GTSC2143 Machine Learning for Business

Tutorial 6

Please write down your answers in this document and submit it at iSpace by the end of this tutorial.

Activity 1. Data Loading and Preprocessing

1. Load and Explore the Dataset
   a) Load the Amazon baby product reviews dataset using pandas:

```
import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.feature_extraction.text import CountVectorizer

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, classification_report, confusion_matrix


# Load the Amazon baby product reviews dataset

data = pd.read_csv("GTSC2143-Lecture 6_analyzing-product-sentiment-assignment_amazon_baby.csv",index_col=0)
```

   b) Check basic information:
   - Dataset shape
   - Column names
   - Check any missing value
   - Drop records with missing value

2. Create Sentiment Labels
   a) Create a new column called 'positive' where the value is 1 if the rating is greater than 3, and 0 otherwise
   b) Display the distribution of sentiment labels

Activity 2. Data Splitting and Text Processing

1. Train/Test Split
   a) Split the data into training (80%) and testing (20%) sets using `random_state=42`
   b) Display the shapes of training and testing sets

2. Convert Text to Features
   a) Use `CountVectorizer` to convert review text into word count features
   b) Set `max_features=1000` to limit vocabulary size
   c) Fit the vectorizer on training data and transform both training and test texts
   d) Display the shape of the feature matrices
   e) Analysis: Write 2-3 sentences explaining how text becomes numerical features.

Activity 3. Model Training and Evaluation

1. Logistic Regression Model
   a) Train a logistic regression classifier using the word count features

b) Use `random_state=42` for reproducible results

2. Evaluate the Model
   a) Make predictions on the test set
   b) Calculate and display:
      - Accuracy score
      - Classification report
      - Confusion matrix
   c) Analysis: Write 2-3 sentences interpreting the model's performance.
3. Feature Analysis
   a) Display the top 10 most positive words (highest coefficients)
   b) Display the top 10 most negative words (lowest coefficients)
   c) Analysis: Write 2-3 sentences about which words drive sentiment predictions.

- End of Tutorial 6 -