Comparison of Bayesian classifiers under different data

Zijun Sun, Ruixin Li
Master's Degree in Artificial Intelligence, University of Bologna
{ zijun.sun, ruixin.li }@studio.unibo.it

January 5, 2024

Abstract

This mini-project applies Bayesian classifiers and advanced statistical methods to analyze student stress factors. It preprocesses the Kaggle 'Student Stress Factors' dataset and utilizes the MNIST dataset, employing four Naive Bayes classifiers with a peak accuracy of 91.73% using Gaussian Naive Bayes. A Bayesian Network is constructed for deeper insight, Additionally, a Bayesian Network is constructed using pgmpy. The project concludes with HMC MCMC logistic regression, achieving 89.54% accuracy. This approach demonstrates the efficacy of combining Bayesian methods and statistical analysis in predictive modeling.

Introduction

Domain

In this project, we leverage advanced statistical and machine learning techniques to model student stress factors, addressing the increasing need to understand and predict mental health outcomes in educational environments. Utilizing the comprehensive 'Student Stress Factors' dataset from Kaggle, which encompasses 20 features across five key categories (Psychological, Physiological, Social, Environmental, and Academic), our goal is to analyze the multifaceted influences on student stress.

Recognizing the complexity of mental health in educational settings, our methodology incorporates a variety of Bayesian classifiers(Vikramkumar, B, and Trilochan 2014) (Multinomial, Gaussian, Complement, and Bernoulli Naive Bayes) to effectively handle the diverse data attributes. Moreover, the implementation of HMC MCMC(201 2011) logistic regression through TensorFlow Probability allows for a more detailed analysis, particularly in assessing the uncertainty of model parameters.

Aim

This project delves into modeling student stress factors using the Kaggle 'Student Stress Factors' dataset with several key goals:
1. Data Preprocessing and Analysis: This involves normalizing the data and computing indices for key stress factor categories to highlight their correlation with student stress levels.
2. Model Implementation and Experimentation: The project applies various Naive Bayes classifiers (Multinomial, Gaussian, Complement, Bernoulli) to predict stress levels, assessing their effectiveness in different scenarios.
3. Bayesian Network Construction: A Bayesian Network(Heckerman 2022) is built and visualized using pgmpy. This aspect focuses on exploring network structures and conducting probabilistic inferences to understand the intricate relationships between different factors.
4. Quantification of Uncertainty with TensorFlow Probability: The project utilizes HMC MCMC methods to quantify the uncertainty in model parameters, employing a multinomial logistic regression model for stress level prediction.

Method

The study utilized a range of machine learning techniques to analyze the "Student Stress Factors" dataset from Kaggle. Data was preprocessed and normalized, with indices for stress factors calculated based on stress level correlations. Bayesian Networks were implemented using pgmpy, optimized with HillClimbSearch and BicScore, and managed through blacklists. The study also compared four Naive Bayes classifiers (Multinomial, Gaussian, Complement, Bernoulli) with Bayesian Networks. TensorFlow Probability was used with Hamiltonian Monte Carlo MCMC methods to assess parameter uncertainty in a multinomial logistic regression model. Performance was evaluated using accuracy, precision, recall, F1 score, and confusion matrices. The final phase involved interpreting the results, offering insights into the effectiveness of different classifiers and methodologies in understanding student stress factors. This approach emphasized the potential of machine learning in psychological and social data analysis.

Results

In the analysis of the student stress dataset, the Multinomial and Gaussian Naive Bayes classifiers emerged as top performers, demonstrating strong metrics across the board. In contrast, the Bayesian Network, while showing promise in Accuracy and Precision, needs further development in Recall to be a viable alternative. Notably, the Bernoulli and Complement Naive Bayes classifiers were less effective for this particular dataset.

Regarding the MNIST dataset(Deng 2012), the Bernoulli Naive Bayes classifier stood out, offering a balanced performance, likely due to the dataset's discrete, binary nature. This feature also explains the strong performance of the Multinomial Naive Bayes in this context.

Model

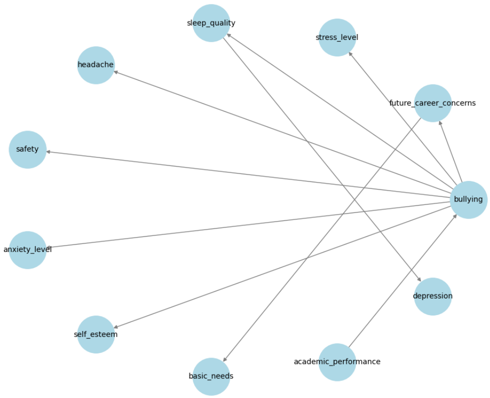

Figure 1: Bayesian network.

- Nodes (Random Variables): Each node in our model represents a variable, like stress-level, anxiety level, self-esteem, etc. These variables are related to psychological or behavioral aspects and are ordinals.
- Conditional Distributions (CPTs): Our model uses Conditional Probability Tables (CPTs) to show how likely each state of a variable is, given the states of its parent variables. For example, a CPT might show how likely different levels of stress level are, depending on anxiety level and self-esteem.
- Model Building Procedure:
  Data Splitting: The dataset was divided into training and testing sets to evaluate the model's performance on unseen data.
  Blacklisting Edges: We decided not to include certain relationships (edges) in Our model based on Our understanding or assumptions about the data.
  Structure Learning: We used Hill Climb Search with the Bayesian Information Criterion to find the best structure for Our network. This method iteratively tries different structures and picks the best.
  Parameter Learning: After setting up Our network's structure, We used the BayesianEstimator to figure out the probabilities for the CPTs.
  Inference: Finally, We used VariableElimination to predict or infer the probabilities of certain variables based on others.

Analysis

Experimental setup

We conducted experiments using diverse Naive Bayes classifiers, a Bayesian Network, and HMC MCMC methods within TensorFlow Probability, aiming to assess how each model performs on datasets related to student stress factors. The evaluation plan involves comparing expected outcomes based on dataset characteristics and classifier properties, analyzing key performance metrics like accuracy, precision, recall, and F1 score, and conducting inter-model comparisons. Additionally, the fit of data distribution to Gaussian models is assessed using MSE approximation, providing insights into the suitability of different classifiers for the data. This comprehensive approach allows for a deep analysis of each model's strengths and limitations in the context of your specific datasets.

Results

The normalization process significantly enhanced model performance, especially for Gaussian classifiers, as anticipated given their preference for normally distributed data. However, a notable surprise was the exceptional performance of the Bernoulli Naive Bayes on the MNIST dataset, which can be attributed to its binary features being well-aligned with the binary nature of the Bernoulli model. Additionally, the Bayesian Networks and HMC MCMC methods demonstrated high effectiveness in these Bayesian applications, despite their complexity and computational intensity. This suggests their suitability for specific dataset types and problem scenarios where advanced probabilistic modeling is beneficial.

Conclusion

In this project, I learned about the versatility and limitations of Bayesian classifiers and advanced statistical methods. The application to both the 'Student Stress Factors' and MNIST datasets was enlightening, demonstrating the methods' adaptability but also revealing their sensitivity to data types. While Gaussian Naive Bayes showed high accuracy, the complexity of Bayesian Networks and HMC MCMC logistic regression highlighted challenges in computational demands and high-dimensional data handling. This experience provided valuable insights into the balance between sophisticated methodological approaches and their practical constraints.

Links to external resources

`https://github.com/DS-KB-lab/Bayesian-classification.git`

References

[201 2011] 2011. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

[Deng 2012] Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* 29(6):141–142.

[Heckerman 2022] Heckerman, D. 2022. A tutorial on learning with bayesian networks.

[Vikramkumar, B, and Trilochan 2014] Vikramkumar; B, V.; and Trilochan. 2014. Bayes and naive bayes classifier.