Independent Study Proposal

Title: Implementations of Probability Theory

Student: Andrew Simonson

Faculty sponsor: Thomas Kinsman

INTRODUCTION

The educational focus of Implementations of Probability Theory surrounds the application of data models that produce non-deterministic insights through probabilistic methodology. By pursuing this study I hope to gain a deeper understanding of how to apply data in risk calculation for mitigation scenarios as they appear in real life, rather than the experimental lab conditions that enable algorithmic certainty.

In contrast to the path of black-box artificial intelligence, this study is tailored to methods designed to produce confidence levels for uncertain events using certain terms, leveraging logical, traceable, and definite, calculations. Current course offerings in the realm of data science focus largely on the storing and management of data, and it is noted that the cluster of data science was until very recently under the branding of data management. Implementations of Probability Theory is intended to extend learnings in previous courses, notably CSCI 420: Principles of Data Mining, for more advanced algorithms used at the intersection of data and computing after the preprocessing stage.

PLANNED WORK

Description:

The planned work of this study will be to first review foundational statistics to form a context to consider probabilistic epistemology. This will formalize the importance of methodological completeness and soundness, which is crucial for later units focused on practical model construction. By bridging the logical gap between correlation and causal inference, we avoid misinterpretation of results from poor model selection caused by misunderstandings of the capabilities and risks of automated data models in uncontrolled environments.

Around the 6th week, the focus will shift to implementing probabilistic models. Starting with classic Bayesian models, the study will progress towards developing an open-ended Monte Carlo engine for computing probability mass and probability density functions. The final weeks will concentrate on engaging with new datasets, applying the course content to select appropriate models. This section will also cover refinement methods, experimenting with tools found in artificial intelligence, such as neural networks, to establish a baseline for identifying patterns and building hypotheses for future guided experimentation.

Schedule:

Each week on this schedule accounts for an estimated 10 hours of work, totalling 150 hours. Dept. expectation is 135 hours for 3 credits.

Weeks 1 - 3: Relevant Statistics Review

- Basic concepts: sample space, events, probability axioms

- Law of large numbers and central limit theorem
- Properties of expectation and variance
- Discrete and continuous random variables
- Probability mass function (PMF) and probability density function (PDF)
- Producing Confidence Intervals
- Statistical Inference
- Hypothesis testing in Data Science

Weeks 4, 5: Understanding Probabilistic Theories and Epistemology:

- Decision Theory
- Info-gap Decisions - Robustness to failure under uncertainty
- Ludic Fallacy

Weeks 6 - 8: Bayesian Statistics

- Bayes' theorem
- Bayesian inference: updating beliefs with new evidence
- Prior, likelihood, and posterior distributions
- Potential reading: Dempster-Shafer Theory

Weeks 9 - 11: Markov Methods

- Markov Chains
- Hidden Markov models (HMMs) and their applications in AI

Week 12: Monte Carlo Methods

- Monte Carlo simulation
- Monte Carlo integration
- Markov Chain Monte Carlo (MCMC) methods

Week 13 - 15: Exploring New Probabilistic Datasets

- Demonstration of method tools with practical datasets
- Final Report

Extra Topics for Future Advanced Study:

- Reinforcement learning with uncertainty (example), Error Driven Learning
- MAP Inference, Estimating a joint probability distribution
- Probabilistic Refinement
- Bayesian networks and probabilistic graphical models
- Spatial Descriptive Statistics
- Polya Process
- Forward-backward algorithm and Viterbi algorithm

DELIVERABLES

Deliverables will be a cumulative development of a series of probabilistic methods (models) collected into a custom Python library to ultimately break down high-dimensional data into a shorter list of calculated indeterminate features.

1. Starting ~week 2 and due by the end of week 6, the first deliverable includes an outline of the library, including standard formats of data and Input/Output that can operate across probabilistic methods, allowing models to be chained together.
2. Week 8: Completion of Bayesian models' implementation
3. Week 10: Markov model implementation
4. Week 12: Implementation of function hooks allowing end-user development of unique monte-carlo simulations
5. Week 15: Final report of product library, stretch goal is to include distribution visualizations and model chain graph UI

Supervision by the advising faculty member will include at minimum weekly discussion to identify progress made each week and potential barriers to continuation, allowing for the schedule to be adapted to actual efforts required for the defined scope of each topic.

EVALUATION

Evaluation will be managed primarily through review of the deliverables and how the understanding of concepts changed in completing the deliverables.

A secondary evaluation method of applying learned concepts to real and synthetic datasets is also accepted, especially when such experimentation occurs through usage of the library constructed in the deliverables.