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 № 4

«Регрессия в анализе данных»

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

Цели: приобрести навыки применения моделей регрессии для решения прикладных задач анализа и сбора данных.

Задачи:

1) Провести регрессию атрибутов исходного набора данных, выполнив следующие шаги:

– определить предметную область решаемой задачи, ею могут выступать прогнозирование цен, продаж и тенденций на рынке, оценка времени протекания определённых процессов (например, выздоровления пациентов, факторов риска, веса, роста и т.д.), системы поддержки принятия решений (акции, риски производства и систем обеспечения, кадровая аналитика);

– выбрать или сгенерировать набор данных для задачи, предварительно проведя предобработку данных (см. Примечание 1);

– провести визуализацию предобработанного набора данных и выбрать для него модель регрессии (см. Примечание 2);

– выбрать метрику ошибку и обучить модель регрессии;

– выполнить задачу предметной области, интерпретировать и визуализировать результаты обучения модели.

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

– решить задачу предметной области с помощью нескольких разных моделей регрессии, визуализировать и сравнить результаты по времени обучения и выполнения задачи, количеству используемой памяти и качеству;

– обучить модель регрессии с помощью градиентного бустинга;

– выполнить понижение размерность исходного набора данных с помощью регрессионной модели (см. Примечание 3).

ПРИМЕЧАНИЕ:

1) Модели регрессии в меньшей степени (как правило, кроме простейшей линейной), чем остальные алгоритмы машинного обучения зависит от аномальных значений, но сильно зависит от мультиколлинеарности, при которой зависимые переменные (входные параметры в модель регрессии) находятся в строгой или нестрогой линейно зависимости друг от друга (например, $x_3 = 2x_1 + x_2$ или $x_3 \approx 2x_1 + x_2$), поэтому, если не прибегать в гребневой регрессии, необходимо хотя бы выполнять корреляционный анализ и/или понижение размерности; пропуски данных, как и слишком аномальные значения, имеет смысл просто удалять.

2) Визуализация данных для регрессии позволяет определить характер распределения признаков в пространстве, это распределение может иметь линейный, нелинейный, монотонный и немонотонный характер, в зависимости от этого характера нужно выбрать модели регрессии.

Классические методы регрессии:

- линейная регрессия: предполагает, что существует линейная связь между независимыми и зависимыми переменными;
- полиномиальная регрессия: используется для моделирования нелинейных отношений между зависимой переменной и независимыми переменными, добавляет полиномиальные члены в модель линейной регрессии для выявления более сложных взаимосвязей;
- регрессия опорных векторов: тип алгоритма регрессии, основанный на алгоритме машины опорных векторов (SVM), работает путем поиска гиперплоскости, которая минимизирует сумму квадратов остатков между прогнозируемыми и фактическими значениями;
- регрессия дерева решений: тип алгоритма регрессии, который строит дерево решений для прогнозирования целевого значения;

– регрессия случайного леса: ансамблевый метод, который объединяет несколько деревьев решений для прогнозирования целевого значения.

Методы регуляризационной линейной регрессии:

– ридж-регрессия: тип линейной регрессии, который используется для предотвращения переобучения путём выполнения регуляризации L2, которая позволяет убирать корреляцию между входными параметрами (мультиколлинеарность).
– лассо-регрессия: тип линейной регрессии, который используется для предотвращения переобучения путем добавления штрафного члена к функции потерь, который заставляет модель использовать некоторые веса и устанавливать другие равными нулю.

3) Ридж-регрессия (она же гребневая) способна убирать коррелирующие значения (борется с мультиколлинеарностью), что по итогу создаёт пространство меньшего количества признаков, что, по сути, выполняет понижение размерности набора данных, убирая коррелирующие значения.

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

Выполнить задачу классификации данных с помощью логистической регрессии и сравнить результаты с результатами практической работы №3.