Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э.Баумана (национальный исследовательский университет)»

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № **121**

по дисциплине **«Технологии машинного обучения»**

Группа **РТ5-61Б** Фамилия Имя Отчество **Забурунов Леонид Вячеславович**

Дата **02.07.2021** Время начала сдачи **14:00** Время окончания сдачи **18:00**

**Вопрос 1. Почему метод ближайших соседей фактически не включает этап обучения?**

Метод ближайших соседей в качестве обучения предполагает измерение расстояний от рабочей точки до всех остальных и, соответственно, упорядочивание всех расстояний по возрастанию, чтобы получить тех самых ближайших соседей. Классический «пайплайн» для модели МО представляет собой фазу обучения, в которой модель заранее обучается на имеющихся данных с целью дать прогноз при поступлении новых. Поскольку в методе ближайших соседей фаза «заготовки» отсутствует (невозможно вычислить расстояние от/до точки, координаты которой мы ещё не знаем), то понятие этапа обучения в принципе ставится под сомнение.

**Вопрос 2. Кластеризация. Метод Affinity Propagation.**

Кластеризация – метод машинного обучения без учителя, при котором множество данных разбивается на подгруппы (кластеры). Модель МО отображает вектор (входные данные) на метку (число, номер кластера).

Методы кластеризации оказываются полезными для предварительной работы с данными: в случае избыточной сложности датасета или его объёма с помощью выделения кластеров можно понять и упростить его структуру. Точнее, кластеризация позволяет разбить выборку на подвыборки, чтобы в дальнейшей анализировать уже эти подвыборки; это упрощает дальнейшую обработку (можно анализировать каждую подвыборку, можно сокращать число образцов для каждой подвыборки и так далее).

Алгоритмы кластеризации могут получать число кластеров на вход как гиперпараметр или же определять его самостоятельно. Примером последней группы является алгоритм *Affinity Propagation* («распространение схожести»). Данный алгоритм также называют «кластеризацией с помощью обмена сообщениями». Получая на вход матрицу *X*, модель начинает итеративный процесс выделения кластеров и определения центров этих кластеров. Во время работы алгоритма обучения формируются две дополнительные матрицы:

1. *R* (*responsibility*, матрица ответственности) – здесь элемент показывает, насколько точка i готова быть в кластере с центром в точке k;
2. *A* (*availability*, матрица доступности) – здесь элемент является ответом на то, насколько точка k готова быть центром кластера, содержащего точку i;

Как и в прочих итеративных процессах, здесь устанавливаются критерии выхода в виде предела числа итераций и *threshold*-значения, являющегося критерием существенности изменения.

**Вопрос 3. Задача №5.**

Набор, на основе которого построена предложенная корреляционная матрица, имеет хорошие перспективы для предсказания с помощью моделей машинного обучения. С одной стороны, видим отсутствие линейных зависимостей нецелевых признаков между собой, с другой – наличие сильной зависимости одного из признаков с целевым. Таким образом, признак X2 должен оказаться наибольшее влияние на итоговый вид модели, и при этом одного этого признака может быть достаточно для околоединичных показателей качества.

Билет рассмотрен и утвержден на заседании кафедры ИУ-5 № 8 от «11» марта 2021 года