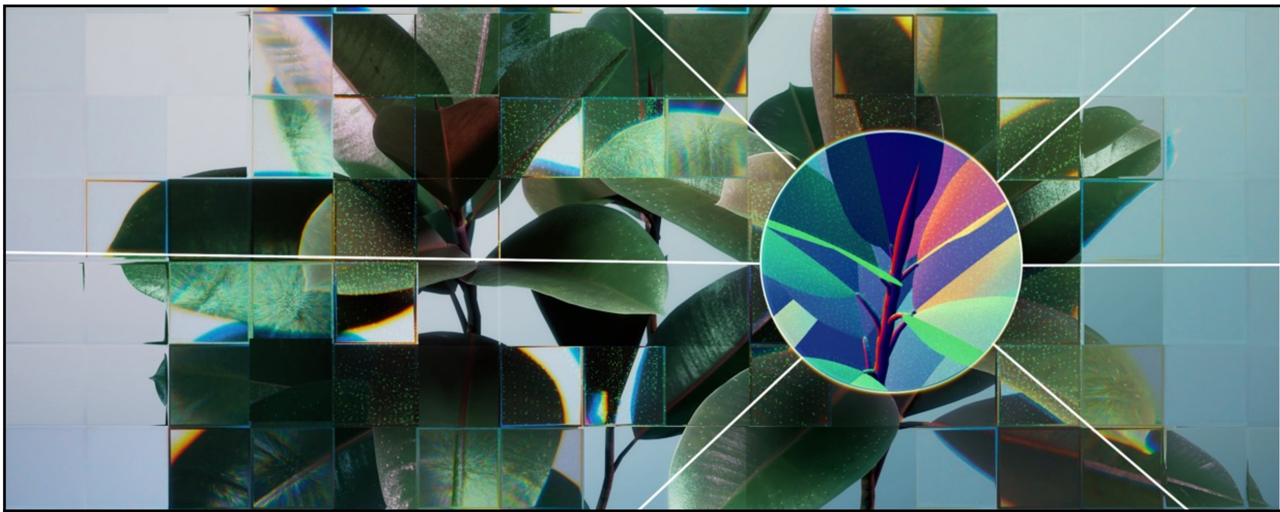


COMP47590: Advanced Machine Learning
Supervised Learning - Evaluation

Dr. Brian Mac Namee

1

Section Outline

In this section we will cover:

- Why evaluate?
- Choosing appropriate evaluation metrics
- Running machine learning experiments

This lecture largely focuses on supervised machine learning.

2

Why Evaluate?

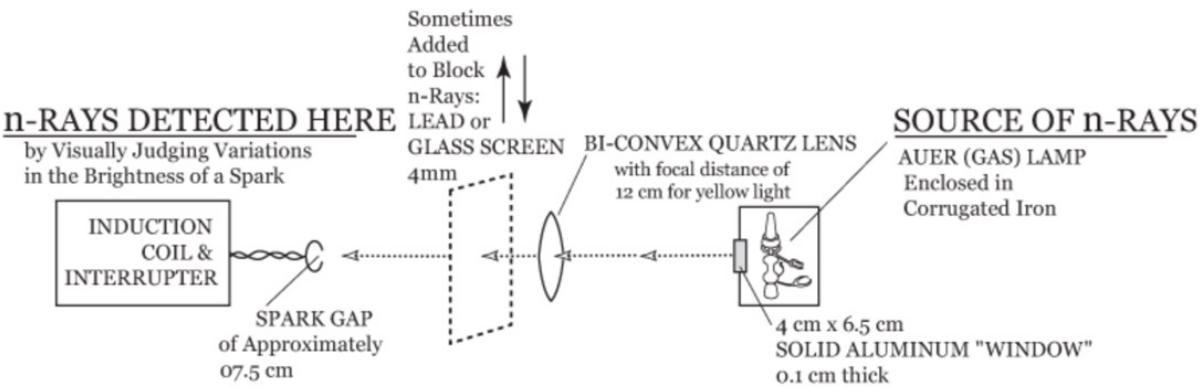
3

Prosper-René Blondlot
Discoverer of n-Rays

https://en.wikipedia.org/wiki/Prosper-Ren%C3%A9_Blondlot

Who has heard of n-rays?

4



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

5

Prosper-René Blondlot
Discoverer of n-Rays

https://en.wikipedia.org/wiki/Prosper-Ren%C3%A9_Blondlot

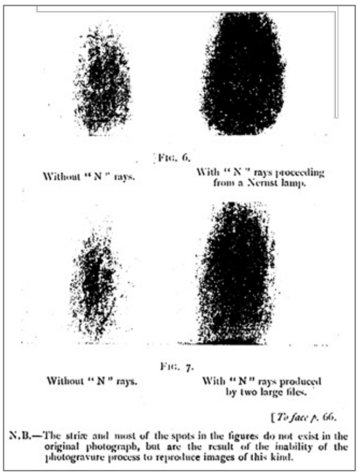


FIG. 6.

Without "N" rays. With "N" rays proceeding from a Nernst lamp.

FIG. 7.

Without "N" rays. With "N" rays produced by two large files.

[To face p. 66.

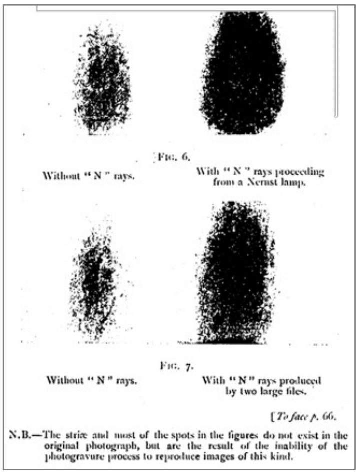
N.B.—The striæ and most of the spots in the figures do not exist in the original photograph, but are the result of the inability of the photogravure process to reproduce images of this kind.

https://en.wikipedia.org/wiki/N-ray

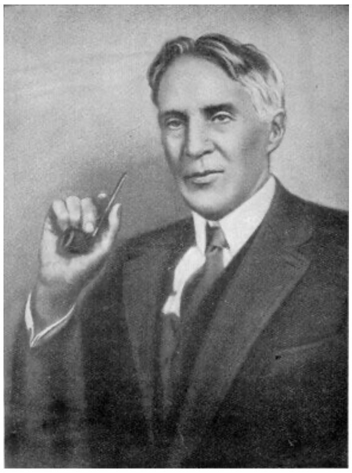
6

Prosper-René Blondlot
Discoverer of n-Rays

https://en.wikipedia.org/wiki/Prosper-Ren%C3%A9_Blondlot



https://en.wikipedia.org/wiki/N-ray



https://www.wired.com/2014/09/fantastically-wrong-n-rays/

APS This Month in Physics History:
September 1904: Robert Wood debunks N-rays
https://www.aps.org/publications/apsnews/200708/history.cfm

Fantastically Wrong: The Imaginary Radiation That Shocked Science and Ruined Its 'Discoverer'
https://www.wired.com/2014/09/fantastically-wrong-n-rays/

7

Why Evaluate?

It is worth distinguishing between two different types of evaluation that we do in machine learning

1. evaluating a model that we would like to deploy for a specific task (**industry**)
2. comparing machine learning methods (**research**)

8

Why Evaluate?

The purpose of evaluation when we want to deploy a model (**industry**) is threefold

1. to determine which algorithm (plus hyperparameter values) is the most suitable for a task
2. to estimate how a model will perform after deployment
3. to convince users that a model will meet their needs

9

Why Evaluate?

The purpose of evaluation when we want to comparing machine learning methods (**research**) can be different

1. to evaluate the performance of a new method against existing baselines
2. to determine the *best* machine learning approach for a specific problem
3. to perform a benchmark experiment

These almost all reduce to an experiment in which we compare multiple approaches using multiple datasets

10

Evaluation For Deployment

11

Hold-Out Test Set

It is important to establish a hold-out test set early on (we'll come back to why)

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

12

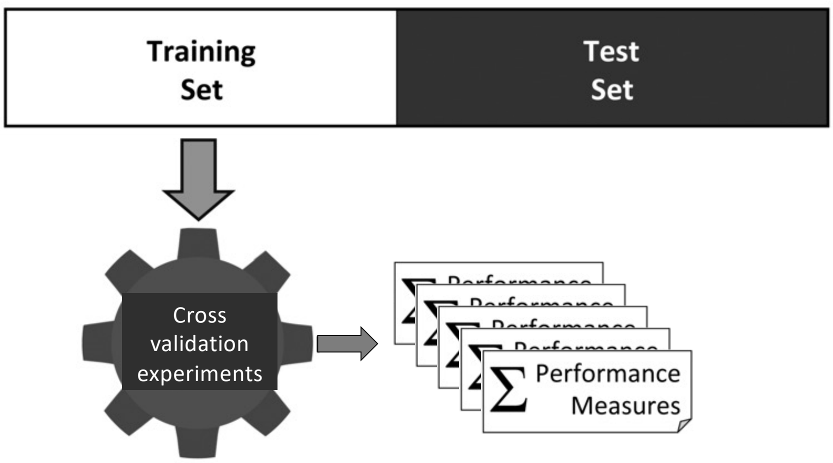
Evaluation For Deployment

The purpose of evaluation when we want to deploy a model (**industry**) is threefold

1. **to determine which algorithm (plus hyperparameter values) is the most suitable for a task**
2. to estimate how the model will perform after deployment
3. to convince users that the model will meet their needs

13

Evaluation For Deployment



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

14

Evaluation For Deployment

k fold cross validation

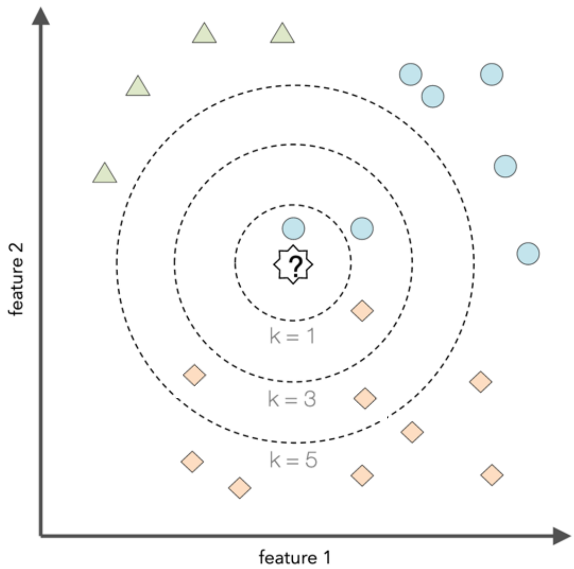
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

15

Evaluation For Deployment

Hyper-parameters

- Hyper-parameters are parameters of the learning algorithm
- All machine learning algorithms have them!
- Hyperparameter values can make a big difference!



Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, Sebastian Raschka
https://arxiv.org/abs/1811.12808

16

Evaluation For Deployment

Hyper-parameter tuning

- A grid search or random search based on cross validations makes sense
- AutoML approaches are good if you have access to them
- After finding best hyper-parameters we should do one last cross validation experiment with a different shuffle

17

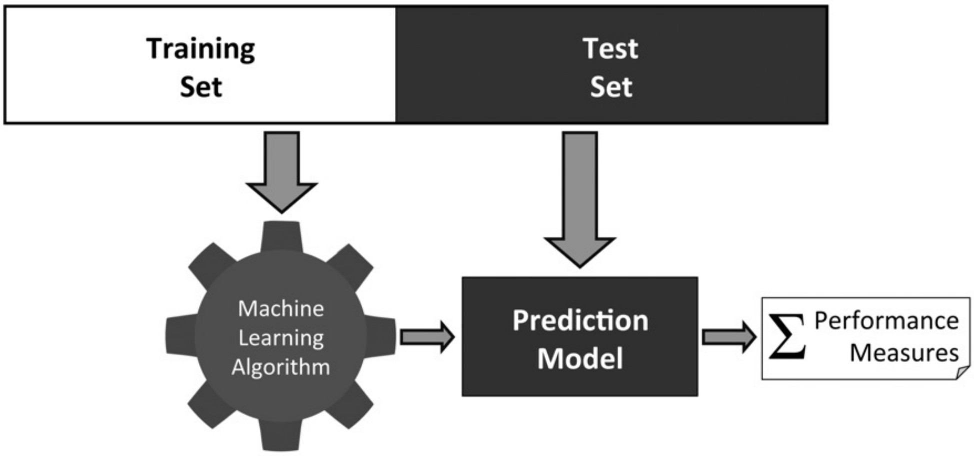
Evaluation For Deployment

The purpose of evaluation when we want to deploy a model (**industry**) is threefold

1. to determine which model is the most suitable for a task
2. **to estimate how the model will perform after deployment**
3. to convince users that the model will meet their needs

18

Evaluation For Deployment



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

19

Evaluation For Deployment

The purpose of evaluation when we want to deploy a model (**industry**) is threefold

1. to determine which model is the most suitable for a task
2. to estimate how the model will perform after deployment
3. **to convince users that the model will meet their needs**

20

Performance Measures

21

Performance Measures

Machine learning offers us a raft of diffrent performance measures for experiments

- Accuracy
- Precison, Recall, F1 score
- Macro-averaged F1 score
- Sensitivity, Specificity
- ROC index
- Dunne index
- Gain & Lift
- Kolmogorov smirnoff
- RMSE
- R^2
- ...

It is really important to pick the right performance measure for the problem you are solving

22

Performance Measures

Model 1 confusion matrix **91%**

		Prediction	
		'non-churn'	*'churn'*
Target	*'non-churn'*	90	0
	'churn'	9	1

Model 2 confusion matrix **78%**

		Prediction	
		'non-churn'	*'churn'*
Target	*'non-churn'*	70	20
	'churn'	2	8

23

Performance Measures

Model 1 confusion matrix $\frac{1}{2}(1 + 0.1) = 55\%$

		Prediction	
		'non-churn'	*'churn'*
Target	*'non-churn'*	90	0
	'churn'	9	1

Model 2 confusion matrix $\frac{1}{2}(0.778 + 0.8) = 78.889\%$

		Prediction	
		'non-churn'	*'churn'*
Target	*'non-churn'*	70	20
	'churn'	2	8

24

Macro Averaging

The specific lesson in this example is to use macro averaging rather than micro averaging when classification datasets are imbalanced

25

Evaluation For Research

26

Why Evaluate?

The purpose of evaluation when we want to comparing machine learning methods (**research**) can be different

1. to evaluate the performance of a new method against existing baselines
2. to determine the *best* machine learning approach for a specific problem
3. to perform a benchmark experiment

These almost all reduce to an experiment in which we compare multiple approaches using multiple datasets

27

Why Evaluate?

The purpos[...]mparing machine le[...]ferent

1. to eva[...]hod agains[...]
2. to det[...]proach for a s[...]
3. to per[...]

We don't care about building one specific model for a problem

These almost all reduce to an experiment in which we compare multiple approaches using multiple datasets

28

Benchmarks

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10
Approach 1										
Approach 2										
Approach 3										
Approach 4										
Approach 5										
Approach 6										
Approach 7										
Approach 8										
Approach 9										
Approach 10										

29

Benchmarks - Performance Measures

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10
Approach 1	0.50	0.42	0.72	0.63	0.58	0.56	0.79	0.83	0.99	0.70
Approach 2	0.45	0.39	0.61	0.53	0.56	0.41	0.75	0.62	0.89	0.59
Approach 3	0.48	0.42	0.66	0.57	0.60	0.45	0.76	0.65	0.91	0.65
Approach 4	0.52	0.57	0.74	0.58	0.61	0.58	0.76	0.67	1.00	0.69
Approach 5	0.65	0.46	0.79	0.69	0.81	0.64	0.89	0.72	1.00	0.76
Approach 6	0.56	0.40	0.67	0.61	0.67	0.50	0.77	0.70	0.94	0.60
Approach 7	0.48	0.40	0.63	0.54	0.58	0.43	0.77	0.63	0.90	0.60
Approach 8	0.51	0.53	0.66	0.57	0.62	0.46	0.88	0.73	0.93	0.61
Approach 9	0.61	0.57	0.77	0.54	0.70	0.59	0.94	0.75	1.00	0.62
Approach 10	0.50	0.48	0.64	0.53	0.63	0.47	0.84	0.64	0.95	0.67

30

Benchmarks - Ranks

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10
Approach 1	6	7	4	2	8	4	5	1	4	2
Approach 2	10	10	10	10	10	10	10	10	10	10
Approach 3	8	6	6	5	7	8	9	7	8	5
Approach 4	4	1	3	4	6	3	8	6	1	3
Approach 5	1	5	1	1	1	1	2	4	1	1
Approach 6	3	8	5	3	3	5	6	5	6	9
Approach 7	9	9	9	7	9	9	7	9	9	8
Approach 8	5	3	7	6	5	7	3	3	7	7
Approach 9	2	2	2	8	2	2	1	2	1	6
Approach 10	7	4	8	9	4	6	4	8	5	4

31

Benchmarks - Average Ranks

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10	Avg Rank
Approach 1	6	7	4	2	8	4	5	1	4	2	4.3
Approach 2	10	10	10	10	10	10	10	10	10	10	10.0
Approach 3	8	6	6	5	7	8	9	7	8	5	6.9
Approach 4	4	1	3	4	6	3	8	6	1	3	3.9
Approach 5	1	5	1	1	1	1	2	4	1	1	1.8
Approach 6	3	8	5	3	3	5	6	5	6	9	5.3
Approach 7	9	9	9	7	9	9	7	9	9	8	8.5
Approach 8	5	3	7	6	5	7	3	3	7	7	5.3
Approach 9	2	2	2	8	2	2	1	2	1	6	2.8
Approach 10	7	4	8	9	4	6	4	8	5	4	5.9

32

Benchmarks - Average Ranks

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6	Data 7	Data 8	Data 9	Data 10	Avg Rank
Approach 1	6	7	4	2	8	4	5	1	4	2	4.3
Approach 2	10	10	10	10	10	10	10	10	10	10	10.0
Approach 3	8	6	6	5	7	8	9	7	8	5	6.9
Approach 4	4	1	3	4	6	3	8	6	1	3	3.9
Approach 5	**1**	**5**	**1**	**1**	**1**	**1**	**2**	**4**	**1**	**1**	**1.8**
Approach 6	3	8	5	3	3	5	6	5	6	9	5.3
Approach 7	9	9	9	7	9	9	7	9	9	8	8.5
Approach 8	5	3	7	6	5	7	3	3	7	7	5.3
Approach 9	2	2	2	8	2	2	1	2	1	6	2.8
Approach 10	7	4	8	9	4	6	4	8	5	4	5.9

33

Benchmarks - Significance Testing

Another characteristic of acdemic evaluations versus industry evaluations is the use of significance testing

- Is the *best* method really *best*?

34

Benchmarks - Significance Testing

We recommend a two step process:

- **Friedman aligned rank test** to first test whether a significant difference between the performance of the algorithms over the datasets exists
  - Compared against a significance level (e.g. 0.05)
- If a difference does exist then a pairwise **Nemenyi test** should be performed to show between which algorithm pairs the significant differences exist
  - Compared against a significance level (e.g. 0.05)

35

Benchmarks - Significance Testing

Nemenyi test gives us a significance matrix

- Which approaches are significantly different from which others?

Significance matrix can give rise to a **critical differences plot** which shows groups of approaches which are significantly different from each other

36

Benchmarks - Significance Testing

Friedman's Aligned Rank Test for Multiple Comparisons

- T = 57.609, df = 9, p-value = 3.862e-09

Nemenyi test

- Critical difference = 4.393, k = 10, df = 90

37

Benchmarks - Significance Testing

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0	-5.6	-2.7	0.25	2.45	-1	-4.05	-0.85	1.55	-1.55
M2	-5.6	0	2.9	5.85	8.05	4.6	1.55	4.75	7.15	4.05
M3	-2.7	2.9	0	2.95	5.15	1.7	-1.35	1.85	4.25	1.15
M4	0.25	5.85	2.95	0	2.2	-1.25	-4.3	-1.1	1.3	-1.8
M5	2.45	8.05	5.15	2.2	0	-3.45	-6.5	-3.3	-0.9	-4
M6	-1	4.6	1.7	-1.25	-3.45	0	-3.05	0.15	2.55	-0.55
M7	-4.05	1.55	-1.35	-4.3	-6.5	-3.05	0	3.2	5.6	2.5
M8	-0.85	4.75	1.85	-1.1	-3.3	0.15	3.2	0	2.4	-0.7
M9	1.55	7.15	4.25	1.3	-0.9	2.55	5.6	2.4	0	-3.1
M10	-1.55	4.05	1.15	-1.8	-4	-0.55	2.5	-0.7	-3.1	0

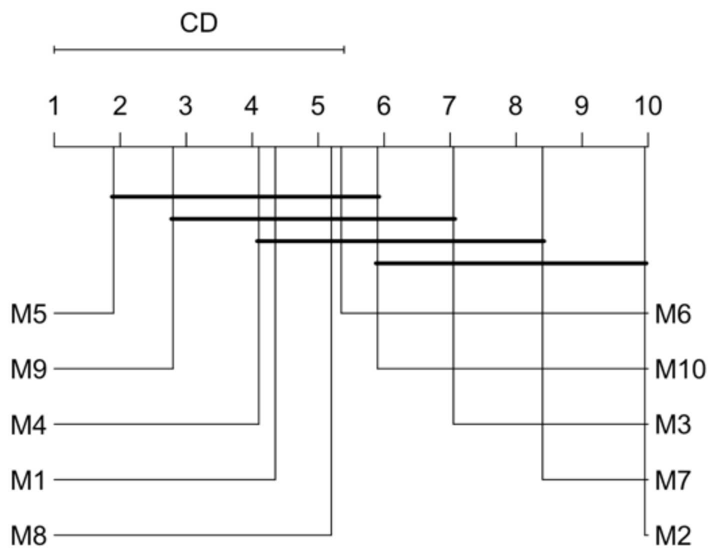
38

Benchmarks - Significance Testing

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	False	True	False	False	False	False	False	False	False	False
M2	True	False	False	True	True	True	False	True	True	False
M3	False	False	False	False	True	False	False	False	False	False
M4	False	True	False	False	False	False	False	False	False	False
M5	False	True	True	False	False	False	True	False	False	False
M6	False	True	False	False	False	False	False	False	False	False
M7	False	False	False	False	True	False	False	False	True	False
M8	False	True	False	False	False	False	False	False	False	False
M9	False	True	False	False	False	False	True	False	False	False
M10	False	False	False	False	False	False	False	False	False	False

39

Benchmarks - Significance Testing



Critical differences plot

40

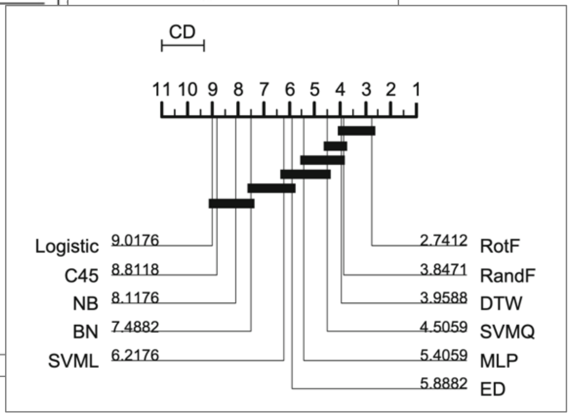
The Great Time Series Bakeoff

Data Min Knowl Disc (2017) 31:606–660
DOI 10.1007/s10618-016-0483-9

classification bake off: a review
luation of recent algorithmic

Table 6 Average accuracy of the best nine classifiers over 85 problems

Datasets	COTE	ST	BOSS	EE	DTW_F	TSF	TSBF	LPS	MSM
Adiac	**0.81**	0.768	0.749	0.665	0.605	0.707	0.727	0.765	
ArrowHead	**0.877**	0.851	0.875	0.86	0.776	0.789	0.801	0.806	
Beef	**0.764**	0.736	0.615	0.532	0.546	0.648	0.554	0.52	
BeetleFly	0.921	0.875	**0.949**	0.823	0.853	0.842	0.799	0.893	
BirdChicken	0.941	0.927	**0.984**	0.848	0.865	0.839	0.902	0.854	
Car	0.899	**0.902**	0.855	0.799	0.851	0.758	0.795	0.836	
CBF	0.998	0.986	**0.998**	0.993	0.979	0.958	0.977	0.984	
ChlorineConcentration	**0.736**	0.682	0.66	0.659	0.658	0.719	0.683	0.642	
CinCECGtorso	**0.983**	0.918	0.9	0.946	0.714	0.974	0.716	0.743	
Coffee	**1**	0.995	0.989	0.989	0.973	0.989	0.982	0.95	
Computers	0.77	0.785	**0.802**	0.732	0.659	0.768	0.765	0.726	
CricketX	**0.814**	0.777	0.764	0.801	0.769	0.691	0.731	0.696	
CricketY	**0.815**	0.762	0.749	0.794	0.756	0.688	0.728	0.706	
CricketZ	**0.827**	0.798	0.776	0.804	0.785	0.707	0.738	0.714	
DiatomSizeReduction	0.925	0.911	0.939	**0.946**	0.942	0.941	0.89	0.915	
DistalPhalanxOAG	0.805	**0.829**	0.815	0.768	0.796	0.809	0.816	0.767	
DistalPhalanxOC	**0.821**	0.819	0.814	0.768	0.76	0.813	0.812	0.742	
DistalPhalanxTW	**0.693**	0.69	0.673	0.654	0.658	0.686	0.69	0.618	
Earthquakes	0.747	0.737	0.746	0.735	**0.747**	0.747	0.747	0.668	
ECG200	0.873	0.84	**0.89**	0.881	0.819	0.868	0.847	0.807	



The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances, Bagnall et al
https://link.springer.com/article/10.1007/S10618-016-0483-9

Time Series Classification
https://www.timeseriesclassification.com/results.php

41

Summary

42

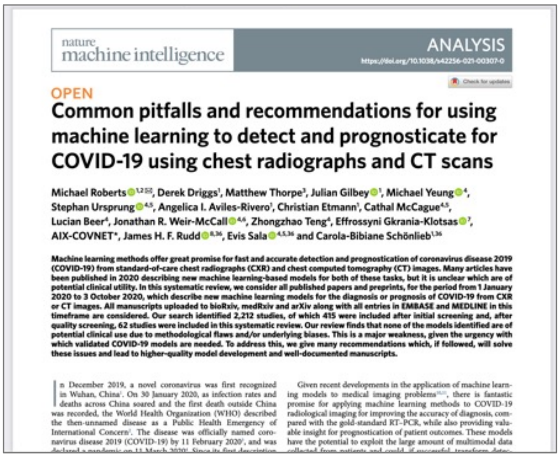
Summary

Evaluation

- Choosing appropriate evaluation mechanisms is crucial in doing machine learning properly
- Macro versus micro averaging is a mistake too often made
- One of the key differences between *industry* evaluations and *research* evaluations is the need for significance testing
- Lots of research evaluations reduce to a benchmark across multiple methods on multiple datasets

43

Discussion



nature machine intelligence

ANALYSIS

https://doi.org/10.1038/s42256-021-00307-0

Check for updates

OPEN

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts[1,2], Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], AIX-COVNET*, James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

In December 2019, a novel coronavirus was first recognized in Wuhan, China[1]. On 30 January 2020, as infection rates and deaths across China soared and the first death outside China was recorded, the World Health Organization (WHO) described the then-unnamed disease as a Public Health Emergency of International Concern[2]. The disease was officially named coronavirus disease 2019 (COVID-19) by 11 February 2020[3], and was

Given recent developments in the application of machine learning models to medical imaging problems, there is fantastic promise for applying machine learning methods to COVID-19 radiological imaging for improving the accuracy of diagnosis, compared with the gold-standard RT-PCR, while also providing valuable insight for prognostication of patient outcomes. These models have the potential to exploit the large amount of multimodal data

"Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 ...", Roberts et al
https://www.nature.com/articles/s42256-021-00307-0

Why do tree-based models still outperform deep learning on typical tabular data?

Léo Grinsztajn, Soda, Inria Saclay, leo.grinsztajn@inria.fr
Edouard Oyallon, MLIA, Sorbonne University
Gaël Varoquaux, Soda, Inria Saclay

Abstract

While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data (~10K samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and neural networks. This leads to a series of challenges which should guide researchers aiming to build tabular-specific neural network: **1.** be robust to uninformative features, **2.** preserve the orientation of the data, and **3.** be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20 000 compute hours hyperparameter search for each learner.

1 Introduction

Deep learning has enabled tremendous progress for learning on image, language, or even audio datasets. On tabular data, however, the picture is muddier and ensemble models based on decision trees like XGBoost remain the go-to tool for most practitioners [Sta] and data science competitions [Kossen et al., 2021]. Indeed deep learning architectures have been crafted to create inductive biases

"Why do tree-based models still outperform deep learning on typical tabular data?", Grinsztajn et al
https://openreview.net/pdf?id=Fp7__phQszn

44

Questions

?

45