

Package ‘CDF’

May 7, 2026

Type Package

Title Centroid Decision Forest for High-Dimensional Classification

Version 0.1.0

Description Implements the Centroid Decision Forest (CDF) as a single user-facing function `CDF()`. The method selects discriminative features via a multi-class class separability score (CSS), splits by nearest class centroid, and aggregates tree votes to produce predictions and class probabilities. Returns CSS-based feature importance as well. Amjad Ali, Saeed Aldahmani, Zardad Khan (2025) <[doi:10.48550/arXiv.2503.19306](https://doi.org/10.48550/arXiv.2503.19306)>.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 3.6)

Imports matrixStats, utils

NeedsCompilation no

RoxygenNote 7.3.3

Author Amjad Ali [aut, cre],
Saeed Aldahmani [aut],
Zardad Khan [aut]

Maintainer Amjad Ali <amjadali@uaeu.ac.ae>

Repository CRAN

Date/Publication 2025-11-10 08:20:32 UTC

Contents

CDF	2
DARWIN	3
Index	5

Description

Trains an ensemble of centroid-splitting trees and predicts for new data. Nodes select top-k features via a multi-class class separability score (CSS), split by nearest class centroid, and aggregate votes.

Usage

```
CDF(xtrain, ytrain, xtest, ntrees = 500, depth = 3, mnode = 3,  
    k = round(2 * log(ncol(xtrain))), mtry = round(0.2 * ncol(xtrain)), seed = NULL)
```

Arguments

xtrain	Numeric matrix or data frame of training predictors.
ytrain	Factor or character vector of class labels (length = nrow(xtrain)).
xtest	Numeric matrix or data frame of test predictors.
ntrees	Integer. Number of trees (default 500).
depth	Integer. Maximum tree depth (default 3).
mnode	Integer. Minimum node size to split (default 3).
k	Integer. Top-k CSS-ranked features per split (default $\text{round}(2 \cdot \log(p))$).
mtry	Integer. Candidate features per node (default $\text{round}(0.2 \cdot p)$).
seed	Optional integer seed for reproducibility.

Value

A list with:

predictions	Character vector of predicted classes for xtest.
probabilities	Numeric matrix of class probabilities (columns are classes).
feature_importance	Named numeric vector of normalized CSS importances.

Author(s)

Amjad Ali, Saeed Aldahmani, Zardad Khan

References

Ali, A., Khan, Z., and Aldahmani, S. (2025). *Centroid Decision Forest*. arXiv:2503.19306.

Examples

```

data(DARWIN)
set.seed(2025)
n <- nrow(DARWIN)
p <- ncol(DARWIN)

# Split the data into training (70%) and test (30%) sets
tr <- sample(seq_len(n), floor(0.7 * n))
te <- setdiff(seq_len(n), tr)

# Prepare training and test matrices
Xtr <- as.matrix(DARWIN[tr, 1:(p - 1), drop = FALSE])
ytr <- DARWIN$Y[tr]
Xte <- as.matrix(DARWIN[te, 1:(p - 1), drop = FALSE])
yte <- DARWIN$Y[te]

# Fit the CDF model
FitCDF <- CDF(Xtr, ytr, Xte, ntrees = 100, seed = 2025)

# Compute classification accuracy
mean(FitCDF$predictions == yte)

# Predicted classes for the test data
FitCDF$predictions

# Predicted class probabilities for the test data
FitCDF$probabilities

# Top 10 most important features
order(FitCDF$feature_importance, decreasing = TRUE)[1:10]

```

DARWIN

DARWIN Handwriting Dataset (P vs H)

Description

Handwriting data from **174** participants for a binary classification task: distinguishing Alzheimer's disease patients (P) from healthy controls (H). The feature matrix contains **450** numeric handwriting-derived measures; the last column Y is the class label with levels P and H.

Usage

```
data(DARWIN)
```

Format

A data frame with **174** rows and **451** columns:

Columns [1–450] Numeric handwriting features (predictors).

Y Factor, class label with two levels: P (patients), H (healthy).

Details

- **Purpose:** Research on predicting Alzheimer's via handwriting analysis.
- **Instances:** 174
- **Features (X):** 450 (numeric)
- **Target (Y):** P/H
- **Missing values:** None

Source

OpenML dataset ID 46606. <https://www.openml.org/search?type=data&status=any&id=46606&sort=runs>

Examples

```
data(DARWIN)
dim(DARWIN)
table(DARWIN$Y)
```

Index

CDF, [2](#)

DARWIN, [3](#)