Лабораторная работа 2. Алгоритмы классификации

Общий порядок выполнения задач

Общий порядок выполнения задач классификации:

1. Понимание проблемы и данных. Посмотрите на ваши данные, понимайте смысл каждого признака и формулируйте цель классификации.
2. Предварительная обработка данных. Это может включать масштабирование признаков, обработку пропущенных значений, преобразование категориальных переменных, а также обработку несбалансированных данных.
3. Разделение данных на обучающую и тестовую выборку. Это помогает оценить, как модель будет работать на новых данных.
4. Выбор и обучение классификаторов. Выберите некоторые модели, которые вы хотите обучить, он могут включать в себя логистическую регрессию, SVM, случайный лес, градиентный бустинг, наивный байесовский классификатор, k-NN, ансамблиевые методы и так далее.
5. Сравнение производительности моделей. Это может включать использование различных метрик, таких как точность, полнота, AUC-ROC. Вы также можете использовать матрицу ошибок для более подробного анализа результатов.
6. Подбор гиперпараметров. Используйте GridSearchCV или RandomizedSearchCV для оптимизации гиперпараметров ваших моделей.
7. Визуализация результатов. Визуализация может помочь вам понять, как ваша модель справляется с данными. Вы можете визуализировать важность признаков, ROC-кривые или любой другой интересующий вас аспект данных или моделей.
8. Интерпретация результатов. Основываясь на ваших результатах, вы можете интерпретировать, как ваша модель справляется с данными и какие признаки наиболее важны для классификации.
9. Оптимизация модели. На основе ваших выводов, вы можете внести улучшения в модели или процесс предварительной обработки данных и повторить процесс обучения, чтобы увидеть, приведет ли это к улучшению производительности.

Наборы данных

Задача 1. Классификация изображений рукописных цифр

Набор данных: [MNIST](http://yann.lecun.com/exdb/mnist/)

Задача 2. Определение мошенничества с кредитными картами

Набор данных: [Credit Card Fraud Detection](https://www.kaggle.com/mlg-ulb/creditcardfraud)

Задача 3. Классификация рака молочной железы

Набор данных: [Breast Cancer Wisconsin (Diagnostic)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

Задача 4. Спам-фильтрация для текстовых сообщений

Набор данных: [SMS Spam Collection](https://www.kaggle.com/uciml/sms-spam-collection-dataset)

Задача 5. Классификация новостных статей

Набор данных: [20 Newsgroups](https://scikit-learn.org/stable/datasets/real_world.html#newsgroups-dataset)

Задача 6. Определение типа стекла

Набор данных: [Glass Identification Dataset](https://archive.ics.uci.edu/ml/datasets/glass+identification)

Большинство наборов данных доступны для скачивания в формате csv или можно загрузить напрямую в Python, используя библиотеки `sklearn`, `seaborn` или `tensorflow`.

Примерный перечень задач

Задача 1. Классификация изображений рукописных цифр

- Набор данных: MNIST.
- Предварительная обработка: масштабирование пикселей, разделение данных на обучение и тестирование.
- Модели: логистическая регрессия, SVM, случайный лес, градиентный бустинг, нейронные сети.

- Сравнение модели: точность, матрица ошибок, ROC-кривые.
- Изучение и подбор гиперпараметров: GridSearchCV или RandomizedSearchCV.
- Визуализация: отображение изображений, важности признаков или активаций нейронной сети.

Задача 2. Определение мошенничества с кредитными картами

- Набор данных: Credit Card Fraud Detection dataset на Kaggle.
- Предварительная обработка: масштабирование, обработка несбалансированных данных.
- Модели: логистическая регрессия, SVM, случайный лес, градиентный бустинг, нейронные сети.
- Сравнение моделей: точность, полнота, ROC-кривые.
- Изучение и подбор гиперпараметров: GridSearchCV или RandomizedSearchCV.
- Визуализация: важности признаков, ROC-кривые.

Задача 3. Классификация рака молочной железы

- Набор данных: Breast Cancer Wisconsin (Diagnostic).
- Предварительная обработка: масштабирование, разделение данных на обучающие и тестовые.
- Модели: логистическая регрессия, SVM, решающие деревья, случайный лес, градиентный бустинг.
- Сравнение моделей: точность, матрица ошибок, ROC-кривые.
- Изучение и подбор гиперпараметров: GridSearchCV или RandomizedSearchCV.
- Визуализация: важности признаков, корреляционная матрица.

Задача 4. Спам-фильтрация для текстовых сообщений

- Набор данных: SMS Spam Collection на Kaggle.
- Предварительная обработка: векторизация текста, масштабирование, разделение данных на обучающие и тестовые.

- Модели: логистическая регрессия, SVM, случайный лес, градиентный бустинг.
- Сравнение моделей: точность, матрица ошибок, ROC-кривые.
- Изучение и подбор гиперпараметров: GridSearchCV или RandomizedSearchCV.
- Визуализация: важности признаков.

Задача 5. Классификация новостных статей

- Набор данных: 20 Newsgroups dataset на sklearn.
- Предварительная обработка: векторизация текста/TF-IDF, масштабирование, разделение данных на обучающие и тестовые.
- Модели: логистическая регрессия, SVM, случайный лес, градиентный бустинг.
- Сравнение моделей: точность, матрица ошибок.
- Изучение и подбор гиперпараметров: GridSearchCV или RandomizedSearchCV.
- Визуализация: важности признаков.

Задача 6. Определение типа стекла

- Набор данных: Glass Identification Dataset на UCI Machine Learning Repository.
- Предварительная обработка: масштабирование, разделение данных на обучающие и тестовые.
- Модели: логистическая регрессия, SVM, k-ближайших соседей, случайный лес, градиентный бустинг.
- Сравнение моделей: точность, матрица ошибок.
- Изучение и подбор гиперпараметров: GridSearchCV или RandomizedSearchCV.
- Визуализация: важности признаков, корреляционная матрица.

Многие из этих задач являются многоклассовыми задачами классификации и обрабатывают данные различной природы (тексты, изображения, структурированные данные), что дает возможность понять широкий спектр подходов к обработке данных и моделированию.