Лабораторная работа 1. Уменьшение объема данных

Статистические методы отбора признаков

1. Задание 1: Отбор признаков с использованием корреляционных матриц.

Для начала, используйте датасет о винах из sklearn.datasets. Найти корреляционную матрицу для датасета и, дополнительно, визуализируйте ее с помощью seaborn.

2. Задание 2: Отбор признаков на основе важности признаков в случайном лесе.

Используйте тот же датасет и подгоните модель RandomForestRegressor к вашим данным, затем используйте атрибут feature_importances_ чтобы определить наиболее важные признаки.

3. Задание 3: Использование выбора признаков на основе p-value.

Для этого задания можно использовать набор данных Boston Housing из sklearn.datasets.

4. Задание 4: Отбор признаков с помощь метода взаимной информации.

Используйте любой датасет, с которым вам приятно работать. Установите библиотеку sklearn и примените функцию mutual_info_classif или mutual_info_regression для отбора признаков.

5. Задание 5: Используйте метод Recursive Feature Elimination на том же датасете.

Постройте модель, например, линейную регрессию или SVM, и используйте RFE для выбора лучшего подмножества признаков.

6. Задание 6: Сравните различные методы отбора признаков.

Примените каждый из методов отбора признаков к одному и тому же датасету, а затем сравните производительность моделей машинного обучения, обученных на этих различных подмножествах признаков.

Каждое задание должно включать в себя следующие шаги: загрузка и предварительная обработка данных, применение метода отбора признаков, обучение модели на выбранных признаках и оценка производительности модели. Для оценки модели могут быть использованы такие метрики, как accuracy для задач классификации и MSE для задач регрессии.

Здесь можно найти больше датасетов для этих задач:

- [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/index.php),
- [Kaggle Datasets](https://www.kaggle.com/datasets),
- [Google's Dataset Search](https://datasetsearch.research.google.com/)

Корреляционные методы отбора признаков

1. Задание 1: Работа с корреляционной матрицей.

Используйте набор данных "Iris" из sklearn.datasets. Вычислите корреляционную матрицу числовых признаков. Затем визуализируйте эту матрицу с помощью heatmap в библиотеке seaborn.

2. Задание 2: Исключение мультиколлинеарных признаков.

Используйте набор данных "Wine" из sklearn.datasets. Вычислите корреляционную матрицу, а затем найдите и исключите признаки, у которых корреляция друг с другом превышает заданный порог.

3. Задание 3: Выбор наиболее значимых признаков.

Используйте набор данных "Boston Housing" из sklearn.datasets. Вычислите коэффициенты корреляции между каждым признаком и целевой переменной, затем выберите n признаков с наибольшим абсолютным значением коэффициента.

4. Задание 4: Применение Ранговой корреляции Спирмена.

Используйте любой набор данных, имеющий порядковые признаки. Примените корреляцию Спирмена для выбора наиболее значимых признаков.

5. Задание 5: Сравнение методов отбора признаков.

Используйте один и тот же набор данных для применения различных методов отбора признаков, включая корреляционные методы, и сравните результаты.

6. Задание 6: Исследование влияния предобработки данных на результаты корреляционного анализа.

Примените различные методы предобработки (например, нормализацию, стандартизацию, логарифмирование) к данным перед вычислением корреляции и сравните полученные результаты.

Каждое задание должно включать в себя следующие шаги: загрузка и предварительная обработка данных, применение метода отбора признаков, и, при необходимости, обучение модели на выбранных признаках и оценка производительности модели.

Здесь можно найти больше датасетов для этих задач:

- UCI Machine Learning Repository: [ссылка](http://archive.ics.uci.edu/ml/index.php)
- - Kaggle Datasets: [ссылка](https://www.kaggle.com/datasets)
- - Google's Dataset Search: [ссылка](https://datasetsearch.research.google.com/)

Методы-обертки

1. Задание 1: Рекурсивное исключение признаков (RFE).

Используйте набор данных "Iris" из sklearn.datasets. Примените метод RFE с использованием модели логистической регрессии. Укажите количество признаков для выбора и сравните производительность модели с и без этих признаков.

2. Задание 2: Sequential Feature Selector.

Используйте набор данных "Boston Housing" из sklearn.datasets. Используйте Sequential Feature Selector для выбора признаков с использованием модели Random Forest. Визуализируйте "важность" признаков.

3. Задание 3: Использование метода-обертки при кросс-валидации.

Используйте любой набор данных на ваше усмотрение. Выберите модель машинного обучения и метод-обертку для отбора признаков. Примените кросс-валидацию, чтобы оценить эффективность этого подхода.

4. Задание 4: Сравнение методов-оберток.

Используйте один и тот же набор данных для применения различных методов-оберток для отбора признаков, например, RFE и Sequential Feature Selector, и сравните полученные результаты.

5. Задание 5: Анализ предсказательной способности признаков.

Используйте набор данных "Wine" из sklearn.datasets. Выберите модель машинного обучения и метод-обертку для отбора признаков и исследуйте, как влияет отбор признаков на предсказательную способность модели.

Каждое задание должно включать в себя следующие шаги: загрузка и предварительная обработка данных, применение метода отбора признаков, обучение модели на выбранных признаках и оценка производительности модели.

Если вы хотите найти больше данных для этих задач, вы можете выйти на такие источники:

- - UCI Machine Learning Repository: [ссылка](http://archive.ics.uci.edu/ml/index.php)
- - Kaggle Datasets: [ссылка](https://www.kaggle.com/datasets)
- - Google's Dataset Search: [ссылка](https://datasetsearch.research.google.com/)

Линейные методы

1. Задание 1: Применение метода главных компонент (PCA).

Используйте набор данных "Iris" из sklearn.datasets. Примените PCA, чтобы снизить размерность до 2-х и визуализируйте результаты.

2. Задание 2: Сравнение PCA и Factor Analysis.

Используйте набор данных "Wine" из sklearn.datasets. Примените PCA и Factor Analysis, чтобы снизить размерность до 2-х и визуализируйте различия в результатах.

3. Задание 3: Исследование влияния предобработки данных на результаты PCA.

Используйте набор данных "Boston Housing" из sklearn.datasets. Примените различные методы предобработки (например, масштабирование, нормализацию) перед применением PCA и сравните полученные результаты.

4. Задание 4: Применение Disciminant Analysis.

Используйте набор данных "Iris" из sklearn.datasets. Примените Linear Discriminant Analysis (LDA) и Quadratic Discriminant Analysis (QDA), чтобы снизить размерность до 2-х и визуализируйте результаты.

5. Задание 5: Сравнение PCA и LDA.

Используйте любой набор данных с классифицирующей моделью. Примените PCA и LDA и сравните, как влияют эти методы снижения размерности на эффективность классификации.

Каждое задание должно включать выгрузку и предварительную обработку данных, применение метода снижения размерности, а также, возможно, обучение модели на полученных признаках и оценку ее производительности.

Если вам нужно больше данных для этих задач, вы можете посетить эти ресурсы:

- - UCI Machine Learning Repository: [ссылка](http://archive.ics.uci.edu/ml/index.php)
- - Kaggle Datasets: [ссылка](https://www.kaggle.com/datasets)
- - Google's Dataset Search: [ссылка](https://datasetsearch.research.google.com/)

Нелинейные методы

1. Задание 1: Применение метода t-SNE.

Используйте набор данных "Iris" из sklearn.datasets. Примените t-SNE, чтобы снизить размерность до 2-х, и визуализируйте результаты.

2. Задание 2: Определение влияния параметров t-SNE.

Используйте тот же набор данных "Iris". Примените t-SNE с разными значениями параметров (например, число итераций, learning rate) и сравните полученные результаты.

3. Задание 3: Сравнение t-SNE и PCA.

Используйте любой набор данных на ваше усмотрение. Примените PCA и t-SNE, чтобы снизить размерность до 2-х, и сравните разницы в результатах визуализации.

4. Задание 4: Применение UMAP.

Используйте набор данных "MNIST" (например, содержащийся в sklearn.datasets). Примените UMAP для снижения размерности и визуализируйте результаты.

5. Задание 5: Сравнение UMAP и t-SNE.

Используйте один и тот же набор данных для применения UMAP и t-SNE. Сравните влияние этих методов снижения размерности на визуальное разделение классов в данных.

Каждое задание должно включать в себя следующие шаги: загрузка и предварительная обработка данных, применение метода снижения размерности, и, при необходимости, обучение модели на полученных признаках и оценка производительности модели.

Если вам нужно больше данных для этих задач, можете использовать эти ресурсы:

- - UCI Machine Learning Repository: [ссылка](http://archive.ics.uci.edu/ml/index.php)
- - Kaggle Datasets: [ссылка](https://www.kaggle.com/datasets)
- - Google's Dataset Search: [ссылка](https://datasetsearch.research.google.com/)