Exercise_3

2025-01-26

Below is a sample function to perform cross-validation error to compare the linear model computed thanks to ordinary least-square and CART algorithm

```
cv_compare <- function(formula, data, k = 10, seed = NULL,
                       loss = function(y, pred) mean((y - pred)^2)) {
    # Check if rpart package is installed
    if (!requireNamespace("rpart", quietly = TRUE)) {
        stop("Package 'rpart' is required for CART. Please install it.")
    }

    # Set seed for reproducibility if provided
    if (!is.null(seed)) {
        set.seed(seed)
    }

    # Number of observations
    n <- nrow(data)
    if (k > n) {
        stop("Number of folds 'k' cannot exceed the number of observations.")
    }

    # Create folds
    folds <- sample(rep(1:k, length.out = n))

    # Initialize vectors to store errors
    ols_errors <- numeric(k)
    cart_errors <- numeric(k)

    # Perform k-fold cross-validation
    for (i in 1:k) {
        # Split data into training and test sets
        test_indices <- which(folds == i)
        train_data <- data[-test_indices, ]
        test_data <- data[test_indices, ]

        # Fit OLS model and compute error
        lm_model <- lm(formula, data = train_data)
        pred_lm <- predict(lm_model, newdata = test_data)
        y_test <- model.response(model.frame(formula, test_data))
        ols_errors[i] <- loss(y_test, pred_lm)

        # Fit CART model and compute error
        cart_model <- rpart::rpart(formula, data = train_data, method = "anova")
        pred_cart <- predict(cart_model, newdata = test_data)
```

```
        cart_errors[i] <- loss(y_test, pred_cart)
    }

    # Return results
    list(
        ols_cv_error = mean(ols_errors),
        cart_cv_error = mean(cart_errors),
        ols_errors_per_fold = ols_errors,
        cart_errors_per_fold = cart_errors
    )
}
```