Classifying fake news using supervised learning with NLP

Katharine Jarmul
Founder, kjamistan
What is supervised learning?

- Form of machine learning
  - Problem has predefined training data
  - This data has a label (or outcome) you want the model to learn
  - Classification problem
  - Goal: Make good hypotheses about the species based on geometric features

<table>
<thead>
<tr>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Petal Width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>I. setosa</td>
</tr>
<tr>
<td>7.0</td>
<td>3.2</td>
<td>4.77</td>
<td>1.4</td>
<td>I. versicolor</td>
</tr>
<tr>
<td>6.3</td>
<td>3.3</td>
<td>6.0</td>
<td>2.5</td>
<td>I. virginica</td>
</tr>
</tbody>
</table>
Supervised learning with NLP

- Need to use language instead of geometric features
- scikit-learn: Powerful open-source library
- How to create supervised learning data from text?
  - Use bag-of-words models or tf-idf as features
IMDB Movie Dataset

<table>
<thead>
<tr>
<th>Plot</th>
<th>Sci-Fi</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>In a post-apocalyptic world in human decay, a ...</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mohei is a wandering swordsman. He arrives in ...</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#137 is a SCI/FI thriller about a girl, Marla,...</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Goal: Predict movie genre based on plot summary
- Categorical features generated using preprocessing
Supervised learning steps

- Collect and preprocess our data
- Determine a label (Example: Movie genre)
- Split data into training and test sets
- Extract features from the text to help predict the label
  - Bag-of-words vector built into scikit-learn
- Evaluate trained model using the test set
Let's practice!
Building word count vectors with scikit-learn
Predicting movie genre

- Dataset consisting of movie plots and corresponding genre
- Goal: Create bag-of-word vectors for the movie plots
  - Can we predict genre based on the words used in the plot summary?
Count Vectorizer with Python

In [1]: import pandas as pd

In [2]: from sklearn.model_selection import train_test_split

In [3]: from sklearn.feature_extraction.text import CountVectorizer

In [4]: df = ... # Load data into DataFrame

In [5]: y = df['Sci-Fi']

In [6]: X_train, X_test, y_train, y_test = train_test_split(df['plot'], y, test_size=0.33, random_state=53)

In [7]: count_vectorizer = CountVectorizer(stop_words='english')

In [8]: count_train = count_vectorizer.fit_transform(X_train.values)

In [9]: count_test = count_vectorizer.transform(X_test.values)
Let's practice!
Training and testing a classification model with scikit-learn
Naive Bayes classifier

- Naive Bayes Model
  - Commonly used for testing NLP classification problems
  - Basis in probability
- Given a particular piece of data, how likely is a particular outcome?
- Examples:
  - If the plot has a spaceship, how likely is it to be sci-fi?
  - Given a spaceship **and** an alien, how likely **now** is it sci-fi?
- Each word from CountVectorizer acts as a feature
- Naive Bayes: Simple and effective
Naive Bayes with scikit-learn

In [10]: from sklearn.naive_bayes import MultinomialNB
In [11]: from sklearn import metrics
In [12]: nb_classifier = MultinomialNB()
In [13]: nb_classifier.fit(count_train, y_train)
In [14]: pred = nb_classifier.predict(count_test)
In [15]: metrics.accuracy_score(y_test, pred)
Out [15]: 0.85841849389820424
## Confusion Matrix

```python
In [16]: metrics.confusion_matrix(y_test, pred, labels=[0,1])
Out [16]:
array([[6410, 563],
       [ 864, 2242]])
```

<table>
<thead>
<tr>
<th></th>
<th>Action</th>
<th>Sci-Fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>6410</td>
<td>563</td>
</tr>
<tr>
<td>Sci-Fi</td>
<td>864</td>
<td>2242</td>
</tr>
</tbody>
</table>
Let's practice!
Simple NLP, Complex Problems

Katharine Jarmul
Founder, kjamistan
Translation

god bless the german language

(source: https://twitter.com/Lupintweets/status/865533182455685121)
Sentiment Analysis

Ex. contexts in r/sports

- “big men are very soft”
- “freakin raging animal”
- “went from the ladies tees”
- “two dogs fighting”
- “being able to hit”
- “insanely difficult saves”
- “amazing shot”
- “he is still crazy good”
- “his stats are insane”

Ex. contexts in r/TwoX

- “some soft pajamas”
- “stuffed animal”
- “lovely ladies”
- “hiking with the dogs”
- “it didn’t really hit me”
- “a difficult time”
- “totally shot me down”
- “overreacting crazy woman”
- “people are just insane”

(source: https://nlp.stanford.edu/projects/socialsent/)
Language Biases

She's a professor. He's a babysitter.

O bir profesör. O bir bebek bakıcısı.

(related talk: https://www.youtube.com/watch?v=j7FwpZB1hWc)
Let's practice!