Pattern Recognition and Machine Learning

Homework 4

Faculty of Interdisciplinary Sciences & Technologies
University Of Tehran
Fall 2023

Part 1

General Homework Policies

1. Due date of this homework is on *Friday 4 Dey 02 (25 Dec. 2023)*, so you need to submit it before the due date[midnight 4 Dey], otherwise you won't get the total score! We consider the following policy for the late homework,

    - Homework is worth full credit at the beginning of class on the due date.
    - It is worth half credit for the next 48 hours.
    - It is worth zero credit after that.

2. You are welcome to collaborate, cooperate, and consult with your classmates provided that you write up the solutions independently.

3. Don't plagiarize! Write everything in your own words, and properly cite every outside source you use. Taking credit for work as well as ideas that are not your own is plagiarism. Students who plagiarize will not get any score and they will be introduced in the class.

4. Please create reference for all sources(books, papers, websites) which you use.

5. Please create a cover letter for your report which is simply the Homework#, title of the course, your name, surname, and student number.

6. You may post questions asking for clarifications and alternate perspectives on concepts on piazza or in the class.

Part 2

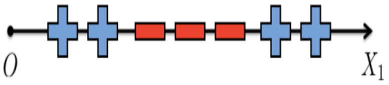
Support Vector Machine (SVM)

1 Theoretical Understanding of SVM

Answer the following questions in detail and put your answers to your final report:

1. *Handling Missing Values:* Define missing values and discuss approaches to overcome this issue. Explain how missing values impact machine learning models and the importance of handling them appropriately. (7 points)

2. *Classification on a Line(Using Kernel):* Given a set of spots on a line with positive and negative labels like below, explain how SVM can be employed for binary classification. Provide two kernel functions designed to effectively map these spots in the second dimension to separate these spots using a linear boundary. (7 points)



3. *KNN vs. SVM Comparison:* Compare KNN and SVM, focusing on boundaries, variance, and parametric/nonparametric aspects. Discuss the interpretability of each model and their suitability for different types of data. (7 points)

4. *SVM for Regression (SVR):* Briefly describe how SVM can be adapted for regression tasks using Support Vector Regression (SVR). Explain the key differences between classification and regression in the SVM context. (7 points)

5. *Hinge Loss in SVM:* Define hinge loss and explain its role in SVM. Discuss how hinge loss is used to train the SVM model and its significance in optimizing the separation between classes. (7 points)

2 Implementation of SVM for Sentiment Analysis

2.1 Introduction

In this section, You are tasked with implementing the Support Vector Machine (SVM) algorithm for sentiment analysis.

Sentiment analysis involves the use of natural language processing and machine learning techniques to determine the emotional tone or sentiment expressed in text data, such as reviews, social media posts, or comments. The goal is to categorize the text as positive, negative, or neutral, providing valuable insights into public opinion and user sentiments. Your objective is to conduct sentiment analysis specifically focused on pinpointing the challenges faced by each major U.S. airline.

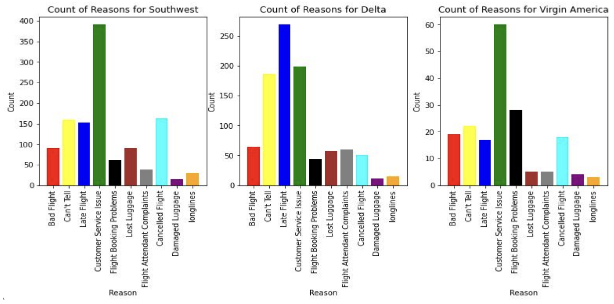
2.2 Data Set

The designated dataset involves US Airline tweets, capturing sentiments associated with each major U.S. airline. This Twitter data, collected since February 2015, underwent initial classification into positive, negative, and neutral categories. Subsequently, contributors were tasked with further categorizing negative sentiments by identifying reasons such as 'late flight' or 'rude service.'

2.3 Impelemtation Steps

Follow the outlined steps to successfully navigate through this analysis:

- *Data Loading and Preprocessing:* Start by loading the US Airline tweets dataset. For preprocessing, address missing values, and drop features with more than 90 percent missing values to enhance model efficiency. Explain the importance of these steps in ensuring the dataset's quality. (10 points)
- *Exploratory Data Analysis (EDA):*
  - Create a bar chart depicting different reasons for negative emotions about airports, categorized by each airline (like the figure below). (5 points)
  - Identify the top three reasons contributing to negative sentiments over all airlines. To do this, you can create another bar chart depicting the reasons regardless of airlines. (5 points)
  - Generate a word cloud for negative reviews to visually highlight the most frequent words. Generate another word cloud for positive reviews. (10 points)



- *Text Processing:* Begin by eliminating neutral opinions and then remove stop words from the dataset. Create a separate copy of the data without stop words. Use the SVM model (for the last part of this section) to compare the impact of this processing on classification accuracy. (10 points)

- *Data Splitting:* Split the data into training and testing sections, allocating 80% to training and 20% to testing. Additionally, create a validation subset within the training data to facilitate further cross-validation. (5 points)

- *SVM Model Creation, Grid Search and Cross-Validation:* Implement SVM models using Radial Basis Function (RBF) with a value of gamma in ['auto', 'scale', 3/n, 6/n], n = the number of samples, Polynomial with a degree in [1, 3, 5], and Linear kernel with the value of c in [0.01, 1, 100]. (you don't need to build separate SVM models.)

  Perform grid search with 5-fold cross-validation to identify the optimal parameters for each SVM model. Report the best model based on accuracy and present the confusion matrix for comprehensive evaluation. Discuss the models' performance, highlighting any signs of overfitting and providing a thorough comparison of results. (20 points)

Good Luck!