

Linear Discriminant Analysis (LDA)

AIM

To improve classification performance by removing near-zero variance predictors and applying LDA on the Ionosphere dataset.

EXPERIMENTAL SETUP

- **Dataset:** Ionosphere (from mlbench)
- **Model:** Linear Discriminant Analysis (LDA)
- **Feature Selection:** Near-zero variance filter
- **Evaluation:** Accuracy using confusion matrix

LIBRARIES REQUIRED

- `caret` – Data partitioning, preprocessing, model training
- `mlbench` – To load the Ionosphere dataset

STEPS

1. Load and split Ionosphere dataset into train/test sets
2. Detect and remove near-zero variance predictors
3. Train LDA model using 5-fold cross-validation
4. Predict on test data
5. Evaluate performance using confusion matrix

CODE SNIPPET

```

library(caret)
library(mlbench)
# Load the dataset
data("Ionosphere", package = "mlbench")

# Split into training and testing sets
set.seed(123)
trainIndex <- createDataPartition(Ionosphere$Class, p = 0.8, list = FALSE)
train_data <- Ionosphere[trainIndex, ]
test_data <- Ionosphere[-trainIndex, ]

# Identify near-zero variance predictors
nzv <- nearZeroVar(train_data, saveMetrics = TRUE)

# Remove near-zero variance predictors
train_data <- train_data[, !nzv$nzv]
test_data <- test_data[, colnames(test_data) %in% colnames(train_data)]

# Train the LDA model using 5-fold CV
trained_model <- train(Class ~ ., data = train_data,
                      method = "lda",
                      trControl = trainControl(method = "cv", number = 5))

# Print model summary
trained_model

```

```

## Linear Discriminant Analysis
##
## 281 samples
## 33 predictor
## 2 classes: 'bad', 'good'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 225, 224, 225, 225, 225
## Resampling results:
##
## Accuracy Kappa
## 0.8720551 0.7045265

```

```

# Predict on test data
predictions <- predict(trained_model, newdata = test_data)

# Evaluate predictions
confusionMatrix(predictions, test_data$Class)

```

```

## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##      bad   13    0
##      good  12   45
##

```

```

##              Accuracy : 0.8286
##              95% CI   : (0.7197, 0.9082)
##      No Information Rate : 0.6429
##      P-Value [Acc > NIR] : 0.0005219
##
##              Kappa   : 0.5821
##
##      McNemar's Test P-Value : 0.0014962
##
##              Sensitivity : 0.5200
##              Specificity : 1.0000
##              Pos Pred Value : 1.0000
##              Neg Pred Value : 0.7895
##              Prevalence : 0.3571
##              Detection Rate : 0.1857
##      Detection Prevalence : 0.1857
##              Balanced Accuracy : 0.7600
##
##              'Positive' Class : bad
##

```

CONCLUSION

Filtering near-zero variance features before LDA improved model efficiency without sacrificing accuracy. The confusion matrix provides insight into classification performance.