Machine learning in Office: exploring opportunities

Bamdev Mishra
Office India Intelligence Team
Microsoft
Team

- Puru
- Atishay
- Bamdev
- Mayank
- Murty
- Pratik
- Ravi
- Rohit
- Venkat
Outline

• What is machine learning? What applications could we consider in Office

• Case study of extracting information from Kaizala

• Learnings
Machine learning and applications in Office
Machine Learning is...

“The ability of a machine (model) to improve its performance based on previous results.”

Machine learning is about predicting the future based on the past.

past

Training Data

learn

model/predictor

future

Testing Data

predict

model/predictor
Machine Learning techniques broadly refer:

- Data mining: machine learning applied to databases
- Pattern recognition and computer vision
- Neural networks, Deep learning for feature representation
- Natural language processing
- Bayesian approaches
- Dimensionality reduction
- Multidimensional analysis
- Statistics
- Optimization techniques: convex, nonconvex, analytical approaches
Two fundamental ways of learning from data: supervised and unsupervised
Supervised learning: given labeled examples.
Supervised learning

Supervised learning: given labeled examples
Supervised learning

Supervised learning: learn to predict new example
Supervised learning: classification

- apple
- apple
- banana
- banana

Classification: a finite set of labels

Supervised learning: given labeled examples
Supervised learning: regression

-4.5

10.1

3.2

4.3

Regression: label is real-valued

Supervised learning: given labeled examples
Regression example: curve fitting

Price of a used car

$x$: car attributes (e.g. mileage)

$y$: price

Can be linear/nonlinear fit to data

$y = wx + w_0$
Supervised learning: ranking

Ranking: label is a ranking

Supervised learning: given labeled examples
Ranking example

Given a query and a set of web pages, rank them according to relevance

Search over documents

Search over chats
Unsupervised learning

Unsupervised learning: given data, i.e. examples, but no labels
Ongoing ML applications in our team

• Mobile document vs Image classification: Office Lens.

• Recognizing digits: entities from cards a part of Kaizala me hub.

• Automatic layout analysis in document images.

• Sensitive / non-sensitive documents/emails.

• Detecting anomalies in IPE.

• Semantic search for Kaizala.
Ongoing machine learning applications in our team - prospective

• Understanding user preference options, e.g., using Visio or PPT.

• Automatic layout detection in Visio.

• Generating recommendations in Kaizala.

• Process mining: understanding flow layouts from data streams.

• Suggesting action cards in Kaizala.
Other machine learning efforts in Office

• My analytics and work place analytics. Understanding productive hours. Generating insights.

• Augmentation loop: a ML framework for different usecases. Smart replies, ...

• There are other ML efforts as well.

• Substrate ML. substrate.ai.

• OML tool

• Office NL tool (natural language extractor)
Case study of extracting information from Kaizala
Kaizala Insights: Analyzing Group Chat Messages in Kaizala

Do **YOU** get *overwhelmed* by **100s** of new messages in a Large Chatty Group?

- Residents of MHV
  - Priyanka: Milk is the only natural Su...
- CaptainAltcoin
  - https://goo.gl/NK49Lv
- HydChat
  - Member joined the group via QR code
Please share vision and dental claim forms
There is no different form for vision
You can download it from HRweb

Any good nursery to buy plants

Do we get discount on Microsoft products?

Kerala needs our help
Today’s snacks
Amex card process
Models and Datasets

• Conversation threads:-
  ➢ Model:- Time Clustering
  ➢ Datasets:- HydChat, DogfoodProd, SBI group, Football world cup group

• Group Sentiment:-
  ➢ Model:- SVM with tf-idf word count features
  ➢ Dataset:- IMDB movie reviews

• Abuse:-
  ➢ Model:- Logistic regression with tf-idf word count features
  ➢ Datasets:- Hate speech and offensive language dataset
KISS: Kaizala Integrated Semantic Search

Search across Kaizala messages and find the most relevant conversation(s) for a user’s query.

Outperforms existing keyword-based search as it considers semantics not spelling.

Results are linked to actual conversation threads to give full context of a message.
Machine learning tech used

- Uses GloVe word vectors to extract 300-dimensional vector for every token in a sentence
- Uses wordnet information content metric as weight vector to give higher weight to important words
- To get sentence embedding, computes weighted average of all tokens in a sentence
- To find relevancy, finds cosine distance between query vector and corpus messages’ vectors
Search Radius

![Diagram showing search radius progression with layers: Web results, Organization, Chats, Bing, SharePoint, Kaizala Messages.](image-url)
Datasets used

• MSR Paraphrase dataset to compare different algorithms to compute similarity metric, we used ROC curve’s area of each algorithm as the main metric.

• HydChat Kaizala group messages.

• SBI Knowledge base group’s messages.
Results: standard embeddings perform good

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Atleast 1 Correct</th>
<th>Zero in More Than 2 Answers</th>
<th>Good Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>azure</td>
<td>0.614107884</td>
<td>0.708609272</td>
<td>608</td>
</tr>
<tr>
<td>dssm</td>
<td>0.456431535</td>
<td>0.900662252</td>
<td>360</td>
</tr>
<tr>
<td>elmo</td>
<td>0.701244813</td>
<td>0.761589404</td>
<td>634</td>
</tr>
<tr>
<td>infersent</td>
<td><strong>0.7593361</strong></td>
<td><strong>0.503311258</strong></td>
<td>823</td>
</tr>
<tr>
<td>azure_clean</td>
<td>0.630705394</td>
<td>0.708609272</td>
<td>627</td>
</tr>
<tr>
<td>use</td>
<td>0.705394191</td>
<td>0.801324503</td>
<td>665</td>
</tr>
<tr>
<td>glove</td>
<td>0.742738589</td>
<td>0.662251656</td>
<td>723</td>
</tr>
<tr>
<td>current</td>
<td>0.730290456</td>
<td>0.761589404</td>
<td>696</td>
</tr>
</tbody>
</table>
Learnings
Learnings from our projects

• Defining the problem statement precisely and properly.

• Data is important. Collecting compliant data more important.

• Designing good metric. business metrics vs machine learning metrics.

• Modeling plays a key role. Tools play a bigger role. Develop tools.

• End-to-end infra for collecting feedback to improve the model.

• Rules are needed, but on top of machine learning models.
Feedback is important

Training Data \(\rightarrow\) model/predictor \(\rightarrow\) Testing Data

past \hspace{1cm} learn \hspace{1cm} future

Feedback
Productionization of a machine learning model takes time

Data Acquisition

Data Cleansing & Data Normalization

Model Training

Model Validation

Inferences & Model deployment

Where is rest 25%?
Most important: 25% to define the problem statement precisely
We looked at the two underlying algorithms with the best performance in the ensemble: Matrix Factorization (which the community generally called SVD, Singular Value Decomposition) and Restricted Boltzmann Machines (RBM). SVD by itself provided a 0.8914 RMSE, while RBM alone provided a competitive but slightly worse 0.8990 RMSE. A linear blend of these two reduced the error to 0.88...
Acknowledgements

• Some slides are inspired from several slides on the web.

• Thanks to David Kauchak’CS 451 – Fall 2013 slides on intro to machine learning.

• Thanks to the team members for sharing pointers.
Join the Kaizala group ‘Data Science News’ for regular data science news. It is discoverable with the nearby feature of Kaizala.

https://join.kaiza.la/g/xlesg_uZTt2qspFMtSITwA
Machine learning in Office: exploring opportunities

Bamdev Mishra
Office India Intelligence Team