Коллоквиум по ML

Tinkoff Generation

17-18 мая 2019

1 Теоретические вопросы

Вам достаётся два случайных из них, вам нужно подготовить по ним ответ, желательно с какими-то записями. Оценивается в 2 балла (по баллу за каждый билет).

Тема №1 Задачи машинного обучения

Обучение с учителем. Обучение без учителя. Регрессия. Классификация. Кластеризация

Тема №2 Приближение функции многочленами

Задача, модель, функция потерь. Проблема переобучения на примере интерполяции. Зачем нужно разбиение выборки на тренировочную и тестовую часть. Кросс-валидация.

Тема №3 Линейная регрессия

Задача и модель линейной регрессии. Регуляризация. Аналитическое решение линейной регрессии.

Тема №4 Логистическая регрессия

Задача и модель бинарной логистической регрессии.

Тема №5 Метрики классификации

Метрики классификации: accuracy, precision, recall, F1, ROC-AUC. Небинарная классификация.

Тема №6 Кластеризация

Задача кластеризации. K-means. DBSCAN. Агломеративная кластеризация.

Тема №7 Уменьшение размерности

Задача уменьшения размерности. SVD. PCA. T-SNE.

Тема №8 Решающие деревья

Алгоритм построения решающего дерева. Кросс-энтропия.

Тема №9 Ансамблирование

Ансамбль алгоритмов. Бэггинг. Случайный лес. Градиентный бустинг.

Тема №10 Обработка текстов

Токенизация, Лемматизация, CountVectorizer, OneHotVectorizer, Стемминг.

Тема №11 Релевантность документа

Задача проверки на релевантность документа на запрос. TF-IDF.

Тема №12 Рекомендательные системы

Постановка задачи. Бейзлайны. Фильтрация на основе содержания (content-based approach). Коллаборативная фильтрация.

2 Задачи

Выдается одна задача, нужно в общих чертах расписать, как бы вы решали практическую задачу, а именно

- как поставить четко задачу
- как должен выглядеть датасет и как его собирать
- какой алгоритм использовать
- какие метрики использовать
- как понять, хорошо ли работает ваше решение на практике?

Оценивается в 2 балла.

Задача №1 Кинопоиск

Вы разрабатываете сайт с базой данных фильмов и хотите ловить ботов, которые накручивают лайки.

Задача №2 Википедия

Вы делаете текстовую энциклопедию и хотите реализовать свой поиск.

Задача №3 Яндекс.Маркет

Вы делаете интернет-магазин и хотите сделать рекомендации для пользователей.

Задача №4 Акинатор

Вы делаете игру-угадайку, которая задает пользователю бинарные вопросы вида “Этот персонаж - животное?” и пытается угадать, кого загадал пользователь.

Задача №5 Тинькофф

Вы разрабатываете алгоритм для банка, который должен определять, выдать ли пользователю кредит.

Задача №6 Яндекс.Директ

Вы разрабатываете рекламную сеть и хотите предсказывать пол пользователя по его поведению в интернете, чтобы потом использовать это для лучшего таргетирования.

Задача №7 Яндекс.Картинки

Вы разрабатываете поиск по картинкам и хотите, например, по запросу “замок” показывать два блока картинок для каждого смысла - фото зАмков и замкОв, то есть выделить для запроса все разные сущности и показывать картинки с ними отдельно.

Задача №8 Google Maps

Вы разрабатываете онлайн-карты и хотите по запросу, по которому надо показать очень-очень много меток, объединять несколько соседних меток в одну большую.

Задача №9 GeekTimes

Вы делаете статью для журнала и хотите нарисовать красивую двумерную картинку, где разные точки - это языки, и похожие языки расположены рядом.

Задача №10 Gmail

Вы делаете почту и хотите автоматически определять спам и кидать в корзину со спамом.

3 Дополнительный вопрос

Также преподаватель спрашивает вас дополнительный вопрос на его выбор из любой части программы. Ответ на него оценивается в 1 балл.