Documentation for intro-ml-for-ecology Repository

Overview

Welcome to the documentation for the **intro-ml-for-ecology** repository. This repository provides educational resources for a brief introduction to machine learning methods in ecology. All scripts are written in R and require packages such as 'caret', 'randomForest', 'xgboost', and 'ipred'.

The repository is structured into the following subfolders:

- **data/**: Contains datasets used in different analyses.
- **functions/**: Includes custom R functions used in the scripts.
- **codes/**: Holds the main scripts for different machine learning applications.

Machine Learning Applications

This repository covers different machine-learning applications relevant to ecology:

1. Tree-based Methods

- Implementation of **classification and regression** using various machine learning algorithms.
- **Functions:** 'MLclass.R' (classification) and 'MLregress.R' (regression).
- **Datasets:** 'biomass_urb_affor.xlsx', 'data_urb_affor.xlsx', and 'pred_urb_affor.xlsx'.
    - These datasets contain functional traits of urban trees, including morphological characteristics and planting suitability for sidewalks and parks.
- **Algorithms Used:**
    - **Logistic Regression**: A statistical method for binary and multinomial classification.
    - **Random Forest**: An ensemble learning method that builds multiple decision trees and combines their outputs for better accuracy.
    - **Support Vector Machines (SVM)**: A method that finds the optimal hyperplane to separate different classes.
    - **k-Nearest Neighbors (kNN)**: A non-parametric method that classifies based on the majority class of its nearest neighbors.
    - **Gradient Boosting Machine (GBM)**: A boosting method that sequentially improves weak learners to enhance model performance.

2. Image Classification

- Classification of urban areas using machine learning algorithms.
- **Dataset:** 'AjuBrazil.tif'.
    - This raster file represents the urban area of Aracaju, Northeast of Brazil.
- **Unsupervised Classification:** Uses clustering algorithms like K-means to group pixels based on similarities without labeled data.
    - **Method:** K-means clustering applied to satellite images.
    - **Process:**
        1. Convert raster image to a data frame.
        2. Perform K-means clustering with a defined number of clusters.
        3. Convert clustered data back to a raster image.
        4. Save and visualize the clustered classification.
- **Semi-Supervised Classification:** Uses manually labeled pixels and Random Forest for classification.
    - **Process:**
        1. Load a raster image.
        2. Provide manually labeled pixels for training.
        3. Extract pixel values for labeled locations.
        4. Train a Random Forest model.
        5. Predict classes for the entire image.
        6. Convert predictions to a raster and save the classified image.

3. Natural Language Processing (NLP)

- Basic introduction to NLP for ecological data.
- **Datasets:** 'article1.pdf' and 'article2.pdf'.
    - These articles contain information on functional trairs of trees.
- **Function:** 'seek_att.R' (located in the functions/ folder).
- **Text Mining (tm)**: Preprocessing textual data for analysis.
- **Vectorization (text2vec)**: Converts text into numerical features suitable for machine learning models.

4. Species Distribution Modelling

- Predicting species distributions using machine learning algorithms.
- **Datasets:** 'climatic_data.xlsx' and 'occurrence.xlsx'.
    - These files contain tree occurrence records and environmental data such as precipitation and temperature.
- **Presence-Absence Models**: Classifies locations based on whether a species is present or not.
- **Environmental Suitability Modelling**: Uses environmental variables to predict species occurrence probability.

5. Other Applications

- **Datasets:** 'climatic_data.xlsx', 'land_use.xlsx', and 'vegetation_data.xlsx'.
    - These files contain climate data, land use information, and forest biomass data.
- **Support Vector Machines (SVM)**: A classification method that finds the optimal hyperplane to separate different classes.
- **Random Forest**: Used here for feature selection and predictive modeling.
- **k-Nearest Neighbors (kNN)**: A simple algorithm that classifies data points based on the majority class of their nearest neighbors.
- **K-Means Clustering**: A clustering method that groups similar data points into a predefined number of clusters.

Functions Overview

The **functions/** subfolder contains custom R functions that simplify data processing and analysis. Below are key functions and their purposes:

Function	Description
MLclass.R	Applies multiple machine learning algorithms for classification tasks.
MLregress.R	Implements various machine learning regression models.
seek_att.R	Automates the search for functional traits in PDF files within a folder.

Required R Packages

To run the scripts successfully, install and load the following packages:

install.packages(c("caret", "randomForest", "xgboost", "ipred", "e1071", "tm", "text2vec", "rpart", "gbm", "kernlab", "ggplot2", "openxlsx", "elasticnet", "raster", "terra", "rasterVis")).

Usage Instructions

1. Clone the repository:

git clone https://github.com/PIBILab/intro-ml-for-ecology.git

2. Open RStudio and navigate to the cloned directory.
3. Load required packages using the library().
4. Run the scripts from the **codes/** folder based on the desired analysis.

For detailed descriptions of individual scripts, functions, and data, refer to Report.pdf in the repository.

Maintainer: PIBILab
Repository URL: [https://github.com/PIBILab/intro-ml-for-ecology]