Fourth Assignment (Data Science)

Due to: 15 Dec 2023

Supervised Learning – Discrete Variable Prediction

Fraud detection is an important aspect of any business or organization. It helps in identifying and preventing fraudulent activities that can cause financial losses. With the advent of technology, fraudsters have become more sophisticated in their methods, making it difficult for traditional methods of fraud detection to be effective. This is where data science comes in. By using machine learning algorithms, data scientists can develop models that can accurately predict fraudulent activities. This assignment will explore different machine-learning methods for fraud detection.

Model Selection

In this section you will perform methods listed below for predicting discrete variables (a.k.a. classification).

1. Logistic Regression Classifier
2. Support Vector Machine (SVM) Classifier
3. K-nearest neighbor (KNN) Classifier
4. Decision Tree Classifier
5. Random Forest Classifier
6. Naive Bayes

Dataset and task

The dataset to perform classification is the credit card fraud detection dataset and the task is to predict whether a transaction is fraud or not. You should train your model on the train dataset and report your models accuracy on both the test set and the train set. . In this case, you should keep 0.2 of the dataset as the test set and fit the model on the rest and report the accuracy of your model on both of these sets. In addition, make sure to have the AUC curve, confusion matrix, F1score, and the decision boundary of your predictions in your report. After the training phase, write **Pipeline** using sklearn library for your code.

NOTE

Bear in mind that preprocessing steps are mandatory for this task, e.g., feature engineering, and feature selection. Don't limit yourself only to the aforementioned methods, based on the quality of your work, extra scores may be granted for observing and testing other classification algorithms. YOU ARE NOT ALLOWED TO USE ANY DEEP LEARNING MODELS. Once again, we emphasize the report; it should contain all your questions and your innovative findings. Use figures, pictures, and tables, and DO NOT PUT ANY CODE IN THE REPORT.

Data

Click to download it.

Handwriting digits(Extra Point)

In this section, you are tasked with predicting handwritten digits using machine learning methods (**Do not use deep learning models**). This presents a unique challenge due to the high dimensionality of image data, which can lead to the **"curse of dimensionality"**. To address this issue, you will explore Feature Reduction Methods(such as PCA, t-SNE, and ...), utilizing their capabilities to efficiently capture the essential information from the handwritten digits while reducing the overall dimensionality of the data. This focus on dimensionality reduction will not only enable successful digit prediction with non-deep learning models but also provide valuable insights into the inherent structure and characteristics of handwritten data. Use different types of approaches to solve other challenge of data and report your achievements.

NOTE

Don't limit yourselves! Instead of seeking a single, definitive solution, embrace the open-ended nature of this challenge.

Remember, there's no one-size-fits-all solution in the world of machine learning. Embrace the freedom to explore, experiment, and forge your own path toward achieving successful digit prediction without deep learning models. This is your opportunity to push the boundaries of traditional techniques and contribute to the advancement of knowledge in the field.

Data

Click to download it.