1. From Data Mining to Knowledge Discovery in Databases

The purpose of this paper is to provide a holistic view to the workflow of data mining to knowledge discover in databases and its interrelated technologies through applications, methodologies and emergent research in the field.

The data used in this study is a subset of a Loan dataset in which each data point represents a person who has received a loan and are classified as: ‘x’ for defaulting a loan or ‘o’ for being in good status. The paper presents the concepts in a hierarchical manner beginning with the data-mining methods such as classification, regression, clustering, summarization, and dependency modeling. The paper then discusses and identifies components of data-mining algorithms such as model representation, model evaluation, and search. Once these components are identified, the paper goes into detail by enumerating various popular techniques. Some of the techniques discussed are decision trees, nonlinear regression, example-based methods, and probabilistic graphic dependency models.

Through the confluence of AI in KDD, businesses can find value in the area of planning and intelligent agents through the combination of such to automate the collection of data from various sources and integrating it into a planning pipeline for quicker prototyping of strategic initiatives.

1. A few useful thing to know about ML

This paper aims to provide an overview of helpful points and heuristics to acknowledge when practicing and implementing machine learning.

Firstly, the paper organizes key insights or common ‘folk knowledge’, as referred to in the text, at the beginning to build an overview of the idea written. Then, the paper provides greater detail into each key insight, their importance and implementation best practices. The paper then ends on the motivation and objectives of a potential case study to implement the insights gathered throughout.

In this paper, the machine learning workflow is divided into 3 comprehensive components: representation, evaluation, and optimization. After this distinction is made, the paper highlights most common difficulties that arise in implementation. We learned that overfitting is the largest problem in machine learning and can be combatted by solutions like cross-validation and regularization. We also learned that the second greatest problem in machine learning is the curse of dimensionality and can be relieved through dimension reduction or feature engineering.

The information provided by this paper will allow the barrier of entry to decrease for aspiring practitioners and researchers of this field and provide a comprehensive ‘cheat-sheet’. By understanding the key insights and heuristics presented in this paper, one may develop habits of best practices and will lead to cleaner and more insightful projects.

B.

One section of the first paper that stood out to me was the section on probabilistic graphic dependency models. It offered a different perspective to me about how to think of Markov chains through a Bayesian manner. I also enjoyed that the paper had a section specifically to emphasize the importance of having an interdisciplinary mindset during the KDD process as I have always been interested in the interrelation between AI and Biology as a lesson in adaptive agents. The Application Issues section was insightful as well as it was a pretty concise list of all possible paths to pursue in the machine learning domain and try to explore to help the machine learning community.

The 2nd paper provides great insight for beginners and experts alike. I believe the points highlighted and condensed in this paper are ideas that must always be referred back to. It is great to have a paper which creates such a concise guide to the fundamentals of the machine learning workflow. I genuinely enjoyed the insight of ‘Learning = Representation + Evaluation + Optimized” as it offered a different view on how I thought about the process/checklist when thinking of the modeling part of machine learning. Also, this paper allowed me to question some of my basic assumptions of the machine learning process, as well as provide further motivation to correct these assumptions. Due to this paper, I was able to clarify the distinction between feature engineering and dimensionality reduction as I previously thought they were different names for the same thing.