ФГАОУ ВО «СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра «Информатика»

Методы анализа данных

Практическая работа № 3
Классификация

Красноярск, 2024

Цель: знакомство с теоретическими основами задачи классификации объектов, формирование навыков решения задачи бинарной классификации с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Предварительная обработку исходных данных
2. Обучение базовых моделей классификации
3. Подбор оптимальных параметров моделей классификации
4. Оценка качества построенных моделей на тестовой выборке

Ссылки на соревнования

Выбор варианта – остаток деления на 3 + 1.

5. https://www.kaggle.com/c/mso-titanic/overview
6. https://www.kaggle.com/c/mso-insurance/overview
7. https://www.kaggle.com/c/mso-churn/overview

Используемые алгоритмы классификации:

1. Логистическая регрессия
2. Метод ближайших соседей
3. Наивный байесовский классификатор
4. Дискриминантный анализ (линейный дискриминантный анализ, квадратичный дискриминантный анализ)
5. Машина опорных векторов.

Для данных методов предусмотреть настройку гиперпараметров. Настройку производить с помощью метода перекрестной проверки.

Общая последовательность действий

1. Ознакомиться с вариантом своего соревнования.
2. Загрузить данные для обучения и для теста.
3. Выполнить предварительную обработку исходных данных (в случае необходимости)
4. Построить модели классификаторов с параметрами, подобранными на перекрестной проверке (cross validation).
5. Провести отбор информативных признаков с помощью разных подходов (встроенные методы, методы фильтрации, методы-обертки).
6. Применить технику сэмплирования к исходным несбалансированным данным (over-sampling, under-sampling, ансамблевые методы).

7. Предсказать целевую переменную для тестовой выборки.
8. Набрать необходимый score для своего варианта.

Точность моделей должна быть больше установленного порогового значения:

- Соревнование «Titanic» - ROC-AUC 0.84
- Соревнование «Customer churn prediction» - ROC-AUC 0.86
- Соревнование «Health Insurance Cross Sell Prediction» - ROC-AUC 0.86

Требования к выполнению практической работы:

1. Написание программного кода и формирование результатов согласно заданию и установленному варианту.

2. Составление отчета, содержащего описание решаемых задач методов решения и полученных результатов.

3. Воспроизводимость полученного результата.

Программный код и отчет должны быть выполнены в среде Jupyter notebook.