|  |
| --- |
| **Explainable AI** |

**Abhishek Bhave**

**UBitName: abhave PersonNumber: 50289049**

Department of Computer Science

University of Buffalo

Buffalo, NY 14214

[*abhave@buffalo.edu*](mailto:hippo@cs.cranberry.lemon.edu)

**Abstract**

This project is to develop a machine learning system that learns explainable features for a task domain while it is learning to answer a variety of queries in that domain. It involves combining deep learning and probabilistic graphical models.

We are given M samples of an input, e.g., a photograph of a person, scanned images of a handwritten word written by a known writer. Our goal is to work with three different sets of features: human determined features, deep learning features, explainable deep learning features.

**1.1 Introduction**

Neural networks are increasingly being used to build programs that can predict and classify in a myriad of different settings. However, we don’t really know how they work. Even if we tune the hyperparameters, we get highly accurate and efficient model with very little loss, we still cannot explain why the model predicted a certain value, or why the model is giving out these sets of results. For this purpose, we use the concept of explainable AI, which is basically explaining the result along with the predictions.

Explainable AI (XAI), Interpretable AI, or Transparent AI refer to techniques in artificial intelligence (AI) which can be trusted and easily understood by humans. It contrasts with the concept of the "black box" in machine learning where even their designers cannot explain why the AI arrived at a specific decision. XAI can be used to implement a social right to explanation. Some claim that transparency rarely comes for free and that there are often tradeoffs between how "smart" an AI is and how transparent it is; these tradeoffs are expected to grow larger as AI systems increase in internal complexity.

Why the need for Explainable Models?

Neural Networks are not infallible.

Besides the problems of overfitting and underfitting that we’ve developed many tools (like Dropout or increasing the data size) to counteract, neural networks operate in an opaque way.

We don’t really know why they make the choices they do. As models become more complex, the task of producing an interpretable version of the model becomes more difficult.

In decision trees or neural networks, the complexity of the method itself removes the ability to carry through the original factors that drive the algorithm’s predicted outcome. Of course, regression models provide an explanation but as previously stated, they aren’t machine learning and suffer from lower precision relative to their machine learning counterparts.

So, Explainable AI (XAI) is basically any machine learning technology that can accurately explain a prediction at the individual level.

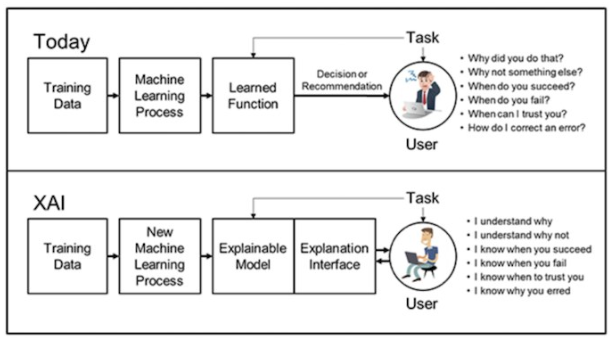


Figure 1 Need of Explainable AI (XAI)

1. Human determined features

We assume that we have human-described variables (and the values that they can take) for an input image.

2. Features learnt using deep learning

Here the raw input (scanned images) are processed by a network to learn a representation, e.g., by means of a several convolutional network and pooling layers. The training could be performed by either by unsupervised learning (e.g., an autoencoder) or by supervised learning (e.g., learning whether a pair of samples is by the same writer).

3. Explainable features learnt using deep learning

These are the representation learnt using deep learning is as similar as possible to the human explainable features described in the first part. Ideally the representation learnt by the deep learner is the same as the human explainable features. This can be learnt using supervised learning where the desired outputs are the explainable features.

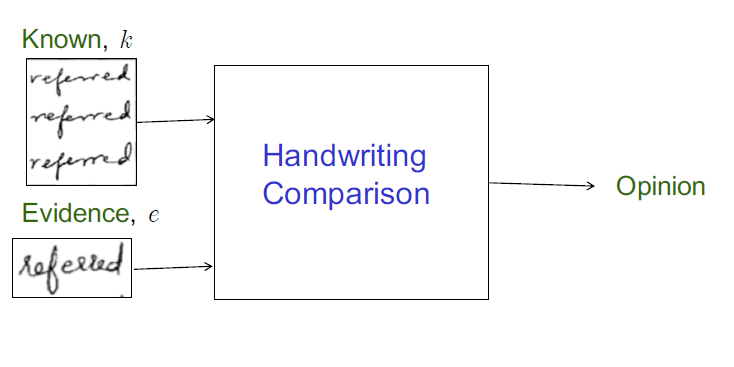


Figure 2 Handwriting Comparison Task

**1.2 Dataset**

In this project we are going to be working on the AND dataset and we are going to create the Bayesian model on the AND dataset.

Data containing handwritten 'AND' images. The details of the AND dataset is as follows:

• Every writer wrote three different pages namely a,b,c.

• Number of writers: 1552

• Number of images: 15260

• Filename Format: XXXXY\_numZ.png [where XXXX is writer number, Y is page number, Z is occurrence of 'AND' in Y]



Figure 3 Various images depicting different structures of AND

For every image of AND in our dataset, we have 15 features corresponding to it, and for every feature we have a feature class associated to it. The 15 features are defined as follows:

Pen Pressure – Pressure applied when writing AND letter.

Letter Spacing – Letter spacing is the distance or the spacing between two letters in the AND word

Size – Size is dependent on dimension and letter spacing.

Dimension – We can classify dimension into high-dimension when the height of the characters are bigger than the width and low-dimension, otherwise low or medium.

Is LowerCase – If sample image is in lowercase or not.

Is Continuous – Is Continuous is dependent on IsLowerCase

Slantness – Can have the classes slight right, very right,left and no slant.

Tilt – Tilt feature is dependent on Slantness.

Entry stroke "A" – Entry stroke of "A" can be nostroke or downstroke.

Staff of "A" – Can have classes like retraced, loopy, teneted or no staff.

Formation "N" – Generally depends on word formation which is depenedent on constancy and size.

Staff of "D" - Staff of "D" is dependent on Staff of "A" and can have retraced, loopy and no staff classes.

Exit stroke "D" – Exit stroke "D" is dependent on Entry stroke "A".

Word formation -Word formation is whether the word has been formed correctly or not and is dependent on constancy .

Constancy – Constancy is dependent on Size which is dependent on dimension and letter spacing.

**2 Task 1 Data Annotation**

**3 Task 2 Bayesian inference**

First, we use thresholding to determine if two variables are independent or not. In this process we eliminate a few conditional probability distributions.

Next, we create Bayesian models using the pgmpy library. These are Directed Acyclic Graphs (DAG) with a directed link between two correlated variables.

We first add the edges to the Bayesian model where each edge is a link between a parent node and a child node. After adding the edges to the model, I have added the corresponding marginal and conditional probabilities in the TabularCPD table.

For all the nodes which do not have any dependency i.e. do not have a parent we add their data from their marginal distribution.

After adding all the variables in the TabularCPD, I have added all the variables to the model. Further I have calculated inference and then conducted sampling using forward\_sample. The number of samples for all the models is kept at 50000

Lastly, I have calculated and printed the K2 score for each model.

Some of the models are as follows:

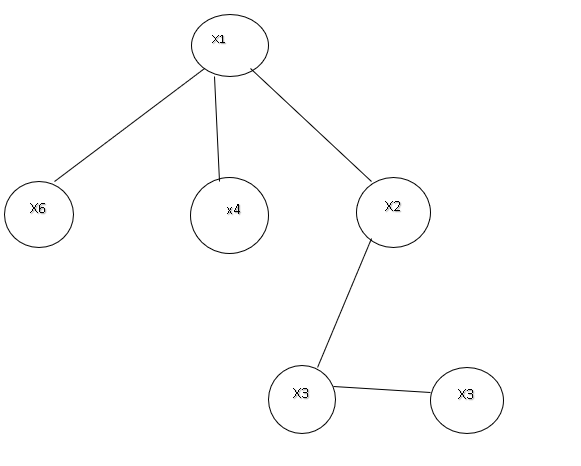


Figure 5

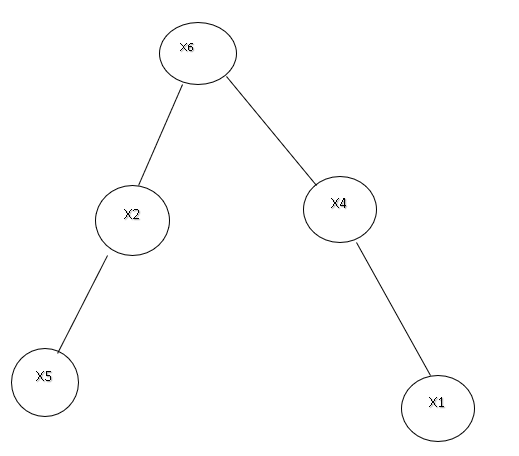


Figure 6

The best model is the model with the best K2 score. The best model among all of my models is:

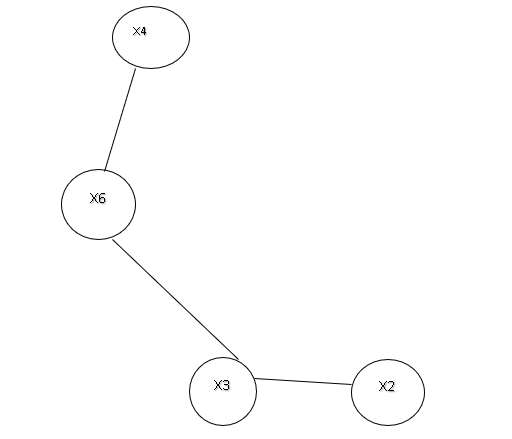


Figure 7 Best Model

In the next step of this task I have calculated the high and low probability of th. Below are the results:

The high probability of th is: 0.049

The low probability of th is: 0.0001

**4 Task 3 Markov Network Construction and Inference**

I have converted my best Bayesian network into a Markov network using moralization. I have calculated Bayesian network inference and the Markov network inference in terms of computation time.

Following are the results from the comparison between Bayesian Model and Markov Model:

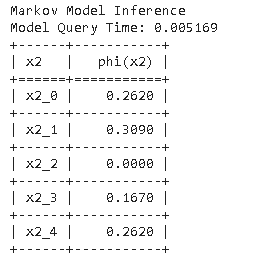
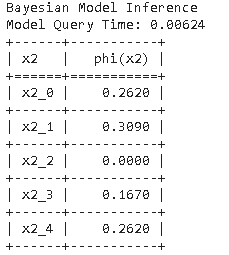


Figure 8 Bayesian Model Inference Figure 9 Markov Model Inference

**5 Task 4 AND Dataset**

In this task I have worked with the AND dataset. Firstly, I extracted the 9 features that I wanted i.e. x1 to x9.

Now to look for the best model among this data I have applied HillClimbSearch which returns the best model.

I have then calculated K2 score of that model which is displayed as below:

[('f3', 'f8'), ('f3', 'f9'), ('f3', 'f4'), ('f5', 'f9'), ('f5', 'f3'), ('f9', 'f1'), ('f9', 'f2'), ('f9', 'f4'), ('f9', 'f6'), ('f9', 'f7'), ('f9', 'f8')]

Model K2 Score: -9462.70489237

**References**

[1]-https://medium.freecodecamp.org/an-introduction-to-explainable-ai-and-why-we-need-it-a326417dd000

[2]- https://simmachines.com/explainable-ai/

[3]- https://en.wikipedia.org/wiki/Explainable\_artificial\_intelligence

[4]-

[5]-