Лабораторная работа 3
«Сравнение методов классификации»

Цель работы

Освоить методы построения и сравнения различных типов классификаторов: линейных, нелинейных и байесовских подходов. Изучить особенности применения линейного и квадратичного дискриминантного анализа.

Задачи

1. Изучить основы методов классификации и их применение для решения задач распознавания образов
2. Освоить построение линейных классификаторов
3. Изучить методы нелинейной классификации
4. Освоить байесовские методы классификации
5. Изучить применение линейного и квадратичного дискриминантного анализа
6. Сравнить качество различных моделей классификации

Исходные данные (выбрать один из датасетов)

- **Вариант 1:** Breast Cancer Wisconsin Dataset
- **Вариант 2:** Iris Plants Dataset
- **Вариант 3:** Wine Recognition Dataset
- **Вариант 4:** Pima Indians Diabetes Dataset

Описание выбранного датасета

- Название датасета: [указать выбранный вариант]
- Описание: [краткое описание задачи классификации]
- Количество наблюдений: [указать количество]
- Количество классов: [указать количество]
- Количество признаков: [указать количество]
- Тип признаков: [указать типы - непрерывные, категориальные]
- Проблема: [сбалансированная/несбалансированная классификация]

Методика выполнения

Часть 1: Подготовка данных и предварительный анализ

1. Загрузка и предобработка данных
2. Анализ распределения классов
3. Исследование корреляций между признаками
4. Разделение на обучающую и тестовую выборки (70/30 или 80/20)
5. Масштабирование признаков для методов, требующих нормализации

Часть 2: Линейные методы классификации

1. Построение модели логистической регрессии
2. Построение линейного SVM (Support Vector Machine)
3. Оценка качества моделей с помощью accuracy, precision, recall, F1-score
4. Построение матриц ошибок (confusion matrix)

Часть 3: Нелинейные методы классификации

1. Построение модели SVM с радиально-базисной функцией (RBF kernel)
2. Оценка качества нелинейных моделей
3. Сравнение с линейными методами

Часть 4: Байесовские методы классификации

1. Построение наивного байесовского классификатора (Gaussian Naive Bayes)
2. Исследование предположения о независимости признаков
3. Оценка качества байесовского классификатора
4. Сравнение с предыдущими методами

Часть 5: Дискриминантный анализ

1. Построение модели линейного дискриминантного анализа (LDA)
2. Построение модели квадратичного дискриминантного анализа (QDA)
3. Сравнение предположений LDA и QDA
4. Оценка качества дискриминантного анализа
5. Визуализация разделяющих поверхностей

Часть 6: Сравнительный анализ

1. Сравнение всех моделей по метрикам качества
2. Анализ стабильности моделей с помощью кросс-валидации
3. Сравнение времени обучения и предсказания
4. Анализ кривых обучения
5. Формулирование практических рекомендаций

Необходимое программное обеспечение

Для Python

```
#
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score, StratifiedKFold
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC, LinearSVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis,
    QuadraticDiscriminantAnalysis
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.metrics import confusion_matrix, classification_report, roc_curve, auc
from sklearn.model_selection import learning_curve
import time
```

Для R

```
#
library(tidyverse)
library(caret)
library(e1071)
library(MASS)
library(randomForest)
library(rpart)
library(pROC)
library(ggplot2)
library(corrplot)
library(klaR)
```

Метрики оценки качества

Для сравнения моделей использовать следующие метрики:

- **Accuracy** - общая точность классификации
- **Precision** - точность положительного прогноза
- **Recall** - полнота классификации
- **F1-score** - гармоническое среднее precision и recall
- **Время обучения** - время построения модели
- **Время предсказания** - время классификации новых объектов

Требования к отчету

Содержание отчета

1. Титульный лист
2. Цель и задачи работы
3. Описание выбранного датасета
4. Методика исследования
5. Результаты по каждой части:

- Код реализации
- Таблицы с результатами
- Графики и визуализации
- Статистические выводы

6. Сравнительный анализ всех моделей
7. Выводы и практические рекомендации

Визуализация

- Матрица корреляций с тепловой картой
- Распределение классов в данных
- Матрицы ошибок для всех моделей
- Кривые обучения для анализа переобучения
- Сравнение метрик качества на графиках

Аналитические таблицы

- Сравнение метрик качества для всех моделей
- Результаты кросс-валидации (среднее и стандартное отклонение)
- Время обучения и предсказания для каждой модели

Сравнение моделей

1. **Сводная таблица результатов**

Модель	Accuracy	Precision	Recall	F1-score	Время обучения (с)
Логистическая регрессия					
Линейный SVM					
SVM (RBF)					
Naive Bayes					
LDA					
QDA					

Таблица 1: Сравнение метрик качества классификаторов

2. **Анализ стабильности моделей**

Модель	Средняя accuracy (CV)	Std accuracy (CV)
Логистическая регрессия		
Линейный SVM		
SVM (RBF)		
Naive Bayes		
LDA		
QDA		

Таблица 2: Результаты кросс-валидации (5-fold)

Вопросы для самоконтроля

1. Какие методы лучше работают на линейно разделимых данных?
2. Когда следует предпочесть нелинейные методы линейным?
3. В каких случаях байесовские методы показывают хорошие результаты?
4. Каковы основные предположения LDA и QDA?
5. Как выбрать между LDA и QDA на практике?
6. Как объем данных влияет на выбор метода классификации?
7. Какие методы требуют больше вычислительных ресурсов?