**Student: Seif Kungulio**

**Date: 04/20/2025**

**Subject: Project 5**

**Class: DSCI 512**

**Section: 01W**

**Instructor: Dr. Nengbing Tao**

**File Name: Project6\_Kungulio\_Seif.docx**

1. Load the dataset bike.csv into memory. Convert holiday to a factor using factor() function. Then split the data into a training set containing 2/3 of the original data (test set containing remaining 1/3 of the original data).

Load the dataset bike.csv into memory

A computer screen shot of a program

AI-generated content may be incorrect.

Convert categorical variables to factor

A computer screen shot of text

AI-generated content may be incorrect.

Split the data into 2/3 training set and 1/3 testing set

A computer screen shot of a program

AI-generated content may be incorrect.

1. Build a support vector machine model.
   1. The response is holiday and the predictors are: season, workingday, casual, and registered. Please use svm() function with radial kernel and gamma=10 and cost = 100.

A computer screen shot of a program

AI-generated content may be incorrect.

The SVM model uses a radial kernel for a binary classification task (Holiday vs. Non-Holiday), with a gamma of 10 and cost of 100. These parameters increase model complexity, potentially improving fit but also raising the risk of overfitting.

The model relies on 633 support vectors (428 from one class, 205 from the other), indicating that the data may not be easily separable. This could also reflect some class imbalance.

* 1. Perform a grid search to find the best model with potential cost: 1, 10, 50, 100 and potential gamma: 1, 3, and 5 and using radial kernel and training dataset.

A screenshot of a computer program

AI-generated content may be incorrect.

Models with gamma = 1 consistently outperformed those with higher gamma values, regardless of the cost setting. Increasing the cost from 1 to 100 led to steadily lower error rates, with the best performance achieved at cost = 100 and gamma = 1. In contrast, higher gamma values caused the model to overfit, resulting in poorer generalization.

Despite a slightly higher error variance, the best model remains stable and reliable. It strikes a solid balance between accuracy and generalization, making it the ideal configuration. It’s recommended to train the final SVM using these parameters and validate it on a test set to ensure robust performance.

* 1. Print out the model results. What’s the best model parameters?

A computer screen shot of a program

AI-generated content may be incorrect.

The SVM model was designed to predict holidays using features like season, workingday, casual, and registered, employing a radial basis function (RBF) kernel to handle non-linear patterns. Through grid search tuning over various cost and gamma values, the optimal parameters were found to be cost = 100 and gamma = 1, resulting in the lowest cross-validation error.

The final model uses 645 support vectors (440 for Non-Holiday, 205 for Holiday), indicating a complex decision boundary and slight class imbalance. This imbalance may affect model predictions, favoring the majority class.

* 1. Forecast holiday using the test dataset and the best model found in c).

A blue screen with white text

AI-generated content may be incorrect.

This summary suggests a strong imbalance in predicted outcomes, with most predictions being “**Non-Holiday**”.

* 1. Get the true observations of holiday in the test dataset.

A screen shot of a computer

AI-generated content may be incorrect.

The summary suggests that most records (3532 out of 3629) are from “Non-Holiday” periods, while only a small subset of records (97) correspond to Holidays. This imbalance in the data might influence modeling results.

* 1. Compute the test error by constructing the confusion matrix. Is it a good model?

A screenshot of a computer program

AI-generated content may be incorrect.

**Test Error = 1 – Accuracy = 1 – 0.9746 = 0.0254**

The model demonstrates high overall accuracy (97.46%) and excellent sensitivity for identifying **Non-Holidays**, but it performs poorly in detecting **Holidays** due to significant class imbalance. With low specificity (7.2%), a poor Kappa score (0.1281), and low balanced accuracy (53.58%), the model is heavily biased toward the majority class. Despite appearing effective at first glance, it fails to reliably identify the minority class (Holidays), making it **unsuitable** if accurate Holiday prediction is important.