Using Deep Learning to Assist in Coding Qualitative Data

This research was supported by grant T15LM011271
About Me

- Upcoming 4th year at UCSD
- Majoring in Computer Science with Bioinformatics
- Desire to learn more about biomedical informatics
- Research Interest in:
  - Precision Medicine
  - Natural Language Processing
  - Machine Translation
Outline

• Deep Learning
• Word2vec
• Metrics
• Model Architecture
• Internal Validation Performance– DARPA Response
• Calibration
• External Validation Performance– FDA Comments
Deep Learning

ARTIFICIAL INTELLIGENCE
Engineering of making Intelligent Machines and Programs

MACHINE LEARNING
Ability to learn without being explicitly programmed

DEEP LEARNING
Learning based on Neural Network


This research was supported by grant T15LM011271
What Is A Neural Network
What Is A Neural Network

Layer 1

Layer 2

Neuron

Weighted Connections
Activation Function

\[ R(x) = \max(0, x) \]

Bias

\[ R(-3.72) = \max(0, -3.72) \]

Diagram:

- Input 1: 2
  - Multiplied by 10: 20
  - Result: 63.284

- Input 2: 10
  - Multiplied by -1.15: -11.5
  - Result: -80.72

Bias: 5

Output: 0
Word2Vec

It is raining cats and dogs.

<table>
<thead>
<tr>
<th>Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>cats</td>
</tr>
</tbody>
</table>

is raining and dogs

![Diagram with input, projection, and output layers showing the process of Word2Vec.](image)
Japan – Tokyo + Russia = Moscow
How Much Data Is Enough

More Data + Bigger Models

Accuracy

1990s

Scale (data size, model size)

neural networks

other approaches

This research was supported by grant T15LM011271
Important Metric

<table>
<thead>
<tr>
<th>Predicted Results</th>
<th>Labeled Disease</th>
<th>Labeled Not Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>a: True positive</td>
<td>b: False positive</td>
</tr>
<tr>
<td>Negative</td>
<td>c: False negative</td>
<td>d: True negative</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{True\ Positives}{True\ Positives + False\ Negatives} \)

Specificity = \( \frac{True\ Negatives}{True\ Negatives + False\ Positives} \)
Important Metric

High Sensitivity
Few False Negatives (blue)

Low Specificity
Many False Positives (red)
Methods

• Goal: Correctly Labeling Disease vs Not Disease
• Models:
  • Convolutional Neural Network (CNN)
  • Log. Reg. with TF-IDF (Justin)
  • Log. Reg. with Word Embedding
• Internally validated using bootstrapping with DARPA responses
• Recalibrated
• Externally validated with FDA comments
• Primary Metrics: AUC ROC, Sensitivity, Specificity
Result

Model Comparison

AUC Score

Neural Network

Log. Reg + TF-IDF

Log. Reg + Word Embedding

This research was supported by grant T15LM011271
DARPA-Calibration Plot (Reliability Curve)

Fraction of Positives

Mean Predicted Value

- Perfectly calibrated
- Original
- Logistic Regression
- Isotonic Regression
Comparsion of Performance at Different Thresholds

<table>
<thead>
<tr>
<th>Predicted Label</th>
<th>True Label</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>O (Youden’s Point)</strong></td>
<td>Disease: 78, Not Disease: 23</td>
</tr>
<tr>
<td></td>
<td>Disease: 23, Not Disease: 77</td>
</tr>
<tr>
<td>Sensitivity = 0.77</td>
<td>Specificity = 0.77</td>
</tr>
<tr>
<td><strong>L (Low Sensitive)</strong></td>
<td>Disease: 71, Not Disease: 17</td>
</tr>
<tr>
<td></td>
<td>Disease: 30, Not Disease: 83</td>
</tr>
<tr>
<td>Sensitivity = 0.70</td>
<td>Specificity = 0.83</td>
</tr>
<tr>
<td><strong>M (Medium Sensitive)</strong></td>
<td>Disease: 85, Not Disease: 30</td>
</tr>
<tr>
<td></td>
<td>Disease: 16, Not Disease: 70</td>
</tr>
<tr>
<td>Sensitivity = 0.84</td>
<td>Specificity = 0.70</td>
</tr>
<tr>
<td><strong>H (High Sensitive)</strong></td>
<td>Disease: 89, Not Disease: 42</td>
</tr>
<tr>
<td></td>
<td>Disease: 12, Not Disease: 58</td>
</tr>
<tr>
<td>Sensitivity = 0.88</td>
<td>Specificity = 0.58</td>
</tr>
</tbody>
</table>

This research was supported by grant T15LM011271
Key Takeaways

• No Literary Works On Applying Deep Learning on Qualitative Data

• Newer ≠ Better
  • Word Embedding

• Different Situation = Different Thresholds
Lessons Learned

• Process of qualitative coding and analysis
• Neural Networks
• Word Embedding
• Calibration and Recalibration
• How to read machine learning papers
Acknowledgements

- Dr. Jejo Koola
- Dr. Cinnamon Bloss
- Dr. Cynthia Schairer
- Justin Castro
- The Bloss Lab