Final Project: Assessing Software Fault Risk With Deep Learning

DATA/COMPSCI 182 Deep Learning
Lecture 20 11/14/2024

What is this project about ?

- Fundamentally, it is focused on …. Deep Learning :)
- In the context of a REAL problem with REAL data
- The problem is in the domain of Automated Software Quality Assessment
  - A DHS funded project called CodeFault
  - Large software development efforts
    - With multiple developers over multiple years
- Problem: Predict which code commits are likely to be faulty, down the road

You will be provided (next week):

- A clear project specification
  - Including direction on which DNN algorithms to develop or apply
  - What to investigate
- Data
  - In structured, ready to load CSVs
- Some (very !) skeletal code
  - Function to load in data into Tensors
    - To which DNNs can then be applied

The Original CodeFault Investigation

- Described in [Ashish, Barish, Minton 2022]
- Was on this very dataset
  - Curated from open repositories on GitHub, as well as (open) profiles of associated developers on GitHub
  - We are providing a partitioned subset of this data for the final project
- Primarily explored feature-driven machine learning algorithms
  - Or Ensembles thereof
  - Some evaluation with AutoML
- The final project is an opportunity to dive into the yet to be investigated: Variety of Deep Learning algorithms to the same problem, and data

Prior Investigation

TO GET YOU FAMILIAR WITH THE PREDICTION PROBLEM, AND COMFORTABLE WITH THE DATA

Introduction

- A data-driven solution for assessing fault (defect) risk in software code
- Curated database
- Commit level problem formulation
- 10-fold increase in fault identification precision
  - And with very sparse fault density (1-2%)
- Software vulnerability identification
  - Multiple levels

Software vulnerability

- Investigated at various levels
  - Code
  - Other metadata
  - (Development) behavior

Feature Engineering

- Three types of features
  - Commit oriented
  - Developer oriented
  - Code oriented

Data

- CSV FORMAT
    - 2 PRIMARY TABLES
- FAULT DENSITY IS SPARSE !

Commit oriented features

Feature	Description	Example
ID	Commit identifier	12345
Commit SHA	Commit hash identifier	2acee567eed8889f7ae
Commit message	The text description included as commit documentation	'Updated PCRE used for win32 builds.'
Modifications count	Modifications in a commit	45
Additions count	Additions in a commit	32
Deletions count	Deletions in a commit	39
Author name, login, ID	Code author name, login, and identifier	John Smith, Jsmith123, 4563
Author email	Code author email	jsmith@microsoft.com
Committer name, login, ID	Code committer name, login, and identifier	Mike Foster, mfoster, 3322
Commit date Hour of day* Day of week*	Date stamp of the commit. Note that (the commit) 'hour of day' and 'day of week' are derived from the commit date stamp	'2019-04-24 13:38:51' 13 Friday

Developer, Code oriented features

Feature	Description	Example
ID	(Unique) developer identifier	1222
login	Developer login name	JSmith123
avatar_url	Profile URL	Github.com/ JSmith123
Company	Organization affiliated with	Microsoft
Blog	Developer has a blog (Y/N)	Y
Location	Geolocation if given	Seattle
Email	Email address	jsmith@microsoft.com
Hireable	Hireable (Y/N)	N
Bio	Developer bio (if any)	
Public repos count	Count of developer's repositories	73
Public gists count	Count of developer's gists	4
Followers count	Count of followers	35
Following count	Count of people followed	22

Feature	Description	Example
ID	File identifier	8
Commit meta ID	Associated commit identifier	2
File path	Directory path to the file	src/main/scripts/ /proxy_module.java
Status	Whether the file has been modified is this commit	modified
Modification, additions, deletions counts	Activity counts	23, 17, 4
Path 1 *	Root folder name	src
Path 2*	File name	proxy_module
Ext	File extension	"java"
Path as text	Path folder xname tokens	"src main scripts proxy_module"
Deleted code		`return X+","+wordLength(str)+ ","+endsNumber(str) +","+ hasYesNo(str)+","+ charFractions(str);`

Machine learning classification

- Three paradigms
  - Feature driven classifier
  - Ensemble
  - Deep learning

Evaluation: Feature driven classification

Classifier	Nginx (Baseline: 0.02)			Apache (Baseline: 0.01)			Wget (Baseline: 0.02)		
	P	R	F	P	R	F	P	R	F
Random Forest	0.11	0.79	0.19	0.03	0.74	0.06	0.09	0.81	0.16
Decision Trees	0.59	0.74	0.66	0.34	0.46	0.39	0.31	0.27	0.29
ABV	0.09	0.71	0.16	0.01	0.56	0.02	0.00	0.00	NA
Gaussian Naïve Bayes	0.20	0.36	0.26	0.03	0.07	0.04	0.00	0.00	NA
SVM	0.00	0.00	NA	0.00	0.00	NA	0.00	0.00	NA
QDA	0.19	0.37	0.25	0.03	0.09	0.04	0.00	0.00	NA
K-Nearest Neighbor	0.10	0.20	0.13	0.00	0.00	NA	0.03	0.08	0.04
Commit message text	0.12	0.40	0.18	0.07	0.19	0.10	0.12	0.11	0.11
File tokens text	0.05	0.84	0.09	0.02	0.74	0.04	0.03	0.43	0.06
Code snippets text	0.08	0.14	0.10	0.02	0.05	0.03	0.29	0.05	0.09

Evaluation: Ensembles

Votes	Nginx (Baseline: 0.02) P	R	F	Apache (Baseline: 0.01) P	R	F	Wget (Baseline: 0.02) P	R	F
2	0.09	0.81	0.16	0.02	0.80	0.04	0.09	0.81	0.16
4	0.20	0.71	0.31	0.15	0.38	0.22	0.31	0.27	0.29
6	0.31	0.37	0.34	0.29	0.03	0.05	0.00	0.00	NA
8	0.75	0.09	0.16	0.00	0.00	NA	0.00	0.00	NA
9	1.00	0.01	0.02	0.00	0.00	NA	0.00	0.00	NA
10	0.00	0.00	NA	0.00	0.00	NA	0.00	0.00	NA

Votes	Nginx (Baseline:0.02) P	R	F	Apache (Baseline: 0.01) P	R	F	Wget (Baseline: 0.02) P	R	F
ALL Classifiers	0.57	0.73	0.64	0.30	0.39	0.34	0.36	0.11	0.17
Only Decision Tree & Random Forest	0.53	0.73	0.61	0.35	0.42	0.38	0.49	0.49	0.49

Evaluation: Deep learning

Dataset	Best In-house			AutoML		
	P	F	R	P	F	R
nginx	0.59	0.74	0.66	0.38	0.25	0.31
apache	0.34	0.46	0.39	0.14	0.04	0.06
curl	0.21	0.22	0.22	0.52	0.08	0.14
wget	0.00	0.00	NA	0.00	0.00	NA
videolan	0.43	0.43	0.43	0.53	0.29	0.38
podofo	0.22	0.62	0.32	0.33	0.33	0.33

Explainability: Feature importance

Feature importance	Feature importance
author_name 0.93816 additions_count_commit 0.04141 modifications_count_commit 0.01663 deletions_count_commit 0.00041	author_name 0.87569 additions_count_commit 0.06819 modifications_count_commit 0.02683 deletions_count_commit 0.01080
committer_date_weekday 0.75268 author_name 0.07270 additions_count_commit 0.05066 committer_date_hour 0.05057 committer_name 0.04726 deletions_count_commit 0.00996 modifications_count_commit 0.00590	committer_date_hour 0.61274 modifications_count_commit 0.16359 committer_name 0.07454 author_name 0.05906 additions_count_commit 0.05896 committer_email_type 0.01382 last_month_faulty_commits 0.00893

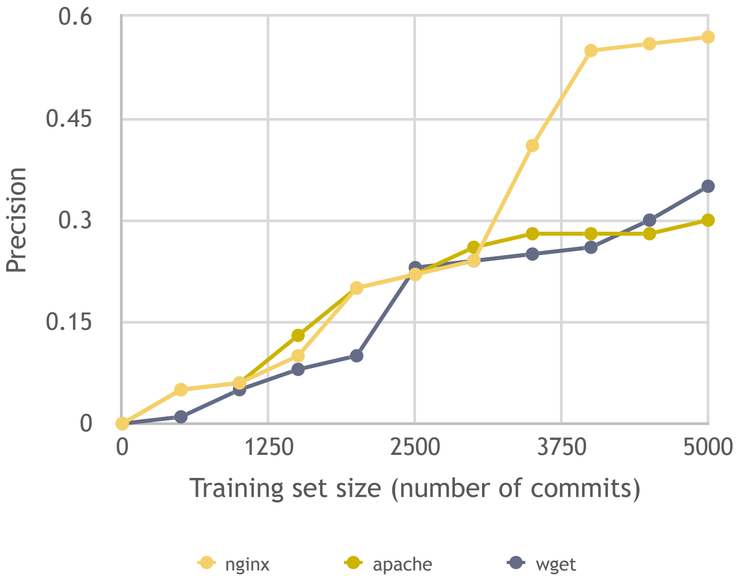
Real-world development risk analysis

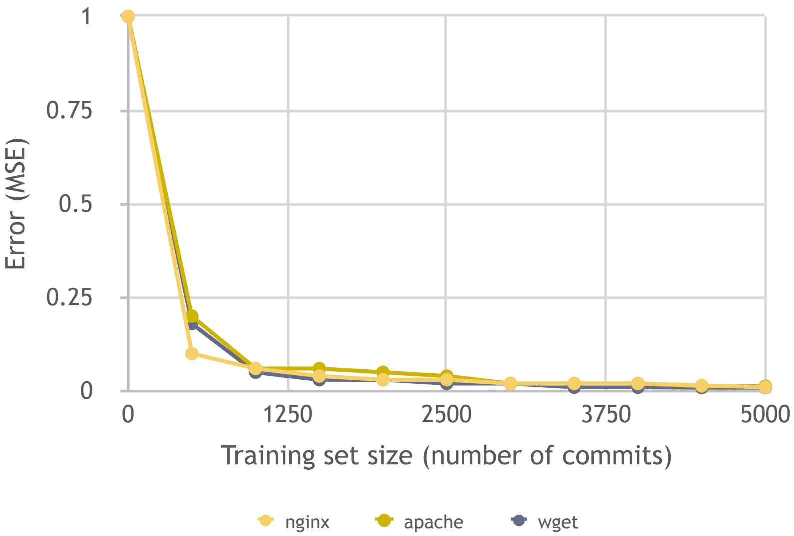
Dataset	Statistical Analysis: Odds ratio	Classifier with no history	Classifier with previous month history	Classifier with activity history but no fault knowledge
apache 1	3	2	15	10
apache 3	4	8	25	20
apache 5	1	5	10	10
i-magick	1	2	6	10
curl 1	5	3	5	5
curl 3	4	0	10	10
curl 5	1	3	7	7
wget 1	1	1	1	1
wget 3	0	3	5	5
wget 5	1	1	3	3
openssl 1	3	5	10	11
openssl 3	13	0	8	16
openssl 5	8	3	15	13
nginx 1	5	5	10	10
libraw 1	4	4	5	5
libraw 5	3	5	6	6

Explainability: History based features

Feature importance		Feature importance	
author_name	0.93816	author_name	0.87569
additions_count_commit	0.04141	additions_count_commit	0.06819
modifications_count_commit	0.01663	modifications_count_commit	0.02683
last_month_commits	0.00324	last_month_commits	0.01550
deletions_count_commit	0.00041	deletions_count_commit	0.01080
last_month_faulty_commits	0.00014	last_month_faulty_commits	0.00295
committer_date_weekday	0.75268	committer_date_hour	0.61274
author_name	0.07270	modifications_count_commit	0.16359
additions_count_commit	0.05066	committer_name	0.07454
committer_date_hour	0.05057	author_name	0.05906
committer_name	0.04726	additions_count_commit	0.05896
deletions_count_commit	0.00996	committer_email_type	0.01382
last_month_faulty_commits	0.00675	last_month_faulty_commits	0.00893
modifications_count_commit	0.00590	deletions_count_commit	0.00447
last_month_commits	0.00325	last_month_commits	0.00390

Learning curve analysis





- ~ 4000 commits in training, for a stable model

Related work

- Recent work in
  - Periodic capture of development behavior
  - Personalized (to developer) defect models
  - Patent incorporating some history based features
- Distinguishing aspects
  - 10-fold precision increase, over sparse fault density
    - Feature engineering
    - Comprehensive machine learning investigation
  - More comprehensive history based features
  - Real-world data, forward-in-time risk prediction

Conclusions, Future work

- Classifiers
  - Decision Tree
- Limited space of relevant features
- Recent history
- Real-world implications

Final Project: Concluding Thoughts

- Please read the associated paper !
  - It is the best way to get the problem context and familiarity with the kind of data you will work with
    - Before the project is released
- Questions :) ?