ДИСЦИПЛИНА	**Проектирование интеллектуальных систем (часть 1/2)**
ИНСТИТУТ	**информационных технологий**
КАФЕДРА	**вычислительной техники**
ВИД УЧЕБНОГО МАТЕРИАЛА	**Материалы для практических/семинарских занятий**
ПРЕПОДАВАТЕЛЬ	**Холмогоров Владислав Владиславович**
СЕМЕСТР	**6, 2023-2024**

Практическая работа № 3

«Классификации в анализе данных»

по дисциплине «Проектирование интеллектуальных систем (часть 1/2)»

Цели: приобрести навыки классификации как инструмента категориального и предикативного анализа данных с контролируемым обучением.

Задачи:

1) Провести классификацию данных исходного набора, выполнив следующие шаги:

– определить предметную область решаемой задачи, ею могут выступать анализ потребительской корзины, предсказание погоды, событий и заболеваний, определение качества товара (или любого другого объекта) на основе атрибутов, автоматическое определение группы или тега объекта, разделение почты по типу и т.д.;

– выбрать или сгенерировать набор данных для задачи и провести над ним предварительную обработку данных (хотя бы минимальную) подобно тому, как она проведена в практической работе №2, однако также необходимо подготовить данные для обучения и тестирования модели (см. Примечание 1);

– выбрать модель классификации, провести её обучение и тестирование, определить качество модели с помощью метрик потерь;

– провести бинарную и/или многоклассовую классификацию данных подготовленного набора.

2) В качестве дополнительного задания реализовать хотя бы один из следующих пунктов (один на оценку 4, два и более – на оценку 5):

– провести множественную и/или несбалансированную классификацию данных, либо любую другую из существующих категорий (см. Примечание 2);

– реализовать ансамбль моделей классификации (см. Примечание 3);

– выполнить задачу предсказательной аналитики/регрессии с помощью алгоритмов классификации (см. Примечание 4).

ПРИМЕЧАНИЕ:

1) Так как классификация является задачей обучения с учителем (из самого определения классификации), то для неё необходим размеченный набор данных с указанием меток классов у каждой записи. Сам процесс предобработки данных совпадает с таковым при кластеризации (замена или удаление пропусков, удаление дублей и т.п., см. Практическую работу №2), некоторые алгоритмы классификации требуют данные в конкретном виде (как правило, в числовом), поэтому необходимо выбирать или преобразовывать уже выбранный набор данных к требуемому виду.

2) Алгоритмы классификации сами имеют достаточно широкую классификацию:

- по способу разделения: линейные и нелинейные;
- по соотношению времени, затрачиваемого на обучение и саму работу: ленивые и энергичные (Lazy Learners and Eager Learners);
- по количеству и качеству принадлежности элемента к классам: бинарная, мультиклассовая, несбалансированная и множественная.

3) Ансамбли моделей искусственного интеллекта позволяют различными способами объединить (как правило, нечётное количество для предотвращения возникновения взаимосвязей) несколько моделей для решения какой-либо (обычно, сложной) задачи, которую один или каждый из методов по отдельности решают недостаточно удовлетворительно, существуют следующие виды ансамблей:

- последовательные ансмабли: однородные модели обучаются последовательно на данных друг друга, что работает как сеть фильтров – бустинг;
- параллельные ансмабли: используется несколько неоднородных (как правило, слабых) моделей, которые обучаются и работают одновременно – стэкинг;

– однородные ансамбли: объединение нескольких однородных моделей, обученных на разных наборах данных и усреднение их результатов, как правило, применяется бэггинг, но может быть и реализация в виде бустинга;
– гетерогенный ансамбль: использует одни и те же данные, но разные признаки из них, так как состоит как из разных групп классификаторов, так и в целом из разных моделей машинного обучения, результаты усредняются.

4) Методы классификации позволяют предсказывать классы, к которым будут относится те или иные элементы на основе обученной на признаках и метках модели, что позволяет реализовать предсказательную аналитику возможных групп ситуаций, например, купит ли конкретный пользователь товар снова, вырастет или спадёт цена на определённую продукцию через несколько месяцев по определённым факторам, будет ли интересен новый продукт пользователям по определённым показателям (цвет, размер, состав и т.п.), что отличается от регрессии, которая предсказывает конкретные значения: какова будет цена на определённый продукт через полгода, сколько за это время потратит наш клиент, сколько будет стоить новый продукт по его параметрам.

ОСОБЫЙ БОНУС (доступен только в том случае, если выполнены пункты 2-го задания):

Выполнить классификацию и кластеризацию на одном и том же наборе данных с преобразованием его вида для каждой из задач (для этого нужен набор данных, содержащий и метки для классификации и признаки для кластеризации, если не удаётся найти такой набор данных, его необходимо сгенерировать или получить с помощью методов интеграции данных), сравнить результаты полученных классов и кластеров.