**Лабораторная работа №1**

**Бинарная классификация фактографических данных**

*Цель работы:*

Получить практические навыки решения задачи бинарной классификации данных в среде Jupiter Notebook. Научиться загружать данные, обучать классификаторы и проводить классификацию. Научиться оценивать точность полученных моделей.

*Задание:*

1. В среде Jupiter Notebook создать новый ноутбук (Notebook)
2. Импортировать необходимые для работы библиотеки и модули
3. Загрузить данные в соответствие с вариантом
4. Вывести первые 15 элементов выборки (координаты точек и метки класса)
5. Отобразить на графике сгенерированную выборку. Объекты разных классов должны иметь разные цвета.
6. Разбить данные на обучающую (train) и тестовую (test) выборки в пропорции 75% - 25% соответственно.
7. Отобразить на графике обучающую и тестовую выборки. Объекты разных классов должны иметь разные цвета.
8. Реализовать модели классификаторов, обучить их на обучающем множестве. Применить модели на тестовой выборке, вывести результаты классификации:
* Истинные и предсказанные метки классов
* Матрицу ошибок (confusion matrix)
* Значения полноты, точности, f1-меры и аккуратности
* Значение площади под кривой ошибок (AUC ROC)
* Отобразить на графике область принятия решений по каждому классу

В качестве методов классификации использовать:

1. Метод к-ближайших соседей (n\_neighbors = {1, 3, 5, 9})
2. Наивный байесовский метод
3. Случайный лес (n\_estimators = {5, 10, 15, 20, 50})
4. По каждому пункту работы занести в отчет программный код и результат вывода.
5. По результатам п.8 занести в отчет таблицу с результатами классификации всеми методами и выводы о наиболее подходящем методе классификации ваших данных.
6. Изучить, как изменится качество классификации, если на тестовую часть выделить 10% выборки, 35% выборки. Для этого повторить п.п. 6 – 10.

**Варианты заданий**

Объем выборки для всех вариантов – n\_samples = 1000 штук.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Вариант** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Вид классов | blobs | blobs | blobs | blobs | moons | moons | moons | moons |
| Random\_state | 34 | 28 | 41 | 23 | 41 | 23 | 77 | 15 |
| cluster\_std | 1.5 | 4.5 | 3 | 5 | - | - | - | - |
| noise | - | - | - | - | 0.25 | 0.3 | 0.25 | 0.2 |
| Centers | 2 | 2 | 2 | 2 | - | - | - | - |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Вариант** | **9** | **10** | **11** | **12** |
| Вид классов | classification | classification | classification | classification |
| Random\_state | 78 | 58 | 15 | 23 |
| class\_sep | 0.45 | 0.7 | 0.6 | 0.35 |
| Для всех вариантов, использующих для генерации make\_classification, дополнительные параметры: n\_features=2, n\_redundant=0, n\_informative=1, n\_clusters\_per\_class=1, |

**Контрольные вопросы**

1. Постановка задачи классификации данных. Что такое бинарная классификация?
2. Общий алгоритм решения задачи классификации данных.
3. Чем отличаются обучающая и тестовая выборки? Какие существуют способы формирования обучающей и тестовой выборок?
4. Как рассчитываются значения полноты и точности классификации?
5. Как рассчитывается значение площади под кривой ошибок?
6. Что показывает и как рассчитывается матрица ошибок?
7. Алгоритм и особенности метода к-ближайших соседей.
8. Алгоритм и особенности метода случайного леса.