Цель лабораторной работы – получение навыков работы с методами классификации.

1. **Изучить набор данных. Создать описание набора данных на русском языке. Описать признаки, используемые в наборе данных (включить полученные описания в отчёт).**
2. **Удалите дубликаты строк в наборе данных;** **приведите размер набора данных до и после данной операции;**
3. **Оцените сбалансированность данных по классам (постройте гистограмму). Используйте полученную информацию при выборе метрики оценки качества классификации (PR или ROC кривая) (пример гистограммы) (про RP и ROC посмотреть, как это должно выглядеть)**
4. **Выполните масштабирование количественных признаков;** **Постройте диаграммы BoxPlot для признаков до и после масштабирования. Выберите способ масштабирования (например, нормализацию или стандартизацию); (примерчик сюда)**
5. **Выполните замену категориальных признаков; выберите и обоснуйте способ замены; (label encoder и one hot encoder)**
6. **Оцените корреляцию между признаков и удалите те признаки, которые коррелируют с наибольшим числом других (удалять признаки нужно только для линейных методов классификации); (матрицу корреляций)**
7. **Заполните пропущенные значения в данных; (imputers)**
8. Решите поставленную задачу классификации в соответствии с заданием. При подборе параметров классификатора используйте метод GridSearchCV и перекрёстную проверку (изучите возможные для изменения параметры классификации). Определите схему построения многоклассового классификатора, используемую по умолчанию (опишите используемую схему кодирования, обоснуйте свой выбор). Постройте, если это возможно, многоклассовую классификацию на основе схем «один-против-всех» и «все-против-всех». Оцените точность классификации для каждой их схем. Постройте кривые PR и ROC (для каждого из классов должны быть построены отдельные кривые, а также кривые для микро и макроусреднения метрик качества). Для линейного классификатора используйте регуляризацию. (проработать этот пункт, слишком большой и много условий)
9. Сравните кривые для классификаторов, указанных в задании, сделайте выводы. (указать методы сравнения кривых)

На практике: рассмотреть перекрёстную проверку, GridSearchCV

Кодировка классификаторов:

1 – классификатор K ближайших соседей (задаётся количество ближайших объектов);

2 – классификатор K ближайших соседей (задаётся радиус для выбора ближайших объектов);

3 – линейный классификатор (персептрон); (указать код классификатора)

4 – логический классификатор (бинарное решающее дерево). DTC

Варианты заданий:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Вариант | Набор данных / Классификаторы | Вариант | Набор данных / Классификаторы | Вариант | Набор данных / Классификаторы |
| 1 | 1 / 1, 3 | 2 | 3 / 1, 3 | 3 | 5 / 1, 3 |
| 4 | 1 / 1, 4 | 5 | 3 / 1, 4 | 6 | 5 / 1, 4 |
| 7 | 1 / 2, 3 | 8 | 3 / 2, 3 | 9 | 5 / 2, 3 |
| 10 | 1 / 2, 4 | 11 | 3 / 2, 4 | 12 | 5 / 2, 4 |
| 13 | 1 / 3, 4 | 14 | 3 / 3, 4 | 15 | 5 / 3, 4 |
| 16 | 2 / 1, 3 | 17 | 4 / 1, 3 | 18 | 6 / 1, 3 |
| 19 | 2 / 1, 4 | 20 | 4 / 1, 4 | 21 | 6 / 1, 4 |
| 22 | 2 / 2, 3 | 23 | 4 / 2, 3 | 24 | 6 / 2, 3 |
| 25 | 2 / 2, 4 | 26 | 4 / 2, 4 | 27 | 6 / 2, 4 |
| 28 | 2 / 3, 4 | 29 | 4 / 3, 4 | 30 | 6 / 3, 4 |
| 31 | 7 / 1, 3 | 32 | 9 / 1, 3 | 33 | 11 / 1, 3 |
| 34 | 7 / 1, 4 | 35 | 9 / 1, 4 | 36 | 11 / 1, 4 |
| 37 | 7 / 2, 3 | 38 | 9 / 2, 3 | 39 | 11 / 2, 3 |
| 40 | 7 / 2, 4 | 41 | 9 / 2, 4 | 42 | 11 / 2, 4 |
| 43 | 7 / 3, 4 | 44 | 9 / 3, 4 | 45 | 11 / 3, 4 |
| 46 | 8 / 1, 3 | 47 | 10 / 1, 3 | 48 | 12 / 1, 3 |
| 49 | 8 / 1, 4 | 50 | 10 / 1, 4 | 51 | 12 / 1, 4 |
| 52 | 8 / 2, 3 | 53 | 10 / 2, 3 | 54 | 12 / 2, 3 |
| 55 | 8 / 2, 4 | 56 | 10 / 2, 4 | 57 | 12 / 2, 4 |
| 58 | 8 / 3, 4 | 59 | 10 / 3, 4 | 60 | 12 / 3, 4 |
| 61 | 13 / 1, 3 |  |  |  |  |
| 62 | 13 / 1, 4 |  |  |  |  |
| 63 | 13 / 2, 3 |  |  |  |  |
| 64 | 13 / 2, 4 |  |  |  |  |
| 65 | 13 / 3, 4 |  |  |  |  |