ML Roadmap

2 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**

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
- Операторы `if, elif, else`
- Циклы `while`, `for`, вложенные циклы
- Локальные, глобальные переменные
- Функции, `lambda`-функции
- ООП
- Исключения, `try, except`, etc.
- Комментарии, docstrings, оформление проекта
- *Некоторые либы*: numpy, pandas, argparse, json, etc. Написание кастомных пакетов

2 Math

По необходимости подтянем и проясним моменты из

- Матстата[1]
- Теорвера
- Линейной алгебры
- По желанию – численные методы, теория оптимизации и тд., потому что это встречается в МЛ на первых порах

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

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, MoE (Mixture of Experts)
- Байесовская теория классификации. Наивный байес. Восстановление плотности распределения: 1. парзеновские окна, 2. метод максимума правдоподобия
- Геометрический смысл предположения нормальности признаков
- Кластеризация иерархическая и спектральная
- Логистическая регрессия
- Обобщенные линейные модели (GLM)

3.2 Deep Learning (DL)

- Переобучение и недообучение и методы борьбы с ними: регуляризация
- Построение простейшей нейросети. Веса. Слои
- Численная минимизация лосс-функции: метод стохастического градиента (SGD) и др.
- Адаптивные градиенты
- Batch normalization, dropout
- 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`

[2]Но 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

- Затухающие градиенты. Проблемы глубоких сеток. 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.) – *эмбеддинги*. CBOW. `word2vec`.
- Многомерное шкалирование данных: SNE, t-SNE.
- Трансформеры. Модели внимания, мотивация их появления. Multihead self-attention. Архитектура трансформера. Позиционное кодирование
- Тематическое моделирование, LDA. TF-IDF, BM25.

4 Soft skills + interviews

- Параллельно с освоением теории и практики – мок-собесы.
- Натаскаю по вопросам с реальных собесов[3]
- Напишем крутое резюме для обхода hr-фильтров

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