

CSC 5741

Lecture 6: Introduction to Machine Learning



Lighton Phiri <lighton.phiri@unza.zm>

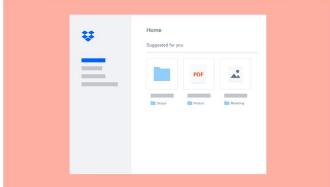
Department of Library and Information Science
University of Zambia

Todos Showcases—May 7, 2019

Using machine learning to predict what file you need next

Neeraj Kumar | 5 days ago

[Twitter](#) [Facebook](#) [LinkedIn](#) [0](#) [Email](#)



As we laid out in our [blog post introducing DBX1](#), Dropbox is building features to help users stay focused on what matters. Searching through your content can be tedious, so we built [content suggestions](#) to make it

<https://blogs.dropbox.com/tech/tag/machine-learning/>

April 30 2019

CSC 5741 L06 - 3

Announcements—May 7, 2019

- **Assessments**

- Class Theory Test: May 21, 2019
- Mini Project Deliverables: May 20, 2019
 - Technical Report
 - Code Repository for Fully Functional Implementation (including interactive Jupyter Notebook)
 - Presentation Slides
- Mini Project Presentations: May 28, 2019
 - Demonstrations [2 minutes]
 - Presentations [10 minutes]
 - Q&A [3 minutes]

April 30 2019

CSC 5741 L06 - 2

Lecture Series Outline

- **Part I: Introduction to Machine Learning**
- **Part II: Datasets**

April 30 2019

CSC 5741 L06 - 4

Lecture Series Outline

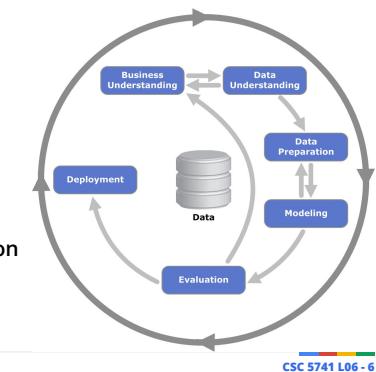
- **Part I: Machine Learning**
 - Introduction
 - Machine Learning
 - Data Attributes
 - Types of Learning
 - Evaluation
- **Part II: Datasets**

April 30 2019

CSC 5741 L06 - 5

Introduction (1/2)

- The Cross-industry standard process for data mining (CRISP-DM) is a model commonly used to highlight approaches in data mining
 - CRISP-DM segments a data mining project into six phases with no strict order of execution
 - Surveys conducted suggest CRISP-DM is the most widely used methodology

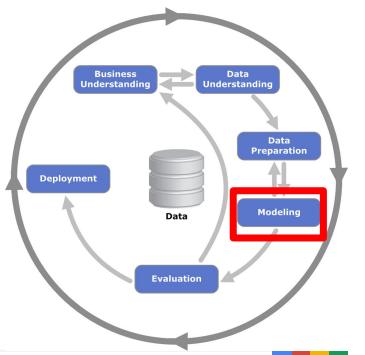


April 30 2019

CSC 5741 L06 - 6

Introduction (2/2)

- Define the model components, features, how it behaves and how to interpret it
- Evaluate the various alternative techniques that can be integrated with the model
 - e.g. Evaluate different classification algorithms

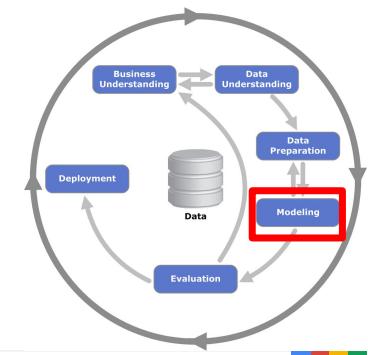


April 30 2019

CSC 5741 L06 - 7

Machine Learning (1/2)

- Finding patterns in data that provide insight or enable fast and accurate decision making
 - Prediction
 - Pattern recognition

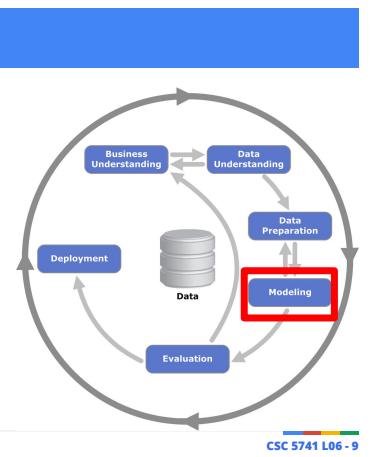


April 30 2019

CSC 5741 L06 - 8

Machine Learning (2/2)

- Aspects of Machine Learning**
 - Input: x
 - Output: $f(x)$
 - $f(x) \rightarrow y$
 - $f(x) \rightarrow$ prediction
 - $f(x) \rightarrow$ pattern
 - Input is generally attributes of a given dataset
 - A key prerequisite to solving a machine learning problem is to have **reliable data** available



April 30 2019

Data Representation

- Input and output data involving machine learning algorithms need to be represented in a mathematical way using data attributes.**
 - The representation is in the form of attribute-value pairs
 - E.g. {gender: "female", minor: "Mathematics"}
 - Attributes type: continuous, categorical and ordinal values
- The representation is dependant on the problem domain and the goals**
 - Predicting is student will pass or fail or JCTR Basic Needs Basket
- Remember that a predictor is a mathematical function**
 - $Y = f(x)$

April 30 2019

CSC 5741 L06 - 10

Data Attributes (1/5)

- Continuous/numeric attributes**
 - Integers or floats
 - Mathematical operations can be performed on values
 - Usually the case that values need normalisation to use a similar scale before applying algorithm

A	B	C	D	E
StudentID	CA	A Grade	Gender	Minor
3283f07fce04d7ce5cd5fd6d191bf	4.78 D	M	Civic	
3aa3919efac1157cc4a2529c3f391fe	23.7 C	M	Languages	
8e5435d4f634804d12512b6caeab4316	25.13 C+	M	Civic	
64e4a90e4dfa32a2877c93d48354c81	17.48 D	M	History	
89407737ae371e6080e88ad29ddd07bd	19.6 D	M	Civic	
b34e7468b181868a48635c0bb376e	32.6 B	M	Civic	
4db474b69680eadalc7d0862e99041	24.43 C	M	History	
a0d3784ea0902162d676fb0ae1fb6c36	20.8 D+	M	Mathematics	
c463632dc537747e3925bc8e4038ad42	26.65 C+	M	Mathematics	
f8b7ffcf26cde897a3ad67a83c99286d	30.55 B	M	Mathematics	
a96fb28013fb01358acfe2e85f598d	25.2 C+	F	Languages	
cad8f19590f7a666094bf5fa3d4c0633	19.6 D	F	History	
94b1c4e2144e735b872faa1078577aca4	25.45 C+	M	Mathematics	
e8c074808b39c3b3e345f4a774696590	25.28 C+	F	Civic	
5e1c7620b6f16173292b3cd1295f38f62	23.55 C	M	Mathematics	
58c274a03eb5b172897c13e209d162b	24.13 C	M	Civic	
d73db0840f8452e20001863997fb5a	25.7 C+	F	Civic	
f30a59bdf6516777b67cfb5d68c0bee1	19.78 D	M	Geography	
lb872461c4668ab02126d905c9ad950d0	28.13 C+	M	Mathematics	

CSC 5741 L06 - 11

Data Attributes (2/5)

- Categorical attributes**
 - Discrete set of values
 - Only one value can be held at a time
 - Categories are mutually exclusive
 - The only numeric operation performed is equality testing

A	B	C	D	E
StudentID	CA	CA Grade	Gender	Minor
3283f07fce04d7ce5cd5fd6d191bf	4.78 D	M	Civic	
3aa3919efac1157cc4a2529c3f391fe	23.7 C	M	Languages	
8e5435d4f634804d12512b6caeab4316	25.13 C+	M	Civic	
64e4a90e4dfa32a2877c93d48354c81	17.48 D	M	History	
89407737ae371e6080e88ad29ddd07bd	19.6 D	M	Civic	
b34e7468b181868a48635c0bb376e	32.6 B	M	Civic	
4db474b69680eadalc7d0862e99041	24.43 C	M	History	
a0d3784ea0902162d676fb0ae1fb6c36	20.8 D+	M	Mathematics	
c463632dc537747e3925bc8e4038ad42	26.65 C+	M	Mathematics	
f8b7ffcf26cde897a3ad67a83c99286d	30.55 B	M	Mathematics	
a96fb28013fb01358acfe2e85f598d	25.2 C+	F	Languages	
cad8f19590f7a666094bf5fa3d4c0633	19.6 D	F	History	
94b1c4e2144e735b872faa1078577aca4	25.45 C+	M	Mathematics	
e8c074808b39c3b3e345f4a774696590	25.28 C+	F	Civic	
5e1c7620b6f16173292b3cd1295f38f62	23.55 C	M	Mathematics	
58c274a03eb5b172897c13e209d162b	24.13 C	M	Civic	
d73db0840f8452e20001863997fb5a	25.7 C+	F	Civic	
f30a59bdf6516777b67cfb5d68c0bee1	19.78 D	M	Geography	
lb872461c4668ab02126d905c9ad950d0	28.13 C+	M	Mathematics	

April 30 2019

CSC 5741 L06 - 12

Data Attributes (3/5)

- **Ordinal attributes**
 - Similar to categorical attributes except that they exhibit a natural order
 - Likert scales
 - Can be encoded as numbers

Student ID	Experience Using Computers	Do you
c4676143b01dd8a53e2d5ea2ee2413db eedfb073ca1dc81ae0420ea3aab81	More than 5 years 1 to 2 years	Yes Yes
98af912970920020748a46fe3b4409fd 4fee5b98c471243d43ec32966d605f8	No Experience Less than 1 year Less than 1 year	Yes Yes No
aa03aa4ad6229df904a49589659703c aa5f685006faa22926896930b7522349	No Experience No Experience No Experience	Yes Yes Yes
5f685006faa22926896930b7522349 45f685006faa22926896930b7522349	No Experience No Experience No Experience	Yes Yes Yes
3732b39a1e12bbc7b4040ade4ed61e1413d tf11c00713c027695a2b433d21e1413d	Less than 1 year More than 5 years	Yes Yes
982a9a3c301db51b12d16d703ba3b 572da2e52fc62a31a9c939966ab	No Experience More than 5 years	Yes Yes
572da2e52fc62a31a9c939966ab sd213be6bc881b1a6217b054870	More than 5 years No Experience	Yes No
058bc71ce93f598c06031ff01acf5 8d2899dc1a99aa53724be4f357fd65f7d	1 to 2 years More than 5 years	No Yes
3fe634efc9967ab9a83ra8faeb110d 8a0cc0506c959e79110e0d7893834c	No Experience Less than 1 year	Yes No
11e18906330f2232926f199b2d65f0d 3975b445569d00942f213b1bd5ee5	No Experience 1 to 2 years	Yes No
021716971e565917e44917130999b LL00004114L4LF-1-7-10FFC7-71A04	1 to 2 years	Yes

April 30 2019

CSC 5741 L06 - 13

Data Attributes (4/5)

- Data such as images can easily be represented using individual pixels
 - Each pixel represents a distinct attribute



April 30 2019

<http://lloydbleekcollection.cs.uct.ac.za>

CSC 5741 L06 - 14

Data Attributes (5/5)

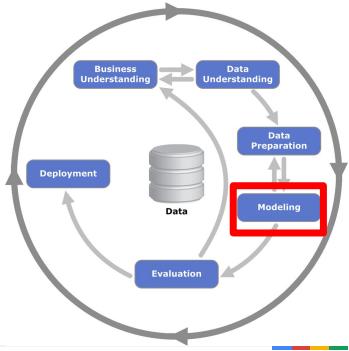
- Textual content could easily be represented using individual terms forming the text
 - Bag-of-words model
 - A numeric variable can be used to signal the relative importance of the word in the document

April 30 2019

CSC 5741 L06 - 15

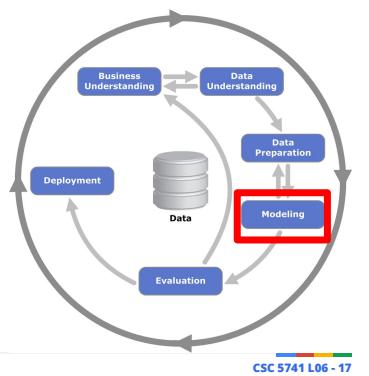
ML Techniques: Supervised Learning (1/8)

- **Supervised learning**
 - Involves the use of labeled data
 - The labels are the resulting output of the algorithm
 - Training is required to teach the algorithm
 - The goal is to predict a specific quantity
 - Accuracy can be measured directly



ML Techniques: Supervised Learning (2/8)

- Predicting an output $f(x)$ given an input x
 - If $f(x)$ is categorical, this would be a Classification problem
 - If $f(x)$ is numerical and/or continuous, this would be a Linear Regression problem



April 30 2019

CSC 5741 L06 - 17

ML Techniques: Supervised Learning (3/8)

Data-Driven Intervention-Level Prediction Modeling for Academic Performance

Mvurya Mgala
Dept of Computer Science
University of Cape Town
HPI School in ICT4D, 7701 Cape Town
mmgala@cs.uct.ac.za

Audrey Mbogho
Dept of Computer Science
University of Cape Town
HPI School in ICT4D, 7701 Cape Town
audrey.mbo@uct.ac.za
Paper Reading Assignment #01

- $f(x)$ —High Intervention/Low Intervention
- x —Test scores, gender, age, student motivation, study time
- Labels—Manually done perhaps?
- Type of learning: Classification

April 30 2019

CSC 5741 L06 - 18

ML Techniques: Supervised Learning (4/8)

Document Type Classification in Online Digital Libraries

Cornelia Caragea,¹ Jian Wu,² Sujatha Das Gollapalli,³ and C. Lee Giles²
¹Department of Computer Science and Engineering, University of North Texas, Denton, TX

²College of Information Sciences and Technology, Pennsylvania State University, University Park, PA
³Institute for Infocomm Research, A*STAR, Singapore

caragea@unt.edu, jxw394@ist.psu.edu, gollapallis@i2r.a-star.edu

Paper Reading Assignment #02

- $f(x)$ —Book/Slides/Thesis/Paper/Resume
- x —File-specific, text-specific, section-specific, containment-specific
- Labels—Manually done
- Type of learning: Classification

April 30 2019

CSC 5741 L06 - 19

ML Techniques: Supervised Learning (5/8)

Moodle Predicta: A Data Mining Tool for Student Follow Up

Igor Moreira Félix¹, Ana Paula Ambrósio¹, Priscila Silva Neves¹,
Joyce Siqueira¹ and Jacques Duilio Brancher²

¹Instituto de Informática, Universidade Federal de Goiás, Goiânia, Brazil

²Departamento de Computação, Universidade Estadual de Londrina

Paper Reading Assignment #03

- $f(x)$ —Students “at risk”
- x —Behaviour and interaction within The Moodle (course, posts, messages, time spent on activities)
- Labels—Manually done perhaps?
- Type of learning: Classification

April 30 2019

CSC 5741 L06 - 20

ML Techniques: Supervised Learning (6/8)

- $f(x)$ —Spam/Not Spam
- x —Email address, subject, email content
- Labels—Mailbox user tagging; Gmail automatic detection
- Type of learning: Classification

April 30 2019 CSC 5741 L06 - 21

ML Techniques: Supervised Learning (7/8)

- $f(x)$ —BnB Amount for specific month next year
- x —USD rate; Rand rate; political climate; ????
- Labels—Already available
- Type of learning: Regression

April 30 2019

Latest JCTR Documents for 2019

- [CLICK HERE to read and download latest Documents](#)
- [Job Alerts Programme Manager and Finance Assistant.pdf](#)
- [Download File](#)
- [KNB Press Statement - March 2019](#)
- [Download File](#)
- [Leads BNS - March 2019](#)
- [Download File](#)
- [Credibility - Strategic Plan, 2019-2022](#)
- [Download File](#)
- [Job Statement For 2019](#)
- [Download File](#)
- [Statement - National Values & Principles Speech](#)
- [Download File](#)
- [2019 National Values Original Speech](#)
- [Download File](#)
- [National on Job Opportunities in the GCF](#)
- [Download File](#)
- [National Social Service Delivery](#)
- [Download File](#)
- [MoE-Zambia Law Development Bill](#)

Fr. Emmanuel Mumba S.J.
Executive Director, Joint Committee for Technical and Vocational Education and Training (JCTR)

Addressing Zambia's Economic Challenges

April 30 2019 CSC 5741 L06 - 22

ML Techniques: Supervised Learning (8/8)

- $f(x)$ —Expected Salary
- x —qualifications, domain, experience, references
- Labels—?????
- Regression

Keywords Location

Choose a category... Advert Consultancy Contract Full Time Internship Part Time
 Temporary

Advert Posted 2 weeks ago Closes: April 30, 2019

UNICAF Zambia

AutoCAD Draughtsman Lubambe Copper Mine Copperbelt, Zambia

Full Time Posted 29 mins ago Closes: May 8, 2019

Document Controller Lubambe Copper Mine Copperbelt, Zambia

Full Time Posted 30 mins ago Closes: May 8, 2019

Sales Analyst Ndola, Zambia

Full Time Posted 30 mins ago Closes: May 8, 2019

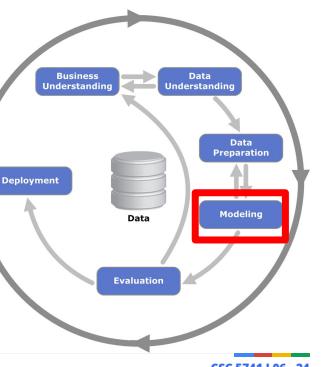
<https://gozambiajobs.com>

CSC 5741 L06 - 23

ML Techniques: Unsupervised Learning (1/2)

- **Unsupervised learning**
 - No specific value to be predicted
 - The goal is pattern recognition, in order to learn more about the data
 - No labelled is required
 - No training required
 - Evaluation is typically qualitative since there are no predefined labels
 - Typically subjective

April 30 2019



ML Techniques: Unsupervised Learning (2/2)

- $f(x)$ —WE DO NOT KNOW
- x —Institution, ETD title, ETD abstract
- Labels—No labels
- Clustering

The screenshot shows the NDLTD Union Archive homepage. It features a header with the NDLTD logo and a search bar. Below the header, there are sections for "INFORMATION", "Recent Submissions", and "Collection Statistics". The "Recent Submissions" section lists several academic papers. The "Collection Statistics" table provides a breakdown of the total number of records by institution.

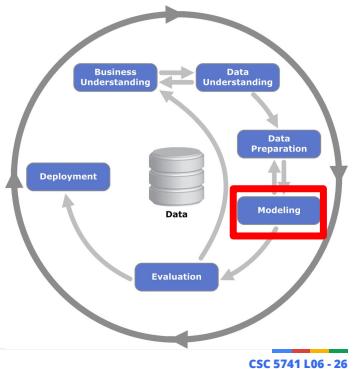
Collection	Total
GRANT-ED	22
URBOLCGMA	0
UDC	6755
Australian Institute University	7772
Atlanta University Center	4392
Australasian Digital Theses Program	59958
Banque de France	10
Bibliothèque Interuniversitaire de la Côte d'Ivoire	10
Bruxelles, Belgique	10

At the bottom of the page, there is a footer with the URL <http://union.ndltd.org> and the identifier CSC 5741 L06 - 25.

April 30 2019

ML Techniques: Reinforcement Learning

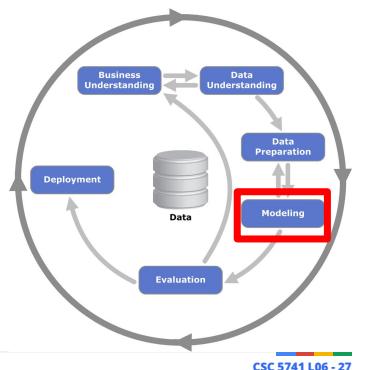
- Semi-supervised/reinforcement learning
 - Uses a combination of supervised and unsupervised learning
 - Unsupervised learning techniques are typically used to improve supervised learning algorithms
 - Example: Scenario where there is fewer labeled data



April 30 2019

Machine Learning Techniques

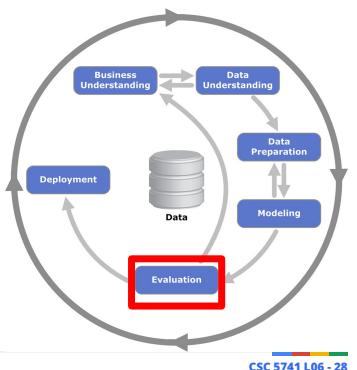
- We will focus on algorithms for solving the following problems
 - Regression problems
 - Classification problems
 - Clustering problems



April 30 2019

Evaluation (1/9)

- Evaluation is a systematic determination of a subject's merit, worth and significance, using a specified criteria.
 - It assess an entity to help in decision making; or to ascertain the degree of achievement or value in regard to the aim and objectives

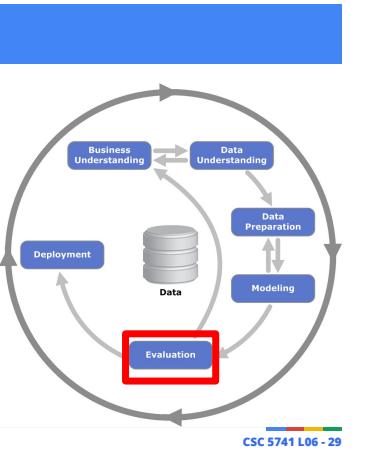


April 30 2019

CSC 5741 L06 - 28

Evaluation (2/9)

- Evaluation is systematic and rigorous.
- Evaluation involves critical assessment of a given set of objectives.
 - How effective is Lighton Phiri at teaching CSC 5741?

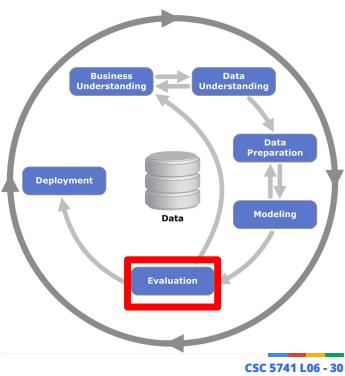


April 30 2019

CSC 5741 L06 - 29

Evaluation (3/9)

- Evaluation forms a crucial part of Machine Learning as it assesses the relative effectiveness of learning
 - How **efficient** is the learning process?
 - Computing resources are finite [...]
 - How **effective** is the learning outcome?
 - Accuracy is key [...]

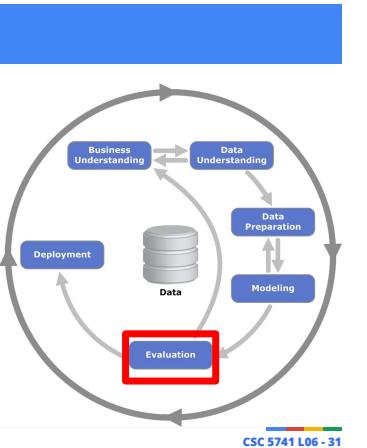


April 30 2019

CSC 5741 L06 - 30

Evaluation (4/9)

- Effectiveness—The extent towards which an ML model is successful at accomplishing its intended objectives.
 - E.g. How accurate is it at classifying at risk students?
- Efficiency—The relative cost implications of an ML model achieving its objectives
 - How long does it take to present the output



April 30 2019

CSC 5741 L06 - 31

Evaluation (5/9)

```
-rw-rw-r-- 1 lightonphiri lightonphiri 4.8M Apr 21 13:31 union_ndlts.org-zajlis-20190420-old_oai_script-batch-0
-rw-rw-r-- 1 lightonphiri lightonphiri 7.4M Sep 14 2018 2019New_Mesh_Tree_Hierarchy.txt
-rw-rw-r-- 1 lightonphiri lightonphiri 8.1M Apr 18 00:00 dspace.unza.zm-20190417-2.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 12M Apr 21 13:38 nb-ir_reclassification.invn
-rw-rw-r-- 1 lightonphiri lightonphiri 63M Apr 30 13:22 paper-eto19-etc_analysis.ipynb
-rw-rw-r-- 1 lightonphiri lightonphiri 1.2G Apr 29 21:21 var_pubmed_mesh_baseline_files.pkl
-rw-rw-r-- 1 lightonphiri lightonphiri 1.3G Apr 29 22:20 var_pubmed_mesh_baseline_clean.pkl
lightonphiri@lightonphiri:~/Lenovo-IdeaPad-Yoga-530-14IKB~$ cd /var/www/html/ml-classification/scripts$
```

- Efficiency is concerned with execution speed and relative usage of computing resources
 - How much RAM is required to build the models?
 - How much processing power is required to build the models?
 - How long does it take to build the model?
 - How much storage space is required?
 - Cost implications?

April 30 2019

CSC 5741 L06 - 32

Evaluation (5/9)

High Performance Computing @ ZAMREN

High Performance Computing(HPC) is an aggregation or clustering of computing power in a way that yields greater performance than could be obtained from a typical workstation or server.

This is done to solve large problems in science, engineering, or business. HPC enables computation and analysis of vast data, for example, in areas such as Water, Energy and Environment, Materials Engineering, Nuclear Physics, Genetics, Neurology, Astrophysics, Bio-informatics, Geosciences, Visualization and Imaging, among the numerous types science and mathematical analyses. They are also intensely used in product development, redesigns and process optimization.

This service at ZAMREN provides an opportunity for our researchers and students to undertake science driven research and be part of the global research communities in their respective research categories and indeed create innovative solutions for social and economic development.

<http://hpc.zamren.zm>

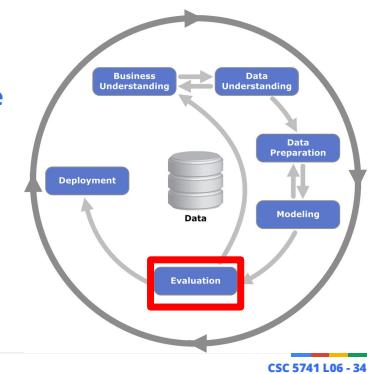
April 30 2019

CSC 5741 L06 - 33

Evaluation (6/9)

- Effectiveness of ML models involves measurement of accuracy in order to determine the relevance of the results

- Recall—The total number of relevant results returned.
 - $\text{Recall} = \frac{\text{relevant retrieved}}{\text{total relevant}}$



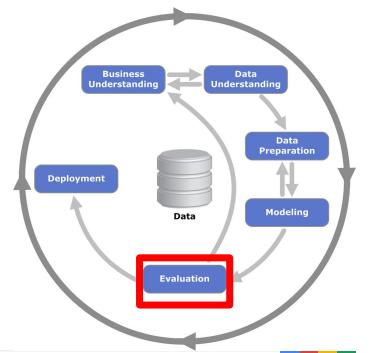
April 30 2019

CSC 5741 L06 - 34

Evaluation (6/9)

Recall

- E.g. If the UNZA institutional repository has a total of 100 documents relevant to a query on "Information Retrieval" and only 70 are retrieved when a query related to IR is issued, then the recall is 70%



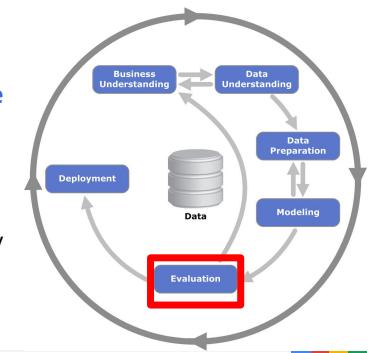
April 30 2019

CSC 5741 L06 - 35

Evaluation (7/9)

- Effectiveness of ML models involves measurement of accuracy in order to determine the relevance of the results

- Precision—The number of returned results that are relevant.
 - $\text{Precision} = \frac{\text{relevant retrieved}}{\text{total retrieved}}$



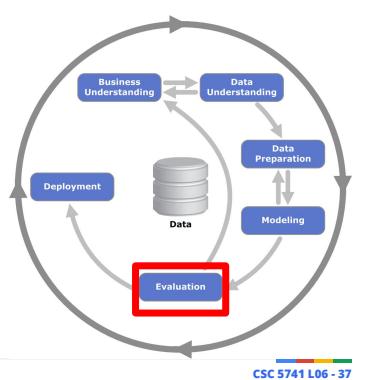
April 30 2019

CSC 5741 L06 - 36

Evaluation (7/9)

- Precision

- E.g. If a search of "Information Retrieval" in the UNZA institutional repository retrieves 100 documents and 40 of those documents are relevant, the precision is 40%



April 30 2019

CSC 5741 L06 - 37

Evaluation (7/9)

false negatives

true negatives

true positives

false positives

April 30 2019

How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

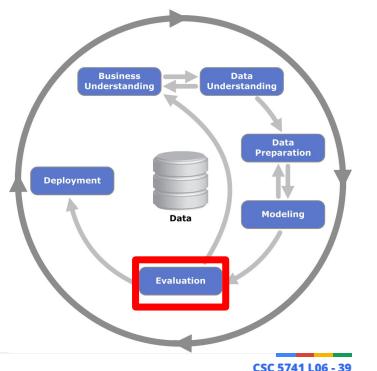
<https://www.wikipedia.org>

CSC 5741 L06 - 38

Evaluation (8/9)

- Effectiveness of ML models involves measurement of accuracy in order to determine the relevance of the results

- F1-score provides a comprehensive measure of a test's accuracy.
 - It considers both the precision p and the recall r



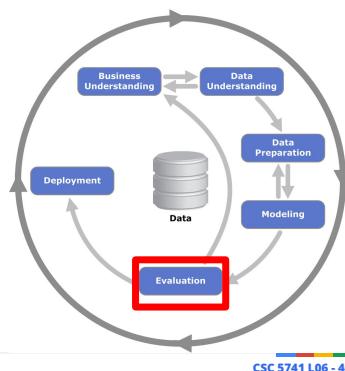
April 30 2019

CSC 5741 L06 - 39

Evaluation (9/9)

- The F1 score conveys the balance between the precision and the recall.

$$\text{F1 Score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

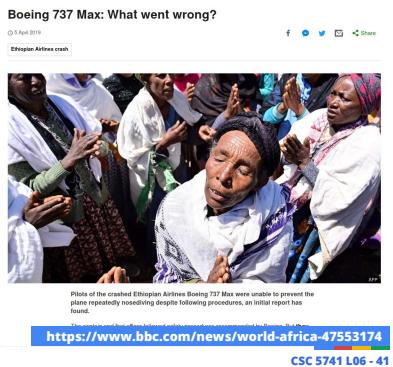


April 30 2019

CSC 5741 L06 - 40

Implications of Evaluation (1/3)

- Careful attention needs to be placed on evaluation of learning algorithms, especially for sensitive domains
 - Domains with safety concerns

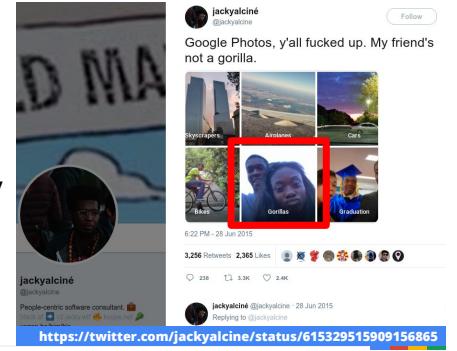


April 30 2019

CSC 5741 L06 - 41

Implications of Evaluation (2/3)

- Careful attention needs to be placed on evaluation of learning algorithms, especially for sensitive domains
 - Domains where socially accepted norms are compromised

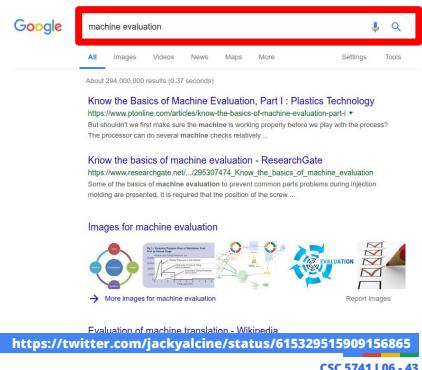


April 30 2019

CSC 5741 L06 - 42

Implications of Evaluation (3/3)

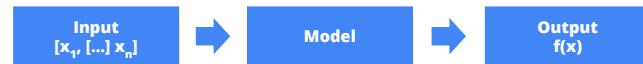
- Some application domains are flexible
 - E.g. wrong search results are generally acceptable



April 30 2019

CSC 5741 L06 - 43

Model Evaluation Process



- A common evaluation technique for supervised learning involves the use of labeled data, split into training and testing datasets
 - Model uses training dataset to train estimators
 - Model uses testing dataset to determine effectiveness of predictions
 - Model evaluation involves determining the proportion of accurate predictions relative to the training set

April 30 2019

CSC 5741 L06 - 44

Model Testing Techniques (1/3)

- The holdout method reserves a representative proportion of the dataset as testing data
 - While there is not prescribed training/testing ration, 80/20, 90/10 and 70/30 ratios are common
 - The samples might not be representative enough and so stratification might be necessary



April 30 2019

CSC 5741 L06 - 45

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

CSC 5741 L06 - 46

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

CSC 5741 L06 - 47

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

CSC 5741 L06 - 48

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (2/3)

- K Fold cross-validation is a more viable alternative to the holdout method**
 - Dataset is randomly split up into "k" equal groups
 - One of the groups (fold) is used as the test set and the rest are used as the training set
 - The model is trained on the training set and scored on the test set.
 - Then the process is repeated until each unique fold has been used as the test set.
 - In 10-fold cross-validation, the dataset is split into 10 groups



April 30 2019

Model Testing Techniques (3/3)

- Leave-one-out is an extreme version of K Fold cross-validation**
 - It is essentially n-fold cross-validation where n = number of data points
 - Each instance is predicted, while training on the remaining (n-1) instances
 - Very comprehensive
 - Computationally intensive
 - The balance of training and testing sets is compromised



April 30 2019

Model Evaluation Techniques (1/3)

- The confusion matrix is commonly used during classification
 - Actual labels are compared against predictions to determine the number of True Positives, False Positives, True Negatives and False Negatives

		Actual Classes	
		Pass	Fail
Predicted Classes	Pass	60	5
	Fail	3	10

April 30 2019

CSC 5741 L06 - 57

Model Evaluation Techniques (2/3)

- Complex non-binary classification problems are sometimes easily interpreted using the confusion matrix

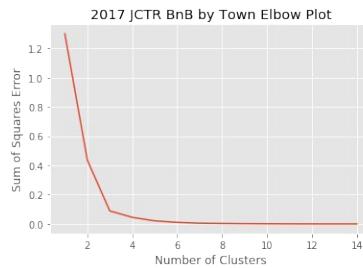
ED	MD	NS	AG	HS	LW	VM	EG	MN	LB	ID
140	2	0	0	15	0	0	0	0	0	0
3	144	0	0	4	0	0	0	0	0	0
4	8	11	1	20	0	0	0	0	0	0
1	5	0	21	10	0	0	0	0	0	0
18	12	1	0	116	1	0	0	0	0	0
1	1	0	0	39	0	0	0	0	0	0
0	19	0	0	1	0	0	0	0	0	0
0	0	1	1	16	0	0	0	0	0	0
0	3	4	0	12	1	0	0	0	0	0
5	1	0	0	4	1	0	0	0	0	0
11	0	0	0	3	0	0	0	0	0	0

April 30 2019

CSC 5741 L06 - 58

Model Evaluation Techniques (3/3)

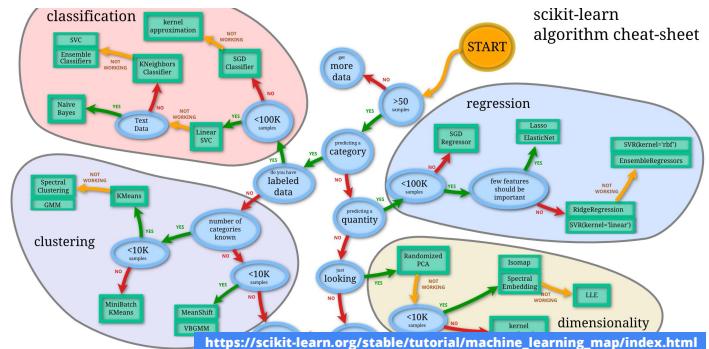
- Unsupervised learning approaches use specific evaluation techniques
 - For instance with K Means clustering, the elbow method is commonly employed to determine the appropriate number of clusters
 - Remember that no training is involved during unsupervised learning



April 30 2019

CSC 5741 L06 - 59

Summary (1/2)



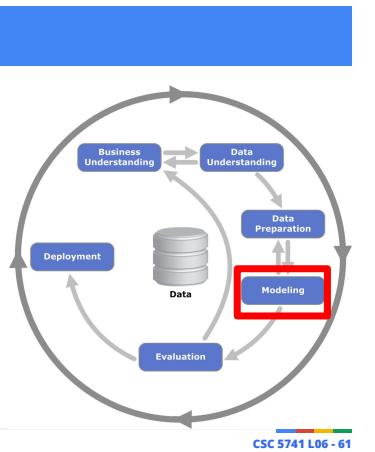
April 30 2019

CSC 5741 L06 - 60

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Summary (2/2)

- We will attempt to answer the following questions:
 - What does the approach/technique/estimator/algorithm do?
 - What are the inputs: x ?
 - What are the outputs: $f(x)$?
 - How do we evaluate the approach/technique/estimator/algorithm?



Q & A Session

- Comments, concerns and complaints?

April 30 2019

CSC 5741 L06 - 62

Lecture Series Outline

- Part I: Machine Learning
- Part II: Datasets
 - Scikit-learn Standard dataset
 - JCTR Basic Needs Basket
 - University of Zambia Electronic Theses and Dissertations
 - B.ICTs Ed. ICT 1110 Continuous Assessment Scores
 - Jupyter Notebook Walkthrough

April 30 2019

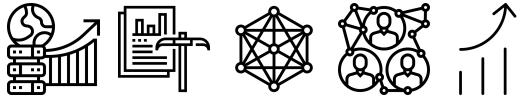
CSC 5741 L06 - 63

Bibliography

- [1] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2017) Data Mining: Practical Machine Learning Tools and Techniques. Chapter 2
<https://www.cs.waikato.ac.nz/ml/weka/book.html>
- [2] An introduction to machine learning with scikit-learn
<https://scikit-learn.org/stable/tutorial/basic/tutorial.html>

April 30 2019

CSC 5741 L06 - 64



CSC 5741

Lecture 6: Introduction to Machine Learning



Lighton Phiri <lighton.phiri@unza.zm>

Department of Library and Information Science
University of Zambia