Methodology

- Data collection:
  - Used java SAX to parse through dataset
  - found 28010 machine learning posts with machine learning tag
- entire analysis in this paper is based on SOTorrent

Topic Modeling using LDA:

- LDA is a method of topic modelling
- java based package for topic modelling "Mallet"
- Requires set of documents and parameters α, β, and # of topics (k)
- Used results from data collections stage
- topic word matrix displayed top words
- document word matrix displaying described topic weights for each input document
- LDA dont have a label but are rather probabilistic distribution of words only

2022-05-11

Topic Categorization 2022-05-10

- LDA was used to identify the topics of machine learning post (28,010)
- Tried categorization with 20 topics and 50 topics ($k = 20$ & $k = 50$)
- Set α & β to 0.01, 0.25, 1.00 for each k value
- when $\alpha/\beta = 0.01$: - hard to identify leading topics
  - important topics remain hidden
  - many topics get associated with document when they arent related

- When α/β = customed
- Mallet returned topics that are close to actual topics
- Assigned labels to topics based on multiple author's consensus
- LDA does not label topics
- Hindel et al: labeling some topics might not be possible
- grouped topic under 4 broad categories (44 topics)

Manual Exploration of LDA
- For validity, randomly sampled 230 Machine learning posts and manually read them
- authors read questions, answer, and comment from each post
- verified if LDA predicts right topic for the post

Evaluation and Results

Research Question 1
- Code Errors topic is most dominant
- developers are trying to adopt new machine learning tools without enough understanding
- Labeling, algorithms, traing datasets, Neural networks are higher ranking topics
- The category groups are Framework, Implementation, Sub-domain, Algorithm
- Framework category includes posts relevant to machine learning frameworks
- Examples of Frameworks Numpy, Panda, ScikitLearn, keras, Caret, Google-Cloud, azure-clad
- Implementation has 51% of topics

- Not many topics fall into Subdomain and Algorithm category
- Example of subdomain: Neural network, Image Processing and Sentiment Analysis
- Examples of Algorithms are ~~Neural Networks, Image Processing and Sentiment Analysis~~ classification and clustering algorithms, convolutional Neural network

Research Question 2
- Most developers are interested in feature selection, selection of more appropriate algorithm. or even how they should train the data set
- lack of answer to questions showing availability of community support for these topics
- most of the posts there is only one answer at most and some of them only had comments
- Classification questions are well-answered
- Theoretical questions on SO remain unanswered
- Programming related questions get more answers
- Developers mostly care about short term solutions
- Developers face issues in identifying the right format of their input data files
- Some developers lack basic understanding of partitioning data into training validation, and testing and the concept of over-fitting
- computer vision and performance issues are more difficult
- developers need better introductory machine learning training and artificial intelligence training

Research Question 3:
- developers are asking more machine learning questions than before, and that is true for all four broad categories defined in the study

- Are machine learning questions different than others in terms of views and accepted answers?
- Compared view counts and answer counts of machine learning posts with posts that are not related to machine learning
- Kolmogorov-Smirnov test and found the differences for both variables to be statistically significant at $\alpha = 0.01$ (p-value = 2.2e-16 for both)
- answer counts show machine learning questions are not answered frequently.
- 56% machine learning questions dont have accepted answer
- 19% have no answers at all
- machine learning questions are harder to answer and demand more work

machine learning posts

Research Question 4

- If tags for machine learning posts are accurate because this is directly related to the number of views and answers a post can recieve
- written (manual) tags matching LDA considered correct
- 66.5% accuracy with LDA
- 72% ~~by~~ accuracy by removing outliers with LDA
- Many SO users do not have the domain knowledge for writing appropriate machine learning tags.
- LDA suggested tag for that post got immediately accepted by the SO community

Threats to Validity

- Internal Validity: There might be posts on SO that are about machine learning but do not have a tag
- Conclusion validity: manual labeling of topics is problematic
  - was minimized by using >1 author

Related Work

- Pinto et al: analyzed SO posts to understand what developers know about software energy consumption
- Yang et al: investigated security related questions from SO these type of studies are beneficial to educators & academic
- Patel et al: Statistical machine learning with expert researchers

Conclusion

- employed topic modelling techniques to identify key areas of interest to developers
- LDA gives 44 topics reclassified into 4 categories
- developers lack proper introductory understanding of machine learning and they dont recieve enough feedback from community
- evident that more intro education in machine learning should be given to developers is required.

- topic modeling approach towards a better tagging system for machine learning questions
- a tagging system would help developers to reach the right people in the community, and would possibly bring earlier feedback on their questions

What do developers ask about ML libraries! A large-scale study using Stack Overflow

- we dont yet understand difficulties faced by software developers when learning about ML libraries and using them within their systems
- Study the questions and perform statistical analysis to explore the answers to our research objectives (finding the most difficult stage, understanding the nature of problems, nature of libraries and studies whether the difficulties stayed consistent over time)
- Both static & dynamic analyses are absent, which are needed to find errors earlier
- API design improvement are needed
  API Misuses are prevalent
- Providing higher levels of abstraction needed
- Understanding behaviour of the trained model is prevalent

Introduction

- ML is becoming an essential computational tool
- ML can introduce unique software development problems
- Stack overflow can give significant insights
- The following libraries were examined: caffe
  H2O
  Keras
  Mahout
  MLlib
  Scikit learn
  Tensor flow
  Theano
  Torch
  weka

- Caffe is a deep learning library
- H2O is a deep learning library 3 to pride a workflow like system for building ML Models
- Keras deep learning library for Python for high levels of abstraction for making neural networks easier
- Mahout is aimed dt providing scalable ML facilities for Hadoop clusters
- MLlib is aimed at providing scalable ML facilities Br Spark clusters
- Scikit-learn is a python library that uses Tensorflow or Theano as the backend This library proide set of abstract APIs
- Tensorflow provides facilities to present a ML from the user in an effort
- Theano and Torch are aimed at scaling ML algorithm using
- Weka is a ML library for java. it provides API support for data preparation, classification, regression, clustering and associa

RQ1 Difficult Stage which Stages are more difficult in a ML pipeline?

RQ2 Nature of problems. which problems are more specific to library and which are inherent to ML?

RQ3 Nature of libraries which libraries face problems in specific Stages and which ones face difficulties in all Stages?

RQ4 Consistency Did the problems stay consistent over the time?

Methodology

of upvote
# of downvote

- Score awarding system: $S = |N_u| - |N_d|$
- Higher score is an indicator of better question

Classification of question

- Classify questions into top-level categories: ML? or not?
- Data preparation: converting raw data into input data
  - Data adaption: Questions under this subcategory are about reading raw data into the suitable data format required by the library
    - converting data, encoding

- Featuring: Questions under this category are about feature extraction and selection
  - Extraction: reduce dimensionality (# of existing features)
  - selection: reduce dimensionality of informative features

- Type mismatch: when type of data by user doesnt match ML requirement

- Shape mismatch: when dimension of tensor or matrix provided by layer doesnt match dimension needed by the next layer

- Data cleaning: removal of null values, handling missing values, encoding data

- Modelling:
  - Model selection: questions related to the choice of best model and choice of API version
  - Model creation: question related to creating ML model using the APIs
  - Model conversion: question related to conversion of a model trained using one library and then using the the trained model for prediction in an environment using another library
  - Model load/store: questions about storing models to disk and loading them to use later

- Training:
  - Error/Exception: Questions that appear in training phase fall into this sub categories.
  - Parameter selection: Some frameworks have optional parameters, and developers have to choice appropriate values for thes parameters & pass relevant values to regd Parmetrs
  - Loss function: questions related to choosing and creating loss function fall into this category
  - Optimizer: question related to choice of optimizer
  - Performance questions related to long training time and/or high memory consumption

- Accuracy: Questions related to training accuracy and/or convergence

- Evaluation
  - Evaluation method selection: Question related to the problems in the usage of APIs for doing validation.
- Visualizating model learning: Questions about visualizing behavior of model to get a better understanding of the training process & the effects of evaluation on the change of loss function & accuracy

- Hyper-Parameter Tuning
  - Improving models performance
  - Tuning strategy selection: Question about choosing among APIs for different tuning methodologies
  - Tuning parameter selection: Discussions related to the selection of parameters for tuning 2022-05-12

2022-05-11

- Prediction
  - After model trained and evaluated, the model it's used to predict new input data.
  - Prediction accuracy: Questions related to prediction accuracy
  - Model reuse: Developers might have difficulty in reusing existing models with their own datasets
  - Robustness: Question about stability of models with slight change, possibly noise, in the datasets