Evaluation of Machine Learning Courses for Data Science Curriculum in American Higher Education

Shilpa Balan
College of Business and Economics, Department of Information Systems
California State University, Los Angeles
sbalan@calstatela.edu

Divya Pakhale
College of Business and Economics, Department of Information Systems
California State University, Los Angeles
dpakhal@calstatela.edu

Neha Shashidhara Guli
College of Business and Economics, Department of Information Systems
California State University, Los Angeles
sneha@calstatela.edu

Abstract

Recent progress in machine learning has been driven by the development of new algorithms and low-cost computation. The adoption of machine learning methods can be found in science, technology, and commerce including health care, manufacturing, education, finance, and marketing. In this epoch of data revolution, it has become important to examine machine learning and its influence in higher education. This paper examines initiatives that American universities have taken to implement machine learning as part of their data science educational curriculum. This paper also explores popular categories of courses offered by universities as part of their machine learning curriculum. Previous studies have shown that there is currently a shortage of data scientists. This study finds that many American universities are now offering machine learning courses to train students with the necessary skills to meet this shortage.

Introduction

Machine learning is a rapidly growing field at the intersection of computer science and statistics and is an integral component of artificial intelligence and data science. The evolution of data has driven recent progress in machine learning, which has been further driven by the current low cost of computation.

Information is now the most valuable commodity in multiple fields. Knowledge gained via data extraction is extremely useful in today's competitive environment. Machine learning techniques are part of the process of knowledge extraction and are fundamental in the data mining process. Data mining is considered to be a key component in knowledge discovery (Paralič et al., 2003). The process of knowledge discovery is crucial for sharing obtained knowledge for effective decision-making (Brachman et al., 1996, Fayyad et al., 1993, Han and Kamber, 2000).

The processing of huge amounts of data includes methods and techniques of data mining, and the design of prediction models. These skills are necessary for business analytics and data science. To master these techniques, it is necessary to have sufficient mathematical and statistical knowledge, as well as computing skills. This knowledge is essential for the skillset of a data scientist. Therefore, teaching these skillsets at universities as part of the data science program is essential. Previous studies have shown that 81% of companies with analytical talent claimed that business analytics creates a competitive advantage (Ransbotham, 2015). Therefore, it is important to examine machine learning and its influence in higher education.

Data science is experiencing rapid growth, urged by the rise of complex and rich data in science, industry, and government. The McKinsey report (McKinsey Global Inst., 2011) forecasted a need for hundreds of thousands of data science jobs over the next decade. This has led to an explosion in the number of data science programs being offered in academics as universities are now rushing to meet this demand. This paper examines initiatives that American universities have taken to implement machine learning as part of their data science curriculum. This paper also explores popular categories of courses offered by universities as part of their data science and machine learning curriculum.

The background section of this paper describes the importance of teaching machine learning in higher education as part of data science. The methodology used for the study is outlined in the research methodology section. The results section details the results of the study, which are then examined in the discussion section.

Research Background

Machine learning is a multidisciplinary area involving artificial intelligence, statistics, information theory, and psychology. The goal of machine learning is to solve real-world problems using models that provide good data approximations (Dhage and Raina, 2016).

Machine learning has progressed over the past few decades. Machine learning is a part of artificial intelligence that emerged as a method to develop applications such as computer vision, speech recognition, natural language processing, and robot control to

name a few (Jordan and Mitchell, 2015). The influence of machine learning has also been seen in industries concerned with data-centric issues, such as consumer services, the diagnosis of faults in complex systems, and logistic chains (Jordan and Mitchell, 2015).

There is currently a shortage of workers with data mining skills. In 2018, the demand for deep analytical skills in the United States is projected to be 50% to 60% higher than the supply (Bucko et al., 2017). The use of machine learning methods is currently necessary in all fields. Machine learning skills are seen to be a competitive advantage for students when applying for jobs (Bucko et al., 2017). In our paper, we review the design of data science curriculum with a focus on machine learning techniques.

Currently, there are several machine learning methods that can be integrated into the study plans of universities in data science (Dhage and Raina, 2016). In terms of university courses, we include methods that are supported by proper software and are useful for industry practice.

While previous studies have shown that the teaching process for machine learning includes artificial intelligence, computer science, and information science (Someren, 2016), we were unable to find any papers describing the teaching process for machine learning curriculum in the United States. Further, we observed that most universities in the United States offer their machine learning courses as part of a data-mining course, possibly because machine learning is essentially mining data (Lynch, 2018). We detail these findings in the results section.

In a viewpoint given by Baker et al. (2009), data mining is categorized as follows: prediction (classification and regression); clustering; relationship mining (association rule mining, correlation mining, sequential pattern mining, and causal data mining); and discovery with models.

Further, Romero and Ventura (2007) categorize work in educational data mining into the following categories: statistics and visualization, and web mining, i.e., clustering, classification, and outlier detection; association rule mining and sequential pattern mining; and text mining. In this context, previous research on analysis in education has been performed using machine learning. Duzhin and Gustafsson (2018) describe a machine learning based app they created to compare clickers and traditional handwritten homework. It was found that clickers were more effective than traditional handwritten homework.

As seen by the above classified data mining categories, machine learning topics including prediction are incorporated into data mining. This is further supported by our results shown in the results section.

Research Methodology

To compare machine learning software used by various universities in the United States, we identified several important keywords. For example, we used the keyword ‘machine learning tools’ and looked at a list of the top universities in the United States that showed up in a Google search. As of September 2018, the keyword ‘machine learning certification’ generated 170,000,000 results in a Google search.

In the results section, we show the top machine learning software retrieved in our search results using the search discussed above. We also show the popular categories of courses offered by different universities in the United States as part of their machine learning course curriculum.

Results

We used the keyword 'machine learning tools', which generated 442,000,000 results in a Google search as of September 2018. We selected the top results from this search. With the help of the top Google search results and those from KDNuggets (2015) we created a table of machine learning tools frequently used by various universities (Table 1). In the table, we see that Weka, MATLAB, Octave, SAS, and TensorFlow are the top five tools used for machine learning courses at American universities.

Further, Figure 1 shows the results of a search for the keywords '*software name* machine learning course'. For example, a search for 'Weka machine learning course' generated 197,000 results as of September 2018. Similarly, we searched for other machine learning software in Table 1 to determine the ranking of their popularity of use at universities. We found that SAS is the most popularly used machine learning and data mining software, followed by MATLAB and then TensorFlow.

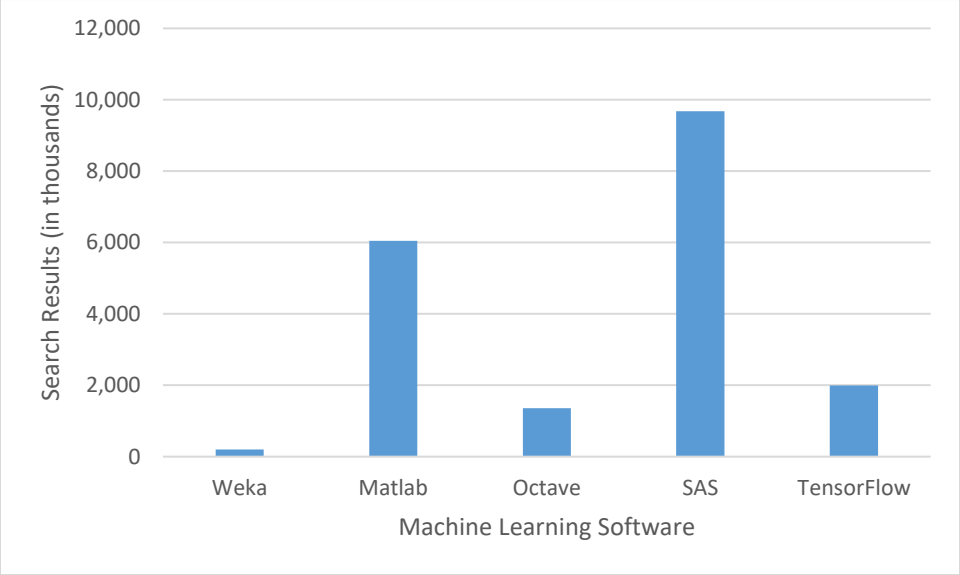

Figure 1: Search results for '*Software name*_Machine Learning course' as of September 2018.

Table 1. Data mining and machine learning tools implemented in data science curriculum at American universities.

Serial Number	Name of Tool	Benefit of the Tool	Source for Additional Information
1.	Weka	Used in machine learning for data mining tasks. This tool was developed at the University of Waikato in New Zealand.	https://www.cs.waikato.ac.nz/ml/weka/
2.	MATLAB	Used for predictive maintenance, sensor analytics, finance, and communication electronics.	https://www.mathworks.com/solutions/machine-learning.html
3.	Octave	Used as a high-level language that is primarily intended for numerical computations.	https://www.gnu.org/software/octave/about.html
4.	SAS	Supports the data-mining and machine-learning process with a visual and programming interface that handles all tasks in the analytical life cycle.	https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html
5.	TensorFlow	An open source software library for high performance numerical computation that was originally developed by researchers and engineers from the Google Brain team within Google's AI organization.	https://www.tensorflow.org/

Table 2 describes popular categories of courses offered by universities in the United States as part of their data science curriculum. We searched for 'machine learning courses' using the Google search engine and compiled a list of popular courses offered as part of the machine learning courses at various American universities. The range of

courses listed in Table 2 are offered by several universities in the United States, such as Columbia University, Harvard, University of Arizona, University of Texas at Austin, Stanford University, and Northwestern University. Table 2 shows that the common courses offered for data mining and machine learning by universities are courses in R, data mining for business, decision analytics, and data warehousing. Note that few universities offered machine learning courses with course titles such as 'Machine Learning for Big Data'.

Table 2. Common data science courses taught at various universities in the United States.

Serial number	Course Name	Brief Course description
1.	Core Statistics Using R	The structure of this course includes statistical analyses such as logistic regression performed in the R language.
2.	Data Mining for Business	This course includes an introduction to data mining algorithms such as neural networks, decision trees, discriminant analysis, and association analyses to extract hidden information in data.
3.	Decision Analytics	This course includes an overview of models available to analyze decision problems for finance, marketing, and operations.
4.	Data Warehousing	This includes course data used to support managerial decisions, developing data warehouses, which includes ETL (extraction, transformation, and loading), and dimensional modeling.

Discussion

The field of machine learning is still rapidly expanding. Practical applications have driven the invention of new formalizations of machine learning problems (Jordan and Mitchell, 2015). The adoption of machine learning methods can be found in science, technology, and commerce, including health care, manufacturing, education, finance, and marketing. Table 3 lists popular machine learning applications in healthcare, manufacturing, and finance.

Table 3. Applications of machine learning.

Serial No.	Industry	Applications
1	Healthcare	• Diagnoses in medical imaging • Drug discovery
2	Finance	• Portfolio management • Algorithmic trading • Fraud detection
3	Manufacturing	• Optimizing semiconductor devices • Quality control • Perfecting the supply chain

Computer-aided detection and diagnoses performed using machine learning algorithms can assist physicians to interpret medical-imaging findings (Schoepf and Costello, 2004). Machine learning is also used in drug discovery (Wale, 2010). Further, machine learning has benefited the field of finance, as analysts have used advanced mathematical and statistical models to determine the relationships between the future values of a stock price and its fundamental quantities (Nuti et al., 2011).

Conclusion

Undoubtedly, big data and analytics have become a popular topic of discussion (IBM, 2015). The demand for skilled workers in big data is growing as the amount of data is rapidly increasing. In 2011, McKinsey (2011) estimated that there would be a need for several scientists with deep analytic skills, to support better decision-making. A growing number of students obtaining a Masters degree in analytics program will address the gap for these analytic skills.

Data scientists employ mathematical models. This necessitates that data scientists have a strong foundation in mathematics in addition to having an understanding of basic statistical theory (Veaux et al., 2017). The recent growth of statistics programs is impressive. While still small in absolute numbers, such programs nearly doubled between 2010 and 2013 (Wasserstein, 2015). Data science graduates should also be skillful in various software skills. Data science offers the opportunity to integrate both computational and statistical courses to solve problems rather than emphasizing one over the other (Veaux et al., 2017).

As more universities move toward developing programs in analytics, there are different perspectives of how to best prepare students to meet the industry demand for data analysts/scientists with advanced skills: some offer analytics certificates or specializations in an MBA program, while others are opting for an MS degree in analytics (Gupta et al., 2015).

From the results section, we can conclude that universities in the United States are beginning to explore programs in data science and machine learning and beginning to implement data analytics and machine learning tools that are frequently used in industry.

References

Baker, Ryan, and Kalina Yacef. "The State of Educational Data Mining in 2009: A Review and Future Visions." *Journal of Educational Data Mining*, vol. 1, no. 1, 2009, pp.3-16.

Brachman, Ronald J., and Tej Anand. "The Process of Knowledge Discovery in Databases." *Advances in Knowledge Discovery & Data Mining*, edited by Usama Fayyad, et al., AAAI/MIT Press, 1996.

Bucko, J., and L. Kakalejčík. "Machine learning techniques in the education process of students of economics," *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2017, pp. 750-754.
URL: http://ieeexplore.ieee.org.mimas.calstatela.edu/stamp/stamp.jsp?tp=&arnumber=7973522&isnumber=7973374

De Dhage, Sandhya N., and Charanjeet Kaur Raina. "A review on Machine Learning Techniques." *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 4, no. 3, 2016, pp. 395-399.

Duzhin, F., and Gustafsson, A., "Machine Learning-Based App for Self-Evaluation of Teacher-Specific Instructional Style and Tools." *Educ. Sci.* 2018, *8*, no. 1, 7.

Fayyad, Usama, and Keki B. Irani. "Multiinterval discretization of continuous-valued attributes for classification learning." *IJCAI*, pp. 1022-1027, 1993.

Gupta, Babita, Michael Goul, and Barbara Dinter. "Business Intelligence and Big Data in Higher Education: Status of a Multi-Year Model Curriculum Development Effort for Business School Undergraduates, MS Graduates, and MBAs." *Communications of the Association for Information Systems*, vol. 26, no. 23, 2015.

Han, Jiawei, Micheline Kamber, and Jian Pei. *Data Mining - Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

Horton, Nicholas J., and Johanna S. Hardin. "Teaching the next generation of statistics students to "Think with Data": Special issue on statistics and the undergraduate curriculum." *The American Statistician*, vol. 69, no. 4, 2015, pp. 259-265.

Jones, Alex. "Top 10 Data Analysis Tools for Business." *KDNuggets*, 2015, http://www.kdnuggets.com/2014/06/top-10-data-analysis-tools-business.html. Accessed 24 December 2015.

Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science*, vol. 349, no. 6245, 2015, pp. 255-260. http://science.sciencemag.org.mimas.calstatela.edu/content/349/6245/255

KDNuggets. “Education in Data Mining, Analytics and Data Science in USA/Canada.” 2015, http://www.kdnuggets.com/education/usa-canada.html. Accessed 14 November 2015.

Lin, Hsuan–Tien, Malik Madgon-Ismail, and Yaser S. Abu-Mostafa. “Teaching machine learning to a diverse audience: the foundation-based approach.” *Teaching Machine Learning Workshop at the 25th International Conference on Machine Learning (ICML)*, 2012.

Lynch, Matthew. “8 Ways Machine Learning Will Improve Education.” *The Tech Advocate*, 2018, https://www.thetechedvocate.org/8-ways-machine-learning-will-improve-education/. Accessed on

Manyika, James, et al. “Big Data: The next frontier for innovation, competition and productivity.”2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Accessed on 14 November 2015.

McKinsey Global Inst. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. New York: McKinsey & Co., 2011. http://www.mckinsey.com/business-functions/digital-mckinsey/ourinsights/big-data-the-next-frontier-for-innovation

Nuti, Giuseppe et al. “Algorithmic Trading.” Computer, vol. 44, no. 11, pp. 61-69. https://ieeexplore.ieee.org/document/5696713/

Paralič, J., “Knowledge Discovery in Databases”. Elfa, Kosice, 2003. ISBN 80-89066-60-7.

Ransbotham, Sam, David Kiron, and Pamela K. Prentice. “The talent dividend: Analytics talent is driving competitive advantage at data-oriented companies.” *MIT Sloan Management Review*, 2015. ISSN 1532-9194.

Romero, Cristobal, and Sebastian Ventura. “Educational Data Mining: A Survey from 1995 to 2005.” *Expert Systems with Applications*, vol. 33, no. 1, 2007, pp. 135-146.

Schoepf, U. Joseph, and Philip Costello. “CT Angiography for Diagnosis of Pulmonary Embolism: State of the Art.” *Radiology*, vol. 230, no. 2, 2004, pp. 329-337.

Van der Vlist, Bram, et al. “Teaching machine learning to design students.” *International Conference on Technologies for E-Learning and Digital Entertainment*, Springer Berlin Heidelberg, vol. 5093, 2008, pp. 206-217. ISBN 978-3-540-69736-7.

Van Someren , Maarten. *Teaching Machine Learning at University of Amsterdam*. 12, 2016. https://dtai.cs.kuleuven.be/events/Benelearn2010/submissions/benelearn2010_submission_22.pdf.

Veaux, Richard, et al. “Curriculum Guidelines for Undergraduate Programs in Data Science.” *The Annual Review of Statistics and Its Application*, vol. 4, pp.15-30.

Wale, Nikil. “Machine Learning in Drug Discovery and Development.” *Drug Development Research*, vol. 72, no. 1, 2011, pp.112-119.