приведенный ниже план – не монолит, всегда можно либо пропустить, либо добавить материал.**

Разумеется, если ты и так знаешь что-то из списков ниже, то мы будем это скипать и разбирать только нужное

1 Python

Пройдемся по основам питона, чтобы более-менее уверенно владеть синтаксисом, знать сильные и слабые стороны языка, а также уметь быстро накидать MVP-решение той или иной задачки, создать скрипт, который будет удобно запускаться из терминала, и так далее.

Примерный план:

- Синтаксис, основные операторы
- Типы данных `list`, `tuple`, `dict`, `set`. Мутабельные и немутабельные типы данных. Отличие `list`, `tuple`
- Условные операторы `if`, `elif`, `else`
- Отличие `is`, `==`
- Итераторы и генераторы, их отличия
- Оператор `in`
- Циклы `while`, `for`, вложенные циклы
- Локальные, глобальные переменные
- Функции, `lambda`-функции
- ООП. Дандер-методы
- Исключения, `try`, `except`, etc.
- Декораторы
- Проверки: `assert`, `isinstance`
- Взаимодействие с атрибутами: `hasattr`, `getattr`, `setattr`, `delattr`
- Оформление проекта: комментарии, `docstrings`, рефакторинг, модули, PEP 8
- Либры `numpy`, `pandas`, `argparse`, `json`, etc. Написание кастомных пакетов

2 Math

По необходимости подтянем и проясним моменты из

- Линейной алгебры
- Матстата¹
- Теорвера
- По желанию – численные методы, теория оптимизации и тд., потому что это встречается в МЛ на первых порах

3 ML

Основы по классическому и глубокому обучению. С дальнейшей специализацией в интересующую область (CV, NLP, etc.)

3.1 Classic ML

- Задача обучения с учителем
- Типы задач в МЛ и их постановка
- Функции потерь, метрики бинарной классификации: Accuracy, Precision, Recall, F_1 , ROC AUC, PR-curve, confusion matrix
- Метод Наименьших Квадратов (МНК)
- Линейная классификация, регрессия
- Кластеризация: KNN
- SVM – метод опорных векторов. Разные ядра SVM
- Кросс-валидация: обучение на разных фолдах
- Сокращение размерности: Lasso (L_1). Геометрический смысл отбора признаков. PCA
- Решающие деревья. Критерии ветвления
- Случайный лес. Метод случайных подпространств
- Линейные ансамбли, стохастическое ансамблирование, bagging, boosting.
- Градиентный бустинг. CatBoost, LightGBM.
- Stacking, blending
- Байесовская теория классификации. Наивный байес. Восстановление плотности распределения: 1. Парзеневские окна, 2. Метод максимума правдоподобия
- Геометрический смысл предположения нормальности признаков
- Кластеризация иерархическая и спектральная
- Логистическая регрессия

3.2 Deep Learning (DL)

- Переобучение и недообучение и методы борьбы с ними: регуляризация. Dropout на трейне и инференсе
- Построение простейшей нейросети. Веса. Слои
- Численная минимизация лосс-функции: метод стохастического градиента (SGD) и др.
- Адаптивные градиенты
- BatchNorm

¹Из матстата наиболее вероятно могут встретиться понятия вроде таких: статистические гипотезы, ошибки 1, 2 рода, доверительные интервалы, статистические тесты, уровень значимости α . Разберем по необходимости

- Сингулярное разложение матриц (SVD)²
- Мультиколлинеарность и переобучение. Гребневая (Ridge (L_2)) регрессия
- Гиперпараметры, какие они бывают, их подбор
- Глубокое обучение. Сверточные нейросети: `conv`, `pool`
- Затухающие градиенты. Проблемы глубоких сетей. Остаточные связи (ResNet), skip-connections.
- Аугментирование данных.
- Рекуррентные нейросетки (LSTM, GRU). Карты Кохонена
- Автокодировщики (AE): Линейный AE, sparse AE, variational AE, etc.
- GAN
- Предобучение, дообучение. Fine-tuning
- Дистилляция моделей. Квантование. Прунинг.
- Векторные представления данных (картинок, текста, etc.) – *эмбединги*. CBOW. word2vec.
- Многомерное шкалирование данных: SNE, t-SNE.
- Трансформеры. Модели внимания, мотивация их появления. Multihead self-attention. Архитектура трансформера. Позиционное кодирование
- Отличия BatchNorm, LayerNorm, GroupNorm, InstanceNorm
- Тематическое моделирование, LDA. TF-IDF, BM25.

4 Soft skills + interviews

- Параллельно с освоением теории и практики – мок-собеседы.
- Натаскаю по вопросам с реальных собеседов³
- Напишем крутое резюме для обхода hr-фильтров

²Но это можно к матеше отнести. Просто это важная штука в МЛ, много где встречается. Здесь заодно можно будет поговорить про собственные числа и векторы матриц

³Разберем классические и каверзные вопросы, чтобы по максимуму повысить шансы устройства по вакансии