Обработка данных для машинного обучения

Цель работы

Познакомиться с основными приемами дескриптивного анализа данных для подготовки к машинному обучению

Задания для выполнения

На сайте Каггл выбрать набор данных каждому студенту индивидуально. Лучше всего выбирать датасеты в предметной области исходя из личных предпочтений. Не стоит выбирать слишком большие наборы данных (лучше не более 100 Мб).

В начале анализа следует привести общую характеристику набора данных: количество наблюдений, признаков, предметная область, содержание признаков, характер целевой переменной, задача обучения (регрессия, классификация, количество классов).

Провести в Юпитер средствами пандас, СКлерн, Сибон описательный анализ выбранного набора включающего (но не ограниченного) следующие шаги:

исследование шкалы измерения каждого существенного признака;

построение эмпирического распределения каждого существенного признака;

анализ количества и распределения отсутствующих значений;

построение совместного распределения каждого признака и целевой переменной;

построение корреляционной матрицы.

По каждому пункту анализа необходимо сделать содержательный вывод.

Провести подготовительную обработку данных, включающую следующие шаги:

6. удаление или заполнение отсутствующих значений;

7. приведение всех признаков к бинарной либо числовой шкале;

8. удаление несущественных либо избыточных признаков;

9. другие необходимые действия в зависимости от набора данных и задачи (группировка, удаление аномалий, и др.).

Количественные характеристики датасета необходимо также привести после его обработки.

Контрольные вопросы

Какие существуют методы устранения отсутствующих значений в наборе данных? Исходя из чего нужно применять различные методы?

Что показывает коэффициент корреляции? В чем его ограничения?

Какие признаки можно считать несущественными?

Какие методы преобразования категориальных признаков в численные существуют?

Какие статистические распределения чаще всего встречаются в реальных задачах?

Что такое совместное и условное распределение?

Дополнительные задания

Использовать несколько моделей машинного обучения для решения поставленной задачи.

Использовать методику замеры времени обучения каждой использованной модели.

Сформировать сводную таблицу результатов обучения, включающую следующие данные:

время обучения модели;

избирательность;

Точность;

Полнота;

Ф1-параметр.