Logistic Regression

April 28, 2025

[ ]:
```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression

# Load the dataset
df = pd.read_csv("/Users/zhengfeibian/Desktop/5630final/MyOwnChooseDataSets/
 ↪Datasets_SYSEN5630FinalProject_Zhengfei/Symptom_Forgetfulness_Datasets/
 ↪alzheimers_abstracts__symptoms.csv")  #

# Randomly sample 20 abstracts, using a fixed random seed for reproducibility
sample_df = df.sample(n=20, random_state=42)
sample_texts = sample_df["Abstract"].tolist() # Extract the abstract texts
sample_labels = sample_df["MentionsSymptom"].tolist() # Extract the␣
 ↪corresponding labels (0 or 1)

# Split the data into a training set and a temporary set (for dev/test)
from sklearn.model_selection import train_test_split
train_df, temp_df = train_test_split(df, test_size=0.3, random_state=42,␣
 ↪stratify=df["MentionsSymptom"])

X_train = train_df["Abstract"] # Training features (abstract texts)
y_train = train_df["MentionsSymptom"] # Training labels (mention symptom or not)

# Convert the text into TF-IDF features, keeping only the top 5000 frequent␣
 ↪words
vectorizer = TfidfVectorizer(max_features=5000)
X_train_vec = vectorizer.fit_transform(X_train)

# Initialize and fit the model, allowing up to 200 iterations for convergence
clf = LogisticRegression(max_iter=200)
clf.fit(X_train_vec, y_train)

# Vectorize the sampled abstracts using the trained TF-IDF vectorizer
X_sample_vec = vectorizer.transform(sample_texts)

# Predict the labels for the sampled abstracts
sample_preds = clf.predict(X_sample_vec)
```

```
# Merge the original abstracts, true labels, and predicted labels into a single␣
↪DataFrame
results_df = pd.DataFrame({
    "Abstract Snippet": sample_texts,
    "True Label (MentionsSymptom)": sample_labels,
    "Logistic Regression Prediction": sample_preds
})

# Display the resulting DataFrame
pd.set_option('display.max_colwidth', 200)
print(results_df)
```

```
                                                Abstract Snippet  \
0   Macrophages accumulate lipid droplets (LDs) under stress and inflammatory conditions. Despite the presence of LD-loaded macrophages in many tissues, including the brain, their contribution to neur…
1   The endoplasmic reticulum (ER) plays a fundamental role in maintaining cellular homeostasis by ensuring proper protein folding, lipid metabolism, and calcium regulation. However, disruptions to ER…
2   C-truncating variants in the charged multivesicular body protein 2B (CHMP2B) gene are a rare cause of frontotemporal lobar degeneration (FTLD), previously identified only in Denmark, Belgium, and …
3   Sex differences in patterns of cortical thickness and neuropsychiatric symptom (NPS) burden were examined among individuals with Alzheimer's disease (AD) and two copies (homozygote carriers) of th…
4   Nanotechnology has significantly impacted drug discovery and development over the past three decades, offering novel insights and expanded treatment options. Key to this field is nanoparticles, ra…
5   Long-term potentiation (LTP) and long-term depression (LTD) are widely used to study synaptic plasticity. However, whether proteins regulating LTP and LTD are altered in cognitive disorders and co…
6   The aggregation of  -amyloid (A ) peptides has been associated with the onset of Alzheimer's disease (AD) by causing neurotoxicity due to oxidative stress and apoptosis. Cordycepin is a natural de…
7   As a transmembrane protein, DPP6 modulates the function and properties of ion channels, playing a crucial role in various tissues, particularly in the brain. DPP6 interacts with potassium channel …
8   Tau aggregation in early affected regions in the asymptomatic stage of Alzheimer's disease marks a transitional phase between stable asymptomatic amyloid positivity and the clinically manifest sta…
9   Mild Cognitive Impairment (MCI) is marked by a measurable decline in cognitive function that exceeds typical age-related changes but does not yet qualify as dementia. The brain's Default Mode Netw…
10  The metabolic syndrome or syndrome X is a clustering of different components counting insulin resistance (IR), glucose intolerance, visceral obesity, hypertension and dyslipidemia. It has been sho…
11  Global life expectancy has steadily increased in recent decades, resulting

    True Label (MentionsSymptom)  Logistic Regression Prediction
0                              0                               0
1                              0                               0
2                              0                               0
3                              0                               0
4                              0                               0
5                              0                               0
6                              0                               0
7                              0                               0
8                              0                               0
9                              0                               0
10                             0                               0
11                             0                               0
```

in a significant rise in the number of individuals aged 80 years and older. This trend is also evident in Latin America, …
12 Syphilis, caused by Treponema pallidum, presents a diagnostic challenge due to its diverse clinical manifestations. Neurosyphilis has seen a resurgence in recent years, particularly among m…
13 BackgroundThe concepts of 'personalized medicine' and 'patient-orchestrated care' in Alzheimer's disease (AD) lack standard conceptualization, which presents challenges for collabora…
14 The apolipoprotein E (APOE) gene's APOE4 variant is frequently associated with an elevated risk of Alzheimer's disease, while APOE3 isoform is found in normal individuals. Both the isoforms differ…
15 Recent evidence suggests that Alzheimer's amyloid-beta (1-40) (A 1-40), an emerging biomarker of cardiovascular disease, may be involved in the heart-brain-renal axis. We aimed to comprehensively …
16 The role of glaucoma in predicting Alzheimer's disease (AD) factors is unknown. This current meta-analysis was aimed at evaluating the risk of AD events in individuals suffering from glaucoma base…
17 Patients with bipolar disorder (BD) are at increased risk of dementia. The underlying mechanisms are debated. FDG-PET elucidates glucose metabolic reductions due to altered neuronal activity in th…
18 Alphaherpesviruses, including herpes simplex virus type 1 (HSV-1), pseudorabies virus (PRV), and bovine herpesvirus type 1 (BoHV-1), are significant pathogens affecting humans and animals. These v…
19 Preventing dementia and Alzheimer's disease (AD) is a global priority. Multimodal interventions targeting several risk factors and disease mechanisms simultaneously are currently being tested worl…

	True Label (MentionsSymptom)	Logistic Regression Prediction
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	1	1
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	1	0