Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Projection

Lab 11

Load dataset

Dataset: Twitter Sentiment Dataset (https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset)

The dataset has three sentiments namely, negative(-1), neutral(0), and positive(+1). It contains two fields for the tweet and label.

- Import libraries

```
import pandas as pd
import numpy as np
import nltk
from scipy.sparse import csr_matrix
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report
from sklearn.svm import SVC
nltk.download('punkt')
```

- Load the twitter sentiment dataset

```
data = pd.read_csv('Twitter_Data.csv')
data.head()
```

Prepare dataset

- Convert data in clean_text column to list

```
X = data["clean_text"].values.tolist()
```

- Prepare labels

```
y = data["category"].values.tolist()
```

- Split dataset into training and test dataset

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.3, random_state=42)
```

Extract TF-IDF Features

- Extract TF-IDF feature vectors

```
cv = TfidfVectorizer(stop_words =”english”, min_df=0.25)
X_fv_train = cv.fit_transform(X_train).todense()  # fit the cv using the training dataset and
transform the                                                    training data
X_fv_test = cv.transform(X_test).todense()     # apply the cv to the test dataset
```

- Scale the features with z-score standardization

```
scaler = StandardScaler()
scaler.fit(X_fv_train)
X_fv_train = scaler.transform(X_fv_train)
X_fv_test = scaler.transform(X_fv_test)
```

- Change the data matrix to sparse matrix

```
X_fv_train = csr_matrix(X_fv_train)
X_fv_test = csr_matrix(X_fv_test)
```

Dimensionality Reduction using SVD

```
from sklearn.decomposition import TruncatedSVD
```

- Decompose the training data matrix using SVD

```
svd = TruncatedSVD(n_components=20, n_iter=7, random_state=42)
svd.fit(X_fv_train)
X_svd_train = svd.transform(X_fv_train)
X_svd_test = svd.transform(X_fv_test)
```

- Train a SVM classifier

```
svm = SVC(gamma='auto’)
svm.fit(X_svd_train,y_train)
```

- Predict the test data

```
y_predict = svm.predict(X_svd_test)
```

- Evaluate the performance

```
print(classification_report(y_test, y_predict))
```

Dimensionality Reduction using PCA

```
from sklearn.decomposition import PCA
```

- Find PC space and transform the data into the space

```
pca = PCA(n_components=20, svd_solver == 'arpack')
pca.fit(X_fv_train)
X_pca_train = pca.transform(X_fv_train)
X_pca_test = pca.transform(X_fv_test)
```

- Train a SVM classifier

```
svm = SVC(gamma='auto’)
svm.fit(X_pca_train,y_train)
```

- Predict the test data

```
y_predict = svm.predict(X_pca_test)
```

- Evaluate the performance

```
print(classification_report(y_test, y_predict))
```

Dimensionality Reduction using LDA

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
```

- Perform LDA

```
pca = PCA(n_components=20, svd_solver == 'arpack')
pca.fit(X_fv_train)
X_pca_train = pca.transform(X_fv_train)
X_pca_test = pca.transform(X_fv_test)
```

- Train a SVM classifier

```
svm = SVC(gamma='auto’)
svm.fit(X_pca_train,y_train)
```

- Predict the test data

```
y_predict = svm.predict(X_pca_test)
```

- Evaluate the performance

```
print(classification_report(y_test, y_predict))
```

Your work!

1. Load the SMS Spam Collection Dataset from https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
2. Split the dataset into training and test dataset (Hint: use *train_test_split* method in sklearn library)
3. Extract feature vectors of samples on both training and test sets
4. Transform the feature vectors into lower dimensional space.
5. Construct a classifier using the training samples for identifying the class of sms (i.e., spam or ham)
6. Evaluate performance of the classifier on the test set,
7. Submit your program to the assignment submission system (http://hw.cs.science.cmu.ac.th/).

Note:

- Put your name and student ID in the first cell using comment tag.
- Name your python notebook file with the pattern Lab_11_XXXXXXXXX.py (XXXXXXXXX is your student ID)

References & Study Resources

- https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection
- https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- https://scikit-learn.org/stable/modules/ensemble.html