Comparison of Machine Learning Algorithms to Build a Predictive Model for Classification of Survey Write-in Responses

Dr. Andrea Roberson
Justin Nguyen

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.
Annual Capital Expenditures Survey (ACES) Research-Overview

- Provides national-level estimates of annual capital investment in new and used buildings, structures, machinery, and equipment by U.S. non-farm businesses

Source: https://www.census.gov/programs-surveys/aces.html
Purpose of the Research

- Prior U.S. Census Bureau studies identified areas of improvement in our editing processes in order to improve the timeliness and quality of our estimates while reducing cost.

- A U.S. Census Bureau Economic Edit Reduction team identified edits and processes that can be automated.

- Suggestions included automating the manual examination of ACES survey write-ins.

- The use of Machine Learning (ML) classifiers was recommended to successfully predict the correct class of capital expenditures.
What is Machine Learning (ML)
Modernization of Statistical Production

- National Statistics Offices should all explore the use of ML (Chu and Poirier, 2015)

- Applications
  - Decision Trees (Portugal): Detection of errors in foreign trade transaction data.
    - Reduced manual examination of records
ML Techniques for Write-In Classification

- Logistic regression [statistics]
  - Training data: Binary response (0:1) and predictors
  - Maximum likelihood leads to model parameters
  - Resulting model is used to predict responses

- Support Vector Machines [non-statistics]
  - Training data: Binary response (0:1) and predictors
  - Hyperplanes in the space of predictors separate responses
  - SVM optimization problem comes from geometry
Data

- 2015 and 2016 ACES Write-in Data

Classification Breakdown

Source: U.S. Census Bureau, 2015 and 2016 Annual Capital Expenditures Survey
Text Classification Overview

▪ Bag of words model

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

The first **Bass** I remember catching was when I was six, my Dad had a performance of Strange Brew by **Cream** on the boat. Recollections indeed: an extremely fashionable **Cream**, with serious sideburns all round. I remember wondering why bassist Bruce was playing a **guitar**. Lost in my thoughts, it happened, there was a beautiful **Bass** hooked to my fishing rod.

When weekend **fishermen** looking to unwind come to Rock Harbor, they take out their rods and reels and angle for striped **bass** one by one. When **commercial fishermen** like Mike Abdow go out on their boats to earn a living, they catch **bass** the same way. But right now, the waters are rough between people who fish **bass** for a living and those who angle for pleasure. Big-**money** sporting interests are trying to stop small-time **commercial fishermen** from pulling in any morestriped **bass**.

**Bank** Negara will change the way it calculates the cost of lending **money** to **commercial** banks and financial institutions, the central **bank** said in a release. In a statement, the **bank** said that as of Nov. 1, the base lending rate, at which **commercial** banks can borrow from the central **bank**, will become more responsive to movements in **money**-market rates. Several weeks ago, the **bank** said it would change the way it calculates the base rate.
**Term-Document Example**

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Write-in 1</th>
<th>Write-in 2</th>
<th>Write-in 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>bass</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>commercial</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>cream</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>guitar</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fishermen</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>money</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CLASS</td>
<td>Equipment</td>
<td>Equipment</td>
<td>Structures</td>
</tr>
</tbody>
</table>

We count the number of times each word is used in each of our write-ins.
Documents in Term Space

Visualizing the word vectors as points in three dimensional space
A.I. Experiments: Visualizing High Dimensional Space

Source: https://experiments.withgoogle.com/ai
Methodology

Prepare the Data

t-SNE visualization of test data

- Equipment
- Not Applicable
- Structures
Logistic Regression (Fit the Model)

- **Grid Search- LR**

```python
import sklearn
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV

pipeline = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', LogisticRegression())])

parameters = {
    'vect__ngram_range': ((1, 1), (1,2), (1, 3)),  # unigrams or bigrams or trigrams
    'clf__penalty': ('l2', 'l1'),
    'clf__C': (1, 10, 12, 14, 15, 20, 25, 30, 35, 40, 45, 50, 55, 70, 75, 100)
}

grid_search = GridSearchCV(pipeline, parameters, n_jobs=-1, verbose=1, refit=True, cv=15, scoring='accuracy')
```
The entries in the confusion matrix have the following meaning in the context of the study:

- $Ee$ is the number of **correct** predictions that a write-in is Equipment.
- $Es$ is the number of **incorrect** predictions that a write-in is a Structure, when in fact it is Equipment.
- $En$ is the number of **incorrect** predictions that a write-in is Not Applicable, when in fact it is Equipment.
Statistics Used to Evaluate Performance

Correct Classification Rate

\[ CCR = \frac{Ee + Nn + Ss}{E.+N.+S.} \]

False Discovery Rate

\[ FDR = \frac{Es + En}{En + Nn + Sn + Es + Ns + Ss} \]

Specificity

\[ Specificity = \frac{Ee}{Ee + Es + En} \]

Sensitivity

\[ Sensitivity = \frac{Nn + Ss}{N.+S.} \]
## Results

The performance statistics for the compared methods on the test data.

<table>
<thead>
<tr>
<th>Model</th>
<th>SVMs</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR</td>
<td>0.9789</td>
<td>0.9794</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0076</td>
<td>0.0076</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.9978</td>
<td>0.9978</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.9146</td>
<td>0.9171</td>
</tr>
</tbody>
</table>
Pros and Cons of SVM

Pros

▪ Can deal with very high dimensional data.
▪ SVMs work very well in practice, even with very small training sets

Cons

▪ Non-Probabilistic: SVMs do not directly provide probability estimates
Pros and Cons of LR

Pros

▪ Wide spread industry comfort for logistic regression solutions trusted
▪ Convenient probability scores
▪ Quick to train

Cons

▪ Logistic regression tends to underperform when there are non-linear decision boundaries.
Conclusions

- LR had a slightly higher Correct Classification rate than SVM.
- LR achieved the highest sensitivity.
- LR was the overall best performing method.
- Recommend ACES staff deploy a LR model into a production system.
References


Questions?

Contact information:

Dr. Andrea Roberson

andrea.roberson@census.gov

Justin D. Nguyen

justin.d.nguyen@census.gov

Thank you for your attendance and attention!