Package ‘CytoProfile’

March 8, 2025

Title Cytokine Profiling Analysis Tool

Version 0.0.0.9000

Description CytoProfile is a comprehensive tool for cytokine profiling analysis.
It supports quality control using biologically meaningful cutoffs on raw cytokine
measurements and tests for distributional symmetry to suggest appropriate transformations.
The package offers exploratory data analysis with summary statistics, enhanced boxplots, and
barplots, along with both univariate and multivariate analysis capabilities for in-depth cytokine profiling.

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

URL https://github.com/saraswatsh/CytoProfile, https://saraswatsh.github.io/CytoProfile/

Depends R (>= 3.5)

Imports mixOmics,
moments,
dplyr,
tidyr,
pROC,
plot3D,
caret,
xgboost,
randomForest,
gplots,
e1071,
ggplot2,
ggrepel,
gridExtra,
reshape2

Suggests BiocManager,
testthat,
knitr,
rmarkdown,
devtools,
Ckmeans.1d.dp

NeedsCompilation no

License GPL (>= 2)

LazyData true

VignetteBuilder knitr

Contents

`cytodata` *Sample Cytokine Profiling Data.*

Description

Contains observed values of cytokines and their respective treatment and groups, derived from:

Usage

```
cytodata
```

Format

A data frame with 297 rows and 29 columns:

Group Group assigned to the subjects.

Stimulation Stimulation received by subjects.

Treatment Treatment received by subjects.

... Additional numeric columns representing measured cytokines.

Source

Example data compiled for cytokine profiling.

References

Pugh GH, Fouladvand S, SantaCruz-Calvo S, Agrawal M, Zhang XD, Chen J, Kern PA, Nikolajczyk BS. T cells dominate peripheral inflammation in a cross-sectional analysis of obesity-associated diabetes. *Obesity (Silver Spring).* 2022;30(10): 1983–1994. doi:10.1002/oby.23528.

Examples

```
data(cytodata)
```

cyt_anova *ANOVA Analysis on Continuous Variables.*

Description

This function performs an analysis of variance (ANOVA) for each continuous variable against every categorical predictor in the input data. Character columns are automatically converted to factors; all factor columns are used as predictors while numeric columns are used as continuous outcomes. For each valid predictor (i.e., with more than one level and no more than 10 levels), Tukey's Honest Significant Difference (HSD) test is conducted and the adjusted p-values for pairwise comparisons are extracted.

Usage

```
cyt_anova(data)
```

Arguments

data	A data frame or matrix containing both categorical and continuous variables. Character columns will be converted to factors and used as predictors, while numeric columns will be used as continuous outcomes.

Value

A list of adjusted p-values from Tukey's HSD tests for each combination of continuous outcome and categorical predictor. List elements are named in the format "outcome_predictor".

Examples

```
# Loading data
data("cytodata")
# Perform ANOVA on selected columns of the cytodata dataset
anova_results <- cyt_anova(cytodata[, c(2:4, 5:6)])
print(anova_results)
```

cyt_bp *Boxtplots for Overall Comparisons by Continous Variables.*

Description

This function creates a PDF file containing box plots for the continuous variables in the provided data. If the number of columns in data exceeds bin.size, the function splits the plots across multiple pages.

Usage

```
cyt_bp(data, pdf_title, bin_size = 25, mf_row = c(1, 1), y_lim = NULL)
```

Arguments

`data`	A matrix or data frame containing the raw data to be plotted.
`pdf_title`	A string representing the name of the PDF file to be created.
`bin_size`	An integer specifying the maximum number of box plots to display on a single page.
`mf_row`	A numeric vector of length two specifying the layout (rows and columns) for the plots on each page.
`y_lim`	An optional numeric vector defining the y-axis limits for the plots.

Value

A PDF file containing the box plots for the continuous variables.

Examples

```
# Loading data
data.df <- cytodata
# Generate box plots for log2-transformed values to check for outliers:
cyt_bp(log2(data.df[, -c(1:4)]), pdf_title = "boxplot_by_cytokine_log2.pdf")
```

---

cyt_bp2 *Boxplot Function Enhanced for Specific Group Comparisons.*

Description

This function generates a PDF file containing boxplots for each combination of numeric and factor variables in the provided data. It first converts any character columns to factors and checks that the data contains at least one numeric and one factor column. If the scale argument is set to "log2", all numeric columns are log2-transformed. The function then creates boxplots using ggplot2 for each numeric variable grouped by each factor variable.

Usage

```
cyt_bp2(data, pdf_title, mf_row = c(1, 1), scale = NULL, y_lim = NULL)
```

Arguments

`data`	A matrix or data frame of raw data.
`pdf_title`	A string representing the title (and filename) of the PDF file.
`mf_row`	A numeric vector of length two specifying the layout (rows and columns) for the plots on each page. Defaults to c(1, 1).
`scale`	Transformation option for continuous variables. Options are NULL (default) and "log2". When set to "log2", numeric columns are transformed using the log2 function.
`y_lim`	An optional numeric vector defining the y-axis l imits for the plots.

Value

A PDF file containing the boxplots.

Examples

```
# Loading data
data_df <- cytodata[, -c(1, 4)]

cyt_bp2(data_df, pdf_title = "boxplot2_test2.pdf", scale = "log2")
```

cyt_dualflashplot *Dual-flashlight Plot.*

Description

This function reshapes the input data and computes summary statistics (mean and variance) for each variable grouped by a specified factor column. It then calculates the SSMD (Strictly Standardized Mean Difference) and log2 fold change between two groups (group1 and group2) and categorizes the effect strength as "Strong Effect", "Moderate Effect", or "Weak Effect". A dual flash plot is generated using ggplot2 where the x-axis represents the average log2 fold change and the y-axis represents the SSMD. Additionally, the function prints the computed statistics to the console.

Usage

```
cyt_dualflashplot(
  data,
  group_var,
  group1,
  group2,
  ssmd_thresh = 1,
  log2fc_thresh = 1,
  top_labels = 15
)
```

Arguments

data	A data frame containing the input data.
group_var	A string specifying the name of the grouping column in the data.
group1	A string representing the name of the first group for comparison.
group2	A string representing the name of the second group for comparison.
ssmd_thresh	A numeric threshold for the SSMD value used to determine significance. Default is 1.
log2fc_thresh	A numeric threshold for the log2 fold change used to determine significance. Default is 1.
top_labels	An integer specifying the number of top variables (based on absolute SSMD) to label in the plot. Default is 15.

Value

A ggplot object representing the dual flash plot for the comparisons between group1 and group2.

Examples

```
# Loading data
data_df <- cytodata[, -c(1, 3:4)]

dfp <- cyt_dualflashplot(
  data_df,
  group_var = "Group",
  group1 = "T2D",
  group2 = "ND",
  ssmd_thresh = -0.2,
  log2fc_thresh = 1,
  top_labels = 10
)
print(dfp)
```

---

cyt_errbp *Error-bar Plot.*

Description

This function draws an error-bar plot for comparing groups to a baseline group. It creates a barplot of the central tendency (mean or median) and overlays error bars representing the spread (e.g., standard deviation, MAD, or standard error). Optionally, p-value and effect size labels (based on SSMD) are added, either as symbols or numeric values.

Usage

```
cyt_errbp(
  data,
  p_lab = TRUE,
  es_lab = TRUE,
  class_symbol = TRUE,
  x_lab = "",
  y_lab = "",
  main = ""
)
```

Arguments

`data` A data frame containing the following columns for each group:

- `name`: Group names.
- `center`: Mean or median values.
- `spread`: Standard deviation, MAD, or standard error.
- `p.value`: P-value for the comparison.
- `effect.size`: Effect size based on SSMD.

Note: The first row of `data` must correspond to the baseline group.

`p_lab` Logical. Whether to label the p-values on the plot. Default is `TRUE`.

`es_lab` Logical. Whether to label the effect sizes on the plot. Default is `TRUE`.

class_symbol	Logical. Whether to use symbolic notation for significance and effect size. Default is TRUE.
x_lab	Character. Label for the x-axis.
y_lab	Character. Label for the y-axis.
main	Character. Title of the graph.

Value

An error-bar plot is produced.

Examples

```
# Load sample data
data_df <- cytodata[, -1]
cyt_mat <- log2(data_df[, -c(1:3)])
data_df1 <- data.frame(data_df[, 1:3], cyt_mat)
cytokineNames <- colnames(cyt_mat)
nCytokine <- length(cytokineNames)
condt <- !is.na(cyt_mat) & (cyt_mat > 0)
Cutoff <- min(cyt_mat[condt], na.rm = TRUE) / 10
# Create matrices for ANOVA and Tukey results
p_aov_mat <- matrix(NA, nrow = nCytokine, ncol = 3)
dimnames(p_aov_mat) <- list(cytokineNames,
                          c("Group", "Treatment", "Interaction"))
p_groupComp_mat <- matrix(NA, nrow = nCytokine, ncol = 3)
dimnames(p_groupComp_mat) <- list(cytokineNames,
                                  c("2-1", "3-1", "3-2"))
ssmd_groupComp_stm_mat <- mD_groupComp_stm_mat <- p_groupComp_stm_mat <-
p_groupComp_mat

for (i in 1:nCytokine) {
Cytokine <- (cyt_mat[, i] + Cutoff)
cytokine_aov <- aov(Cytokine ~ Group * Treatment, data = data_df)
aov_table <- summary(cytokine_aov)[[1]]
p_aov_mat[i, ] <- aov_table[1:3, 5]
p_groupComp_mat[i, ] <- TukeyHSD(cytokine_aov)$Group[1:3, 4]
p_groupComp_stm_mat[i, ] <- TukeyHSD(cytokine_aov)$`Group:Treatment`[1:3, 4]
mD_groupComp_stm_mat[i, ] <- TukeyHSD(cytokine_aov)$`Group:Treatment`[1:3, 1]
ssmd_groupComp_stm_mat[i, ] <- mD_groupComp_stm_mat[i, ] / sqrt(2 *
aov_table["Residuals", "Mean Sq"])
}

results <- cyt_skku(cytodata[, -c(1,4)], print_res_log = TRUE,
                 group_cols = c("Group", "Treatment"))
pdf("bar_error_plot_enriched.pdf")
par(mfrow = c(2,3), mar = c(8.1, 4.1, 4.1, 2.1))
for (k in 1:nCytokine) {
result_mat <- results[1:9, , k]
center_df <- data.frame(
 name = rownames(result_mat),
 result_mat[, c("center", "spread")],
 p.value = c(1, p_groupComp_stm_mat[k, 1:2]),
 effect.size = c(0, ssmd_groupComp_stm_mat[k, 1:2])
 )
cyt_errbp(center_df, p_lab = TRUE, es_lab = TRUE,
         class_symbol = TRUE,
          y_lab = "Concentration in log2 scale",
          main = cytokineNames[k])
}
dev.off()
```

cyt_heatmap *Heat Map.*

Description

This function creates a heatmap using the numeric columns from the provided data frame. If requested via the `scale` parameter, the function applies a log2 transformation to the data (with non-positive values replaced by NA). The heatmap is saved as a file, with the format determined by the file extension in `title`.

Usage

```
cyt_heatmap(data, scale = NULL, annotation_col_name = NULL, title)
```

Arguments

`data` A data frame containing the input data. Only numeric columns will be used to generate the heatmap.

`scale` Character. An optional scaling option. If set to "log2", the numeric data will be log2-transformed (with non-positive values set to NA). Default is NULL.

`annotation_col_name` Character. An optional column name from `data` to be used for generating annotation colors. Default is NULL.

`title` Character. The title of the heatmap and the file name for saving the plot. The file extension (".pdf" or ".png") determines the output format.

Value

The function does not return a value. It saves the heatmap to a file.

Examples

```
# Load sample data
data("cytodata")
data_df <- cytodata
# Generate a heatmap with log2 scaling and annotation based on
# the "Group" column
cyt_heatmap(
  data = data_df[, -c(1,3,4)],
  scale = "log2",  # Optional scaling
  annotation_col_name = "Group",
  title = "Heatmap.png"
)
```

cyt_pca *Analyze Data with Principal Component Analysis (PCA) for Cytokines.*

Description

This function performs Principal Component Analysis (PCA) on cytokine data and generates several types of plots, including:

- 2D PCA plots using mixOmics's `plotIndiv` function,
- 3D scatter plots (if `style` is "3d" or "3D" and `comp_num` is 3) via the plot3D package,
- Scree plots showing both individual and cumulative explained variance,
- Loadings plots, and
- Biplots and correlation circle plots.

The function optionally applies a log2 transformation to the numeric data and handles analyses based treatment groups.

Usage

```
cyt_pca(
  data,
  group_col = NULL,
  trt_col = NULL,
  colors = NULL,
  pdf_title,
  ellipse = FALSE,
  comp_num = 2,
  scale = NULL,
  pch_values = NULL,
  style = NULL
)
```

Arguments

`data`	A data frame containing cytokine data. It should include at least one column representing grouping information and optionally a second column representing treatment or stimulation.
`group_col`	Character. The name of the column containing the grouping information. If not specified and `trt_col` is provided, the treatment column will be used as the grouping variable.
`trt_col`	Character. The name of the column containing the treatment information. If not specified and `group_col` is provided, the grouping column will be used as the treatment variable.
`colors`	A vector of colors corresponding to the groups. If set to NULL, a palette is generated using `rainbow()` based on the number of unique groups.
`pdf_title`	A string specifying the file name of the PDF where the PCA plots will be saved.
`ellipse`	Logical. If TRUE, a 95% confidence ellipse is drawn on the PCA individuals plot. Default is FALSE.

comp_num	Numeric. The number of principal components to compute and display. Default is 2.
scale	Character. If set to "log2", a log2 transformation is applied to the numeric cytokine measurements (excluding the grouping columns). Default is NULL.
pch_values	A vector of plotting symbols (pch values) to be used in the PCA plots. Default is NULL.
style	Character. If set to "3d" or "3D" and comp_num equals 3, a 3D scatter plot is generated using the plot3D package. Default is NULL.

Value

A PDF file containing the PCA plots is generated and saved.

Examples

```
# Load sample data
data("cytodata")
# Subset data to exclude columns 1, 4, and 24, then filter out rows
# where Group is "ND" and Treatment is "Unstimulated"
data_subset <- cytodata[, -c(1, 4, 24)]
data_df <- dplyr::filter(data_subset,
Group != "ND" & Treatment != "Unstimulated")
# Run PCA analysis and save plots to a PDF file
cyt_pca(
  data = data_df,
  pdf_title = "Example_PCA_Analysis.pdf",
  colors = c("black", "red2"),
  scale = "log2",
  comp_num = 3,
  pch_values = c(16, 4),
  style = "3D",
  group_col = "Group",
  trt_col = "Treatment",
  ellipse = FALSE
)
```

cyt_rf *Run Random Forest Classification on Cytokine Data,*

Description

This function trains and evaluates a Random Forest classification model on cytokine data. It includes feature importance visualization, cross- validation for feature selection, and performance metrics such as accuracy, sensitivity, and specificity. Optionally, for binary classification, the function also plots the ROC curve and computes the AUC.

Usage

```
cyt_rf(
  data,
  group_col,
  ntree = 500,
  mtry = 5,
  train_fraction = 0.7,
  plot_roc = FALSE,
  k_folds = 5,
  step = 0.5,
  run_rfcv = TRUE
)
```

Arguments

`data`	A data frame containing the cytokine data, with one column as the grouping variable and the rest as numerical features.
`group_col`	A string representing the name of the column with the grouping variable (the target variable for classification).
`ntree`	An integer specifying the number of trees to grow in the forest (default is 500).
`mtry`	An integer specifying the number of variables randomly selected at each split (default is 5).
`train_fraction`	A numeric value between 0 and 1 representing the proportion of data to use for training (default is 0.7).
`plot_roc`	A logical value indicating whether to plot the ROC curve and compute the AUC for binary classification (default is FALSE).
`k_folds`	An integer specifying the number of folds for cross-validation (default is 5).
`step`	A numeric value specifying the fraction of variables to remove at each step during cross-validation for feature selection (default is 0.5).
`run_rfcv`	A logical value indicating whether to run Random Forest cross-validation for feature selection (default is TRUE).

Details

The function fits a Random Forest model to the provided data by splitting it into training and test sets. It calculates performance metrics such as accuracy, sensitivity, and specificity for both sets. For binary classification, it can also plot the ROC curve and compute the AUC. If `run_rfcv` is TRUE, cross-validation is performed to select the optimal number of features.

Value

A list containing:

`model`	The trained Random Forest model.
`confusion_matrix`	The confusion matrix of the test set predictions.
`importance_plot`	A ggplot object showing the variable importance plot based on Mean Decrease Gini.
`rfcv_result`	Results from Random Forest cross-validation for feature selection (if `run_rfcv` is TRUE).
`importance_data`	A data frame containing the variable importance based on the Gini index.

Examples

```
data.df0 <- cytodata
data.df <- data.frame(data.df0[, 1:4], log2(data.df0[, -c(1:4)]))
data.df <- data.df[, -c(1, 3, 4)]
data.df <- dplyr::filter(data.df, Group != "ND")

results <- cyt_rf(
  data = data.df, group_col = "Group", k_folds = 5, ntree = 1000,
  mtry = 4, run_rfcv = TRUE, plot_roc = TRUE
)
```

cyt_skku *Distribution of the Data Set Shown by Skewness and Kurtosis.*

Description

This function computes summary statistics — including sample size, mean, standard error, skewness, and kurtosis — for each numeric measurement column in a data set. If grouping columns are provided via `group.cols`, the function computes the metrics separately for each group defined by the combination of these columns (using the first element as the treatment variable and the second as the grouping variable, or the same column for both if only one is given). When no grouping columns are provided, the entire data set is treated as a single group ("Overall"). A log2 transformation (using a cutoff equal to one-tenth of the smallest positive value in the data) is applied to generate alternative metrics. Histograms showing the distribution of skewness and kurtosis for both raw and log2-transformed data are then generated and saved to a PDF if a file name is provided.

Usage

```
cyt_skku(
  data,
  group_cols = NULL,
  pdf_title = NULL,
  print_res_raw = FALSE,
  print_res_log = FALSE
)
```

Arguments

`data` A matrix or data frame containing the raw data. If `group.cols` is specified, the columns with names in `group.cols` are treated as grouping variables and all other columns are assumed to be numeric measurement variables.

`group.cols` A character vector specifying the names of the grouping columns. When provided, the first element is treated as the treatment variable and the second as the group variable. If not provided, the entire data set is treated as one group.

`Title` A character string specifying the file name for the PDF file in which the histograms will be saved. If omitted, the plots are generated on the current graphics device.

`printResRaw` Logical. If `TRUE`, the function returns and prints the computed summary statistics for the raw data. Default is `FALSE`.

`printResLog` Logical. If `TRUE`, the function returns and prints the computed summary statistics for the log2-transformed data. Default is `FALSE`.

Details

A cutoff is computed as one-tenth of the minimum positive value among all numeric measurement columns to avoid taking logarithms of zero. When rouping columns are provided, the function loops over unique treatmentlevels (using the first element of `group.cols` as treatment and the second as group, if available) and computes the metrics for each measurement column within each subgroup. Without grouping columns, the entire data set is analyzed as one overall group.

Value

The function generates histograms of skewness and kurtosis for both raw and log2-transformed data. Additionally, if either `printResRaw` and/or `printResLog` is `TRUE`, the function returns the corresponding summary statistics as a data frame or a list of data frames.

Examples

```
# Example with grouping columns (e.g., "Group" and "Treatment")
data(cytodata)
cyt_skku(cytodata[, -c(1, 3, 4)], pdf_title = "Skew_and_Kurtosis.pdf",
  group_cols = c("Group")
)

# Example without grouping columns (analyzes the entire data set)
cyt_skku(cytodata[, -c(1:4)], pdf_title = "Skew_and_Kurtosis_Overall.pdf")
```

`cyt_splsda` *Analyze data with Sparse Partial Least Squares Discriminant Analysis (sPLS-DA).*

Description

This function conducts Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) on the provided data. It uses the specified `group_col` (and optionally `trt_col`) to define class labels while assuming the remaining columns contain continuous variables. The function supports a log2 transformation via the `scale` parameter and generates a series of plots, including classification plots, scree plots, loadings plots, and VIP score plots. Optionally, ROC curves are produced when `roc` is `TRUE`. Additionally, cross-validation is supported via LOOCV or Mfold methods. When both `group_col` and `trt_col` are provided and differ, the function analyzes each treatment level separately.

Usage

```
cyt_splsda(
  data,
  group_col = NULL,
  trt_col = NULL,
  colors = NULL,
  pdf_title,
  ellipse = FALSE,
  bg = FALSE,
  conf_mat = FALSE,
  var_num,
  cv_opt = NULL,
  fold_num = 5,
  scale = NULL,
  comp_num = 2,
  pch_values,
  style = NULL,
  roc = FALSE
)
```

Arguments

`data`	A matrix or data frame containing the variables. Columns not specified by `group_col` or `trt_col` are assumed to be continuous variables for analysis.
`group_col`	A string specifying the column name that contains group information. If `trt_col` is not provided, it will be used for both grouping and treatment.
`trt_col`	A string specifying the column name for treatments. Default is `NULL`.
`colors`	A vector of colors for the groups or treatments. If `NULL`, a random palette (using `rainbow`) is generated based on the number of groups.
`pdf_title`	A string specifying the file name for saving the PDF output.
`ellipse`	Logical. Whether to draw a 95\ figures. Default is `FALSE`.
`bg`	Logical. Whether to draw the prediction background in the figures. Default is `FALSE`.
`conf_mat`	Logical. Whether to print the confusion matrix for the classifications. Default is `FALSE`.
`var_num`	Numeric. The number of variables to be used in the PLS-DA model.
`cv_opt`	Character. Option for cross-validation method: either "loocv" or "Mfold". Default is `NULL`.
`fold_num`	Numeric. The number of folds to use if `cv_opt` is "Mfold". Default is 5.
`scale`	Character. Option for data transformation; if set to `"log2"`, a log2 transformation is applied to the continuous variables. Default is `NULL`.
`comp_num`	Numeric. The number of components to calculate in the sPLS-DA model. Default is 2.
`pch_values`	A vector of integers specifying the plotting characters (pch values) to be used in the plots.
`style`	Character. If set to `"3D"` or `"3d"` and `comp_num` equals 3, a 3D plot is generated using the `plot3D` package. Default is `NULL`.
`roc`	Logical. Whether to compute and plot the ROC curve for the model. Default is `FALSE`.

Value

A PDF file containing the classification figures, component figures with Variable of Importance in Projection (VIP) scores, and classifications based on VIP scores greater than 1. ROC curves and confusion matrices are also produced if requested.

Examples

```
# Loading Sample Data
data.df <- cytodata

cyt_splsda(data.df[,-c(1,4)], pdf_title = "Example sPLS-DA Analysis.pdf",
colors = c("black", "purple", "red2"), bg = TRUE, scale = "log2",
conf_mat = TRUE, var_num = 25, cv_opt = "loocv", comp_num = 3,
pch_values = c(16, 4, 3), style = "3d",
group_col = "Group", trt_col = "Treatment", roc = TRUE)
```

`cyt_ttest` *Two Sample T-test Comparisons.*

Description

This function performs pairwise comparisons between two groups for each combination of a categorical predictor (with exactly two levels) and a continuous outcome variable. It first converts any character variables in `data` to factors and applies a log2 transformation to the continuous variables if specified. Depending on the value of `scale`, the function conducts either a two-sample t-test or a Mann-Whitney U test and prints the resulting p-values. An error is thrown if a categorical variable does not have exactly two levels.

Usage

```
cyt_ttest(data, scale = NULL)
```

Arguments

`data`	A matrix or data frame containing continuous variables and categorical variables.
`scale`	A character specifying a transformation for continuous variables. Options are `NULL` (default) and `"log2"`. When `scale = "log2"`, a log2 transformation is applied and a two-sample t-test is used; when `scale` is `NULL`, a Mann-Whitney U test is performed.

Value

A list of p-values from the statistical tests for each combination of a continuous outcome and a categorical predictor is returned.

Examples

```
data_df <- cytodata[, -c(1, 4)]
data_df <- dplyr::filter(data_df, Group != "ND", Treatment != "Unstimulated")
# Two sample T-test
cyt_ttest(data_df[, c(1, 2, 5:6)], scale = "log2")
# Mann-Whitney U Test
cyt_ttest(data_df[, c(1, 2, 5:6)])
```

cyt_volc *Volcano Plot.*

Description

This function subsets the numeric columns from the input data and compares them based on a selected grouping column. It computes the fold changes (as the ratio of means) and associated p-values (using two-sample t-tests) for each numeric variable between two groups. The results are log2-transformed (for fold change) and -log10-transformed (for p-values) to generate a volcano plot.

Usage

```
cyt_volc(
  data,
  group_col,
  cond1 = NULL,
  cond2 = NULL,
  fold_change_thresh = 2,
  p_value_thresh = 0.05,
  top_labels = 10
)
```

Arguments

`data`	A matrix or data frame containing the data to be analyzed.
`group_col`	A character string specifying the column name used for comparisons (e.g., group, treatment, or stimulation).
`cond1`	A character string specifying the name of the first condition for comparison. Default is `NULL`.
`cond2`	A character string specifying the name of the second condition for comparison. Default is `NULL`.
`fold_change_thresh`	A numeric threshold for the fold change. Default is `2`.
`p_value_thresh`	A numeric threshold for the p-value. Default is `0.05`.
`top_labels`	An integer specifying the number of top variables to label on the plot. Default is `10`.

Value

A list of volcano plots (as `ggplot` objects) for each pairwise comparison. Additionally, the function prints the data frame used for plotting (excluding the significance column) from the final comparison.

Note

If `cond1` and `cond2` are not provided, the function automatically generates all possible pairwise combinations of groups from the specified `group_col` for comparisons.

Examples

```
# Loading data
data_df <- cytodata[,-4]

volc_plot <- cyt_volc(data_df, "Group", cond1 = "T2D", cond2 = "ND",
fold_change_thresh = 2.0, top_labels= 15)
print(volc_plot$`T2D vs ND`)
```

cyt_xgb *Run XGBoost Classification on Cytokine Data.*

Description

This function trains and evaluates an XGBoost classification model on cytokine data. It allows for hyperparameter tuning, cross-validation, and visualizes feature importance.

Usage

```
cyt_xgb(
  data,
  group_col,
  train_fraction = 0.7,
  nrounds = 500,
  max_depth = 6,
  eta = 0.1,
  nfold = 5,
  cv = FALSE,
  objective = "multi:softprob",
  early_stopping_rounds = NULL,
  eval_metric = "mlogloss",
  gamma = 0,
  colsample_bytree = 1,
  subsample = 1,
  min_child_weight = 1,
  top_n_features = 10,
  verbose = 1,
  plot_roc = FALSE
)
```

Arguments

`data` A data frame containing the cytokine data, with one column as the grouping variable and the rest as numerical features.

`group_col` A string representing the name of the column with the grouping variable (i.e., the target variable for classification).

`train_fraction` A numeric value between 0 and 1 representing the proportion of data to use for training (default is 0.7).

`nrounds` An integer specifying the number of boosting rounds (default is 500).

`max_depth` An integer specifying the maximum depth of the trees (default is 6).

`eta` A numeric value representing the learning rate (default is 0.1).

`nfold` An integer specifying the number of folds for cross-validation (default is 5).

`cv` A logical value indicating whether to perform cross-validation (default is FALSE).

`objective` A string specifying the XGBoost objective function (default is "multi:softprob" for multi-class classification).

`early_stopping_rounds`
An integer specifying the number of rounds with no improvement to stop training early (default is NULL).

`eval_metric` A string specifying the evaluation metric (default is "mlogloss").

`gamma` A numeric value for the minimum loss reduction required to make a further partition (default is 0).

`colsample_bytree`
A numeric value specifying the subsample ratio of columns when constructing each tree (default is 1).

`subsample` A numeric value specifying the subsample ratio of the training instances (default is 1).

`min_child_weight`
A numeric value specifying the minimum sum of instance weight needed in a child (default is 1).

`top_n_features` An integer specifying the number of top features to display in the importance plot (default is 10).

`verbose` An integer specifying the verbosity of the training process (default is 1).

`plot_roc` A logical value indicating whether to plot the ROC curve and calculate the AUC for binary classification (default is FALSE).

Details

The function allows for training an XGBoost model on cytokine data, splitting the data into training and test sets. If cross-validation is enabled (`cv = TRUE`), it performs k-fold cross-validation and prints the best iteration based on the evaluation metric. The function also visualizes the top N important features using `xgb.ggplot.importance()`.

Value

A list containing:

`model` The trained XGBoost model.

`confusion_matrix`
The confusion matrix of the test set predictions.

`importance` The feature importance matrix for the top features.

`class_mapping` A named vector showing the mapping from class labels to numeric values used for training.

`cv_results` Cross-validation results, if cross-validation was performed (otherwise NULL).

`plot` A ggplot object showing the feature importance plot.

Examples

```
# Example usage:
data_df0 <- cytodata
data_df <- data.frame(data_df0[, 1:4], log2(data_df0[, -c(1:4)]))
data_df <- data_df[, -c(1, 3, 4)]
data_df <- dplyr::filter(data_df, Group != "ND")

cyt_xgb(
  data = data_df, group_col = "Group",
  nrounds = 500, max_depth = 4, eta = 0.05,
  nfold = 5, cv = TRUE, eval_metric = "mlogloss",
  early_stopping_rounds = NULL, top_n_features = 10,
  verbose = 0, plot_roc = TRUE
)
```

Index