Conditional Mutual Information Estimation and Feature Importance Ranking

Liu Chang

June 2, 2024

1 Introduction

The goal of this project is to implement estimators of conditional mutual information (CMI) and use them to rank features based on their importance.

2 Methodology

2.1 Difference Based Methods

2.1.1 True Conditional Mutual Information

For the true CMI, we use analytical methods based on the properties of Gaussian distributions and their covariance matrices.

I generate three Gaussian variables X, Y, and Z with a known covariance matrix.

For Gaussian variables, mutual information can be calculated using their covariance matrices. The formula for mutual information $I(X;Y)$ between two Gaussian variables is:

$$I(X;Y) = \frac{1}{2} \log \left(\frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{XY}|} \right)$$

where Σ_X and Σ_Y are the covariance matrices of X and Y, respectively, and Σ_{XY} is the joint covariance matrix of X and Y.

Conditional mutual information $I(X;Y|Z)$ is calculated using the chain rule for mutual information:

$$I(X;Y|Z) = I(X,Z;Y) - I(X;Y)$$

2.1.2 Estimated Conditional Mutual Information

For the estimated CMI, we use a neural estimator based on the Donsker-Varadhan (DV) representation of KL divergence.

The DV representation of KL divergence is used to estimate mutual information. A neural network model is constructed to estimate the mutual information. The model is trained using the Adam optimizer, and early stopping is employed to prevent overfitting.

2.1.3 Estimating Mutual Information

The mutual information values $I(X, Z; Y)$ and $I(X; Y)$ are estimated using the neural network model. The estimated conditional mutual information $I(X; Y|Z)$ is derived from these values using the chain rule.

2.1.4 Result Known CMI Formula

- **True Values:**
  - True $I(X, Z; Y) = 0.009156364979020115$
  - True $I(X; Y) = 0.009156364979020115$
  - True $I(X; Y|Z) = 0.0$
- **Estimated Values:**
  - Estimated $I(X, Z; Y) =$ -2.0422613620758057
  - Estimated $I(X; Y) =$ -0.3410928249359131
  - Estimated $I(X; Y|Z) =$ -1.7011685371398926

2.1.5 Simulated Data

The top 10 important features for the simulated data include features 19, 18, 1, 2, 3, 4, 5, 6, 7, and 8. The intersection size of 8 indicates a high level of agreement between the simulated and estimated important features. The absence of inversions suggests that the ranking of these features is consistent. This consistency indicates that the simulated dataset is well-behaved and the CMI method is effectively capturing the important features.

2.1.6 Real Data

The top 10 important features for the real data are features 29, 15, 3, 4, 5, 6, 7, 8, 9, and 10. The intersection size of 7 shows a significant overlap, although slightly less than the simulated data. The absence of inversions indicates a consistent ranking of features, suggesting that the real dataset also provides reliable feature importance estimates. However, the slightly lower intersection size compared to the simulated data might indicate more complexity or noise in the real dataset.

2.2 Comparison of Simulated and Real Data

Both datasets show a substantial overlap in important features, with intersection sizes of 8 and 7 for simulated and real data, respectively. The lack of inversions in both datasets indicates a stable ranking of features. The higher intersection size in the simulated data suggests that the simulated dataset is more straightforward or less noisy compared to the real dataset. This could be due to controlled conditions in the simulation, whereas real-world data often contain more variability and noise.

2.3 Generative and Divergence Based Method

2.3.1 Data Preparation

Two datasets are considered: simulated data and real data. Features and target variables are identified in each dataset.

2.3.2 CMI Calculation

The `mutual_info_classif` function from `sklearn.feature_selection` is used to calculate mutual information. CMI is calculated by conditioning the mutual information on another variable or set of variables.

2.3.3 K-Nearest Neighbors (KNN) Method

KNN is used to estimate the probability densities needed for mutual information calculations. For each data point, the k nearest neighbors are found and used to estimate local density, which helps in calculating CMI.

3 Results

3.1 True CMI vs Estimated CMI

The true CMI line is flat at 0, while the estimated CMI line shows significant fluctuations.

3.2 Feature Importance

3.2.1 Simulated Data

The top 10 important features for the simulated data include features 19, 18, 1, 2, 3, 4, 5, 6, 7, and 8. The intersection size of 8 indicates a high level of agreement between the simulated and estimated important features. The absence of inversions suggests that the ranking of these features is consistent. This consistency indicates that the simulated dataset is well-behaved and the CMI method is effectively capturing the important features.

3.2.2 Real Data

The top 10 important features for the real data are features 29, 15, 3, 4, 5, 6, 7, 8, 9, and 10. The intersection size of 7 shows a significant overlap, although slightly less than the simulated data. The absence of inversions indicates a consistent ranking of features, suggesting that the real dataset also provides reliable feature importance estimates. However, the slightly lower intersection size compared to the simulated data might indicate more complexity or noise in the real dataset.

3.2.3 Comparison of Simulated and Real Data

Both datasets show a substantial overlap in important features, with intersection sizes of 8 and 7 for simulated and real data, respectively. The lack of inversions in both datasets indicates a stable ranking of features. The higher intersection size in the simulated data suggests that the simulated dataset is more straightforward or less noisy compared to the real dataset. This could be due to controlled conditions in the simulation, whereas real-world data often contain more variability and noise.

4 Problems and Possible Reasons

- The true CMI being constant at 0 is unusual and suggests a problem in the definition or calculation of true CMI. The large fluctuations in estimated CMI suggest potential issues with the estimation method.
- The quality of data might be insufficient, leading to unstable CMI estimates.
- Incorrect assumptions in the model or method used for estimating CMI could lead to incorrect results.
- High noise levels or variability in the data can impact the accuracy of CMI estimates.
- The methods or libraries used for calculating CMI might have limitations or specific requirements not being met in the current implementation.
- The neural estimator produced negative values for mutual information, indicating issues with the model's stability and training. The discrepancy between true and estimated values suggests a need for further refinement of the DV loss function and hyperparameter tuning.

5 Conclusion

The project successfully implemented estimators for conditional mutual information and ranked features based on their importance. However, further re-

finements are necessary to improve the accuracy and stability of the mutual information estimates.