|  |
| --- |
| **CSE 674 Project 1: Determining Probabilities of**  **Handwriting Formations using PGMs** |

**Neeraj Ajit Abhyankar**

Department of Computer Science and Engineering

State University of New York

University at Buffalo

*Email – nabhyank@buffalo.edu*

**Abstract**

Probabilistic Graphical Models can be used to determine probabilities of observations which are described using several variables. In this project we will work with two datasets of handwriting patterns viz. for ‘th’ and for ‘and’ since they are commonly occurring. These datasets are based on observations using several variables with minute details of the relationship between both the individual letters and combined. Also this helps in devising methods to identify whether the given handwriting patter is rare or common, further leading to handwriting recognition and analysis with respect to certain individual. In this Project we perform four tasks; in the first we evaluate the pairwise correlation and independencies for multi-categorical data by calculating the cross-entropy, then in second we construct a few Bayesian networks (PGM’s) with maximum likelihood edges using independencies from the first task and compare them, in the third task we convert the Bayesian networks to Markov networks and compare the inferences, in the fourth task we use the ‘and’ dataset to predict the best model and compare it with other models constructed. Thus allowing us a further in-depth analysis into the handwriting domain and data generation w.r.t. CPD (Conditional Probability Distribution) and MPD (Marginal Probability Distribution) values.

1. **Feature Definitions and Details :**

To categorize any data, we need features and details of the data. Similarly, handwritten data too has certain features, the extraction, characterization of these features with the structural details will be our core dataset for analysis and evaluation on the assigned tasks. These features and details are obtained by the document examiners and their assessment of different documents and the difference or similarities found in the data found.

Below is such a characterization of the structure of dataset ‘th’ as given by document examiners as shown in Table (a).

The below table describes the structural details involved in the characterization of a handwritten data. The table can be read as matrix where umber of rows and columns are related with respect to each other at each block in the table. Say we need to find out what is the shape of the arch when the height relationship between ‘t’ to ‘h’ is ‘t’ is taller than ‘h’. We search the table for ‘x1’ and ‘x3’ in the first row and find the respective features, now to find the details we can browse over the columns and find that when ‘x1’ has ‘t’ is taller than ‘h’ ‘x3’ has ‘no set pattern (c)’. So we come to know that there is no set pattern which is observed when the given condition occurs. Similarly, we can use the follow g table as an example to find the structural details of the given dataset which will be further used for the analysis of tasks to be performed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x1  (Height Relationship of t to h) | x2  (Shape of Loop of h) | x3  (Shape of Arch of h) | x4  (Height of Cross on t staff ) | x5  (Baseline of h) | x6  (Shape of t) |
| t shorter than h  (a) | retraced (a) | rounded arch (a) | upper half  of staff (a) | slanting  upward (a) | tented (a) |
| t even with h (b) | curved right side  and straight left  side (b) | pointed (b) lower | lower half  of staff (b) | slanting  downward  (b) | single stroke (b) |
| t taller than h (c) | curved left side  and straight  right side (c) | no set pattern  (c) | above staff  (c) | baseline  even (c) | looped (c) |
| no set pattern  (d) | both sides  curved (d) |  | no fixed  pattern (d) | no set pat-  tern (d) | closed (d) |
|  | no fixed pattern  (e) |  |  |  | mixture of  shapes (e) |

**Table (a) : Characteristics of ‘th’ as provided by document examiners.**

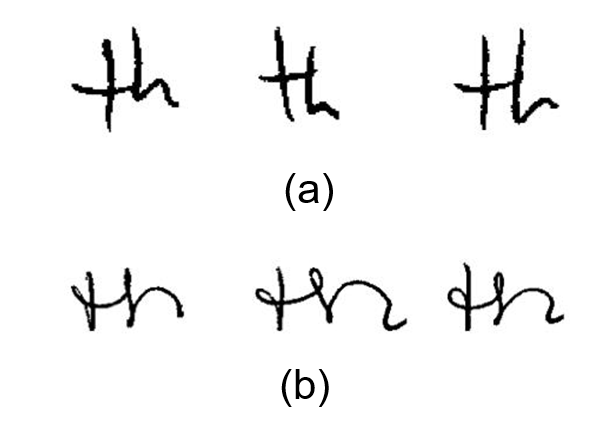
**[In this characterization here are six random variables x1-x6 taking on the values a; b; c; d;e as indicated in parentheses.]**

Let’s look at an example of data :

Below are two samples of images encoded such that their probabilities can be determined by constructing graphical model to determine whether a particular sample was written by a certain individual whose handwriting characteristics are known.

In the below images, the characteristics of

1. individual 1 encoded are as x1 = b; x2 = a; x3 = a; x4 = d; x5 = a; x6 = b
2. individual 2 as x1 = b; x2 = b; x3 = a; x4 = b; x5 = a; x6 = c.



1. **Importance of Distribution Values and Requirement :**

To understand the working of the project we need to understand why it is necessary to have certain values of distributions.

Marginal Probability Distribution (MPD) gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. In our case we need these values to calculate the independencies between two nodes and also when we have independent nodes in the models we will construct.

When we look at the below Table (b) we find all the marginal probability distribution values from x1-x6 and can compare it with Table (a) for similarity in the structural features and characteristics.

These distributions will be considered for the independent nodes in the models we will construct as Bayesian Networks (PGM’s). We have all the MPD’s given in ‘Table2.csv’ file for our use which looks exactly like the below Table (b).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Values | x1 (Relative height of t to h) | x2 (Shape of h loop) | x3 (shape of h arch) | x4 (height of cross of t) | x5 (Baseline of h) | x6 (Shape of t) |
| a | 78%(156) | 27:5%(55) | 18%(36) | 71:5%(143) | 37:5%(75) | 1:5%(3) |
| b | 1.5%(3) | 32%(64) | 66%(132) | 10.5%(21) | 11%(22) | 32%(64) |
| c | 5.5%(11) | 2.5% (5) | 16%(32) | 1%(2) | 10.5%(21) | 14%(28) |
| d | 15%(30) | 17%(34) |  | 17%(34) | 41%(82) | 31.5%(63) |
| e |  |  | 21%(42) |  |  | 21%(42) |

**Table (b) : Marginal distributions of the six features of ‘th’.**

**These Marginal probabilities (verbose) are Based on samples from 200 individuals. From Muehlberger, et. al., "A Statistical Examination of Selected Handwriting Characteristics,", Journal of Forensic Sciences (1977), 205-211.**

Conditional Probability Distribution (CPD) of two jointly distributed random variables X and Y, can be given as; the conditional probability distribution of Y given X is the probability distribution of Y when X is known to be a particular value. In our case, we need these values to calculate the dependencies of two nodes and also when we have dependent nodes in the models we will construct.

We have all these CPD’s in the ‘.csv’ files provided to us. Each ‘.csv’ file provides the CPD’s of x1-x6 in ‘Table3.csv’ to ‘Table8.csv’ for six variables. For details consider the below ‘Table3.csv’ file.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x1 (Parent node) | x01 (t shorter than h) | x11 (t even with h) | x21 (t taller than h) | x31 (No set pattern) |
| Total MPD | 78%(156) | 1.5%(3) | 5.5%(11) | 15%(30) |
| x02 (retraced staff) | 23.1%(36) | 66.6%(2) | 45.5%(5) | 40%(12) |
| x12 (curved right) | 36.5%(57) | 0%(0) | 9.1%(1) | 20%(6) |
| x22 (curved left) | 2.6%(4) | 0%(0) | 0%(0) | 3.3%(1) |
| x32 (both curved) | 17.3%(27) | 0%(0) | 18.2%(2) | 16.7%(5) |
| x42 (no pattern) | 20.5%(32) | 33.3%(1) | 27.3%(3) | 20%(6) |
| x04 (upper staff) | 73.7%(115) | 100%(3) | 72.7%(8) | 56.7%(17) |
| x14 (lower staff) | 7.7%(12) | 0%(0) | 27.3%(3) | 20%(6) |
| x24 (above staff) | 1.3%(2) | 0%(0) | 0%(0) | 0%(0) |
| x34 (no pattern) | 17.3%(27) | 0%(0) | 0%(0) | 23.3%(7) |
| x06 (tented t) | 1.9%(3) | 0%(0) | 0%(0) | 0%(0) |
| x16 (single stroke t) | 28.2%(44) | 66.6%(2) | 54.5%(6) | 40%(12) |
| x26 (looped t) | 12.8%(20) | 33.3%(1) | 9.1%(1) | 20%(6) |
| x36 (closed t) | 35.2%(55) | 0%(0) | 18.2%(2) | 20%(6) |
| x46 (mixed shapes) | 21.8%(34) | 0%(0) | 18.2%(2) | 20%(6) |

**CPD’s for x1 as Parent Node in ‘Table3.csv’**

Consider the above table, in this we consider the conditional probability distributions of node ‘x1’ as Parent Node; the table shows all the values in percentage of probabilities of occurrences and the number of occurrences in brackets for all respective nodes as children.

Now in our Bayesian Model if we have node ‘x1’ as parent at someplace with nodes ‘x2’, ‘x4’, ‘x6’ as children then we will look the corresponding values from this table to find the dependencies and other required parameters for tasks to be performed. Similarly, we have all the CPD’s for given six variables from ‘x1’ - ‘x6’ as parents with respective children and their distributions in the provided ‘.csv’ files.

Say we need to find what are the CPD’s when ‘x1’ is the parent and ‘x4’ is the child so we look at the row with ‘x04 (upper staff)’ to row ‘x34 (no pattern)’ and its corresponding ‘x01 (t shorter than h)’ to ‘x31 (no set pattern)’ columns. These will give us the values of CPD’s for the required parent and child combination for computation of parameter in the tasks.

1. **Probabilistic Graphical Models :**

From known standard definition a graphical model or probabilistic graphical model (PGM) or structured probabilistic model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables.

In our case the PGM will be a Bayesian Network including multiple nodes with parent child relationships. We will construct a few Bayesian Models based on acquired relationships from the independencies and by putting a likelihood threshold for efficient construction of the models.

The main aim of Probabilistic Graphical Models is to provide an intuitive understanding of joint probability among random variables.

1. **Descriptions of Given Tasks :**

**4.1 Task-1 : To Evaluate pairwise correlations and independences in the given data**The basic data pre-processing is performed in Task-1 where we extract the given MPD’s and CPD’s from the ‘.csv’ files. In this task we have used a recursive loop to extract all the data from all given ‘.csv’ files in a list of lists from which we can extract and slice all the required data into parts as per requirement. We have divide the list of lists into seven parts, where the MPD’s are stored separately and the CPD’s for six variables is stored in six different parts. This allows us to easily manipulate the data into locations when needed for further analysis and execution of other tasks.  
  
Since we have multi-categorical variables, we can measure the correlation by calculating the cross-entropy for all variable dependencies using ‘.csv’ files as :

After we calculate the values using above equation we get all the correlations and independences where the values vary from 0.09 to 0.21.  
  
 **4.2 Task-2 : Construct a Bayesian network with the fewest number of edges that maximizes the likelihood**  
  
The PGM’s we will be implementing in our project are Bayesian Networks. To construct the Bayesian Networks, we need nodes and edges of maximum likelihood so as to build models/graphs for representation and analysis of the same. From Task-1 we have the independences of the variables with each other, also the MPD’s and CPD’s. Now we will threshold the obtained values in Task-1 to 0.14 ≥ ; the values which are greater or equal to the threshold are the values which are independent and below are dependent values. Only the values which are independent after threshold will be chosen for further analysis. After choosing the appropriate values we will build approximately six model using the acquired nodes. These will be the fewest edges with maximum likelihood, which will further be used for the construction of Directed Acyclic Graphs (DAG’s). The DAG’s will be our models for sampling data from the given CPD’s and MPD’s to acquire the minimum scoring of likelihood for best model.   
  
To construct the DAG’s we start by considering a parent node and use assigned children from data provided to us. Then we construct a Network of these nodes and join the edges for each parent node to its children; we sample out using ancestral sampling method the parent first and then its children to gather the data for given sample. We need all six nodes to sample the data and around 1,000 samples (in our case 10,000). The parameter for measuring the best and the worst model is the K2 scores; where the model with Maximum K2 score is the Best Model with Maximum likelihood of occurrence with a certain probability, similarly the Worst Model has Minimum K2 score and Minimum likelihood of occurrence with a certain probability which can be calculated based on Maximum Likelihood for both.   
**4.3 Task-4 : Use the ‘and’ image dataset to construct a Bayesian network and evaluate the goodness score of several Bayesian Networks**   
  
\*\*Note : We will be performing Task-4 before Task-3 because Bayesian to Markov conversion will be performed only after analysis of both the ‘th’ and ‘and’ datasets.\*\*  
  
For this Task we are given the dataset of handwriting ‘and’ of nine features for each image. We have to construct several Bayesian networks and evaluate the performance of them all using the K2 scoring method. In this Task we will use the Hill Climb Search Algorithm, which belongs to statistical local search algorithms. Hill Climb Search is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by making an incremental change to the solution. By using this algorithm, we will find the best possible model for the ‘and’ dataset. After we obtain the Best Model we will calculate the K2 score of this model and then design several other models for comparison. Also if we want we can estimate the CPD’s of the edges for additional inferences and calculations.  
  
**4.4 Task-3 : Convert the Best Model Bayesian Network into Markov Network using moralization. Also Compare the inferences of Bayesian and Markov Networks:**   
  
For Markov Models using Probability theory it is assumed that future states depend only on the current state, not on the events that occurred before it. Markov Models are commonly used in the fields of [predictive modelling](https://en.wikipedia.org/wiki/Predictive_modelling) and [probabilistic forecasting](https://en.wikipedia.org/wiki/Probabilistic_forecasting), it is desirable for a given model to exhibit the Markov property.  
  
We will now perform and analyze Task-3 by converting the given Bayesian Networks to Markov Networks. For this Task we will use the pgmpy library function ‘to\_markov’. Then we can also find the probability of maximum likelihood for this model.  
Also we can use the time stamps to find out which of the models have performed well i.e. Bayesian Model or Markov Model.

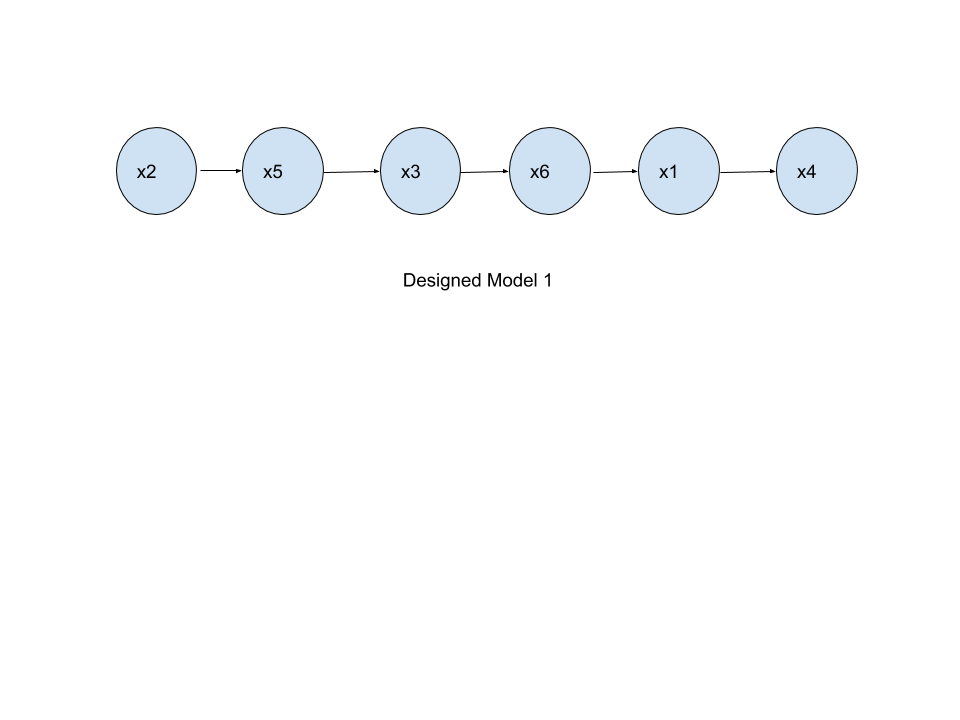
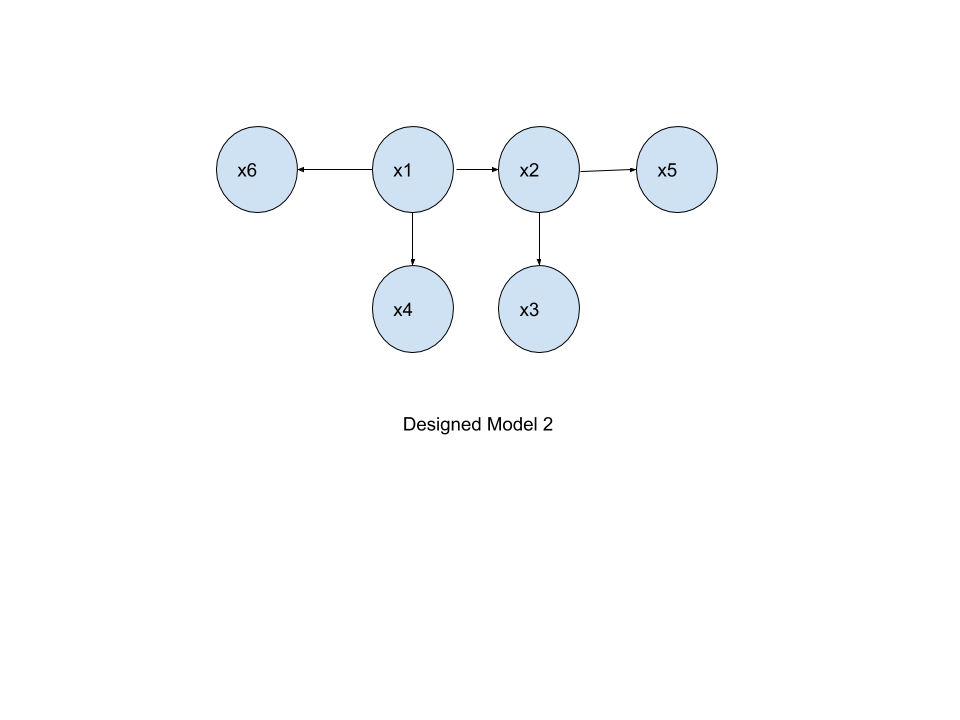
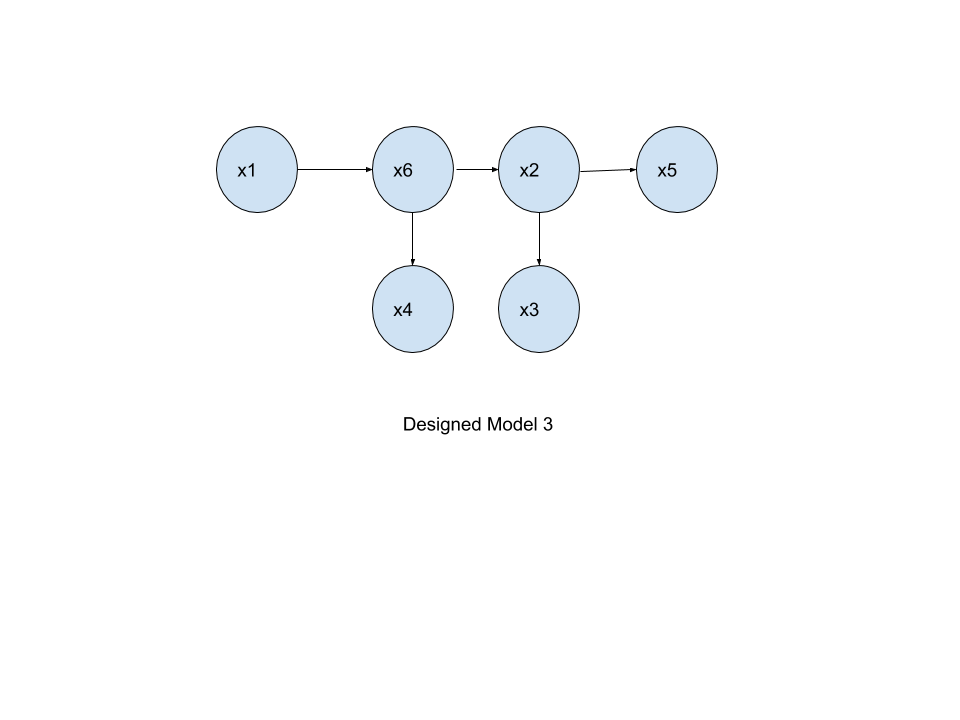
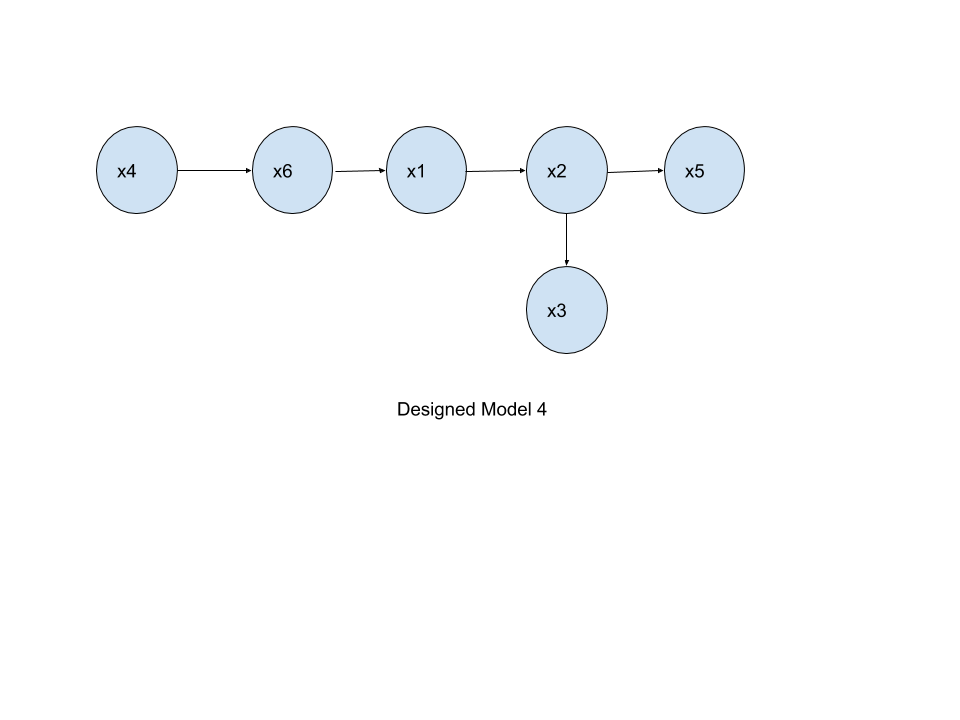
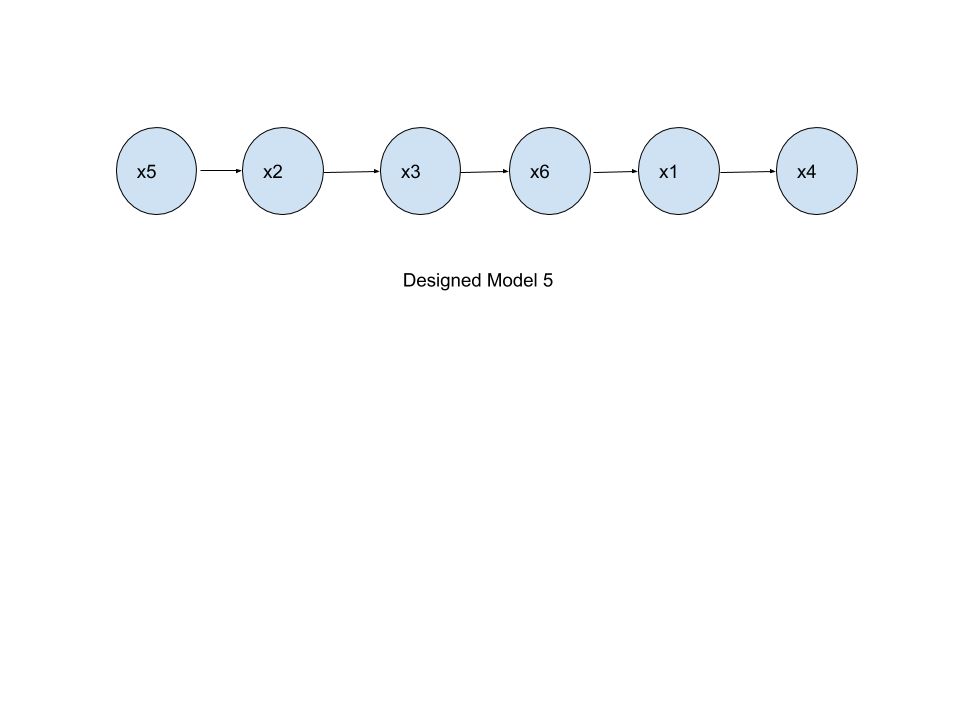
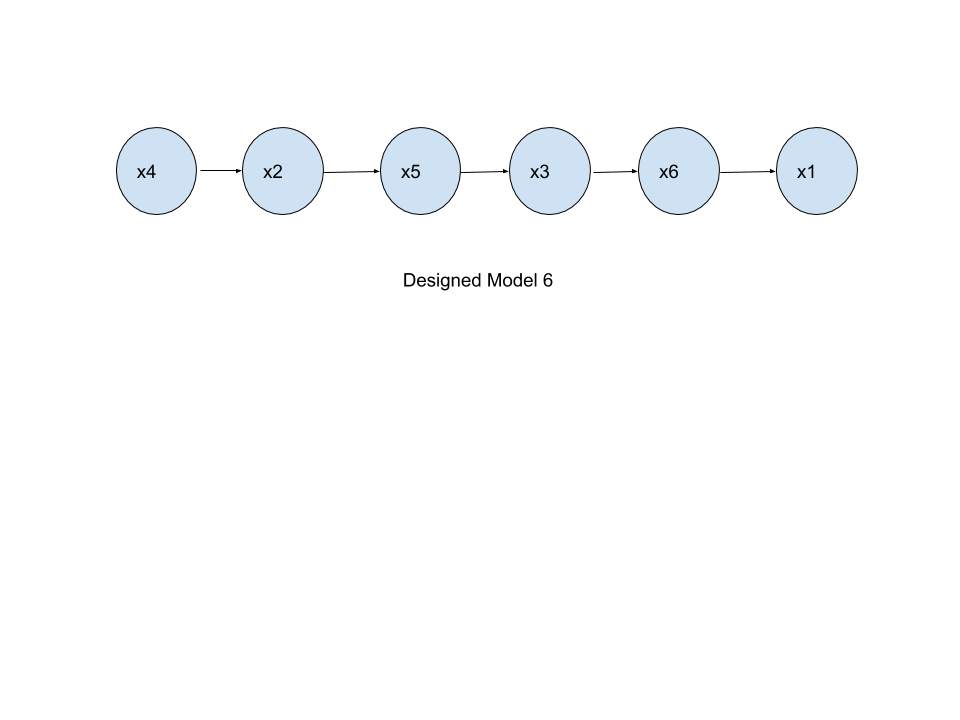
1. **Observations and Inferences for Tasks :**

**5.1 Task-1 : To Evaluate pairwise correlations and independences in the given data**

For Task-1 we got the dependences and independences as below

|  |  |
| --- | --- |
| **Variable Status :** | **Value :** |
| Independent - x1 x2: | 0.15977000000000002 |
| Dependent - x1 x4: | 0.11943000000000005 |
| Independent - x1 x6: | 0.16015500000000005 |
| Independent - x2 x3: | 0.21852500000000008 |
| Dependent - x2 x5: | 0.12926000000000004 |
| Independent - x3 x2: | 0.21875800000000004 |
| Dependent - x3 x5: | 0.11551999999999997 |
| Dependent - x3 x6: | 0.09498000000000004 |
| Dependent - x4 x1: | 0.11957000000000005 |
| Dependent - x4 x2: | 0.11569999999999997 |
| Independent - x4 x6: | 0.14347 |
| Dependent - x5 x1: | 0.12939 |
| Dependent - x5 x2: | 0.11596500000000003 |
| Independent - x6 x1: | 0.16036999999999998 |
| Independent - x6 x2: | 0.17531500000000003 |
| Dependent - x6 x3: | 0.09434000000000006 |
| Independent - x6 x4: | 0.14306999999999997 |

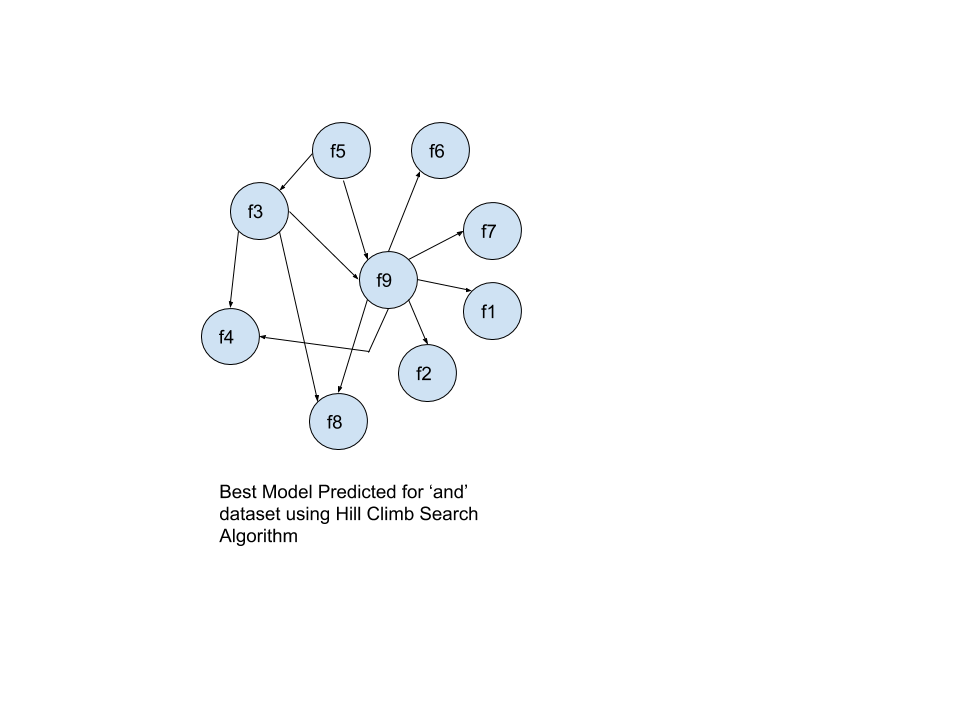
**5.2 Task-2 : Construct a Bayesian network with the fewest number of edges that maximizes the likelihood**

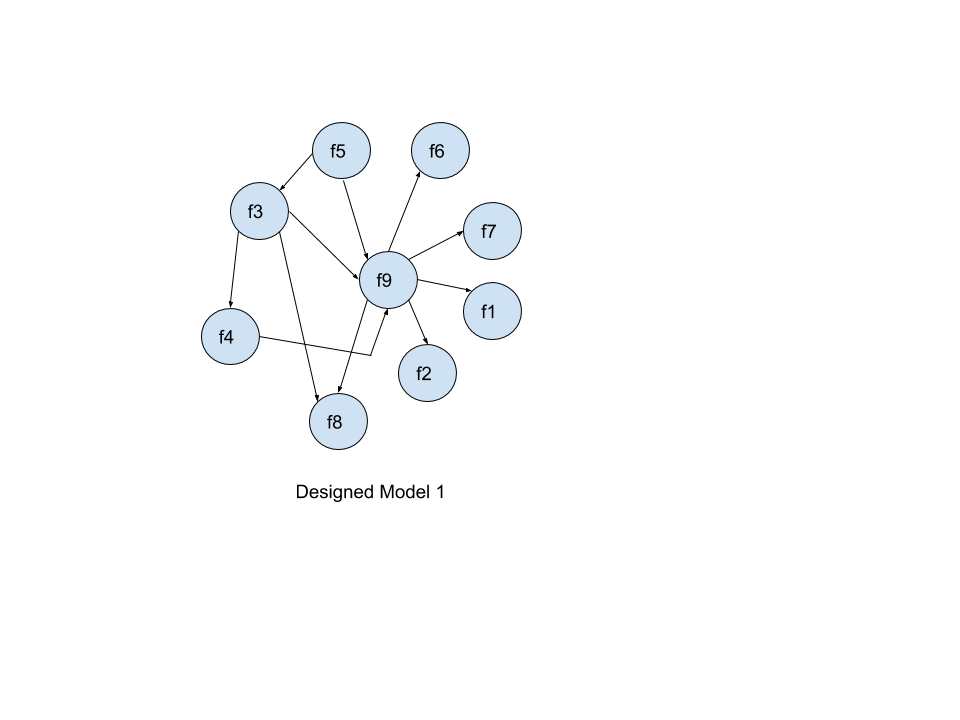
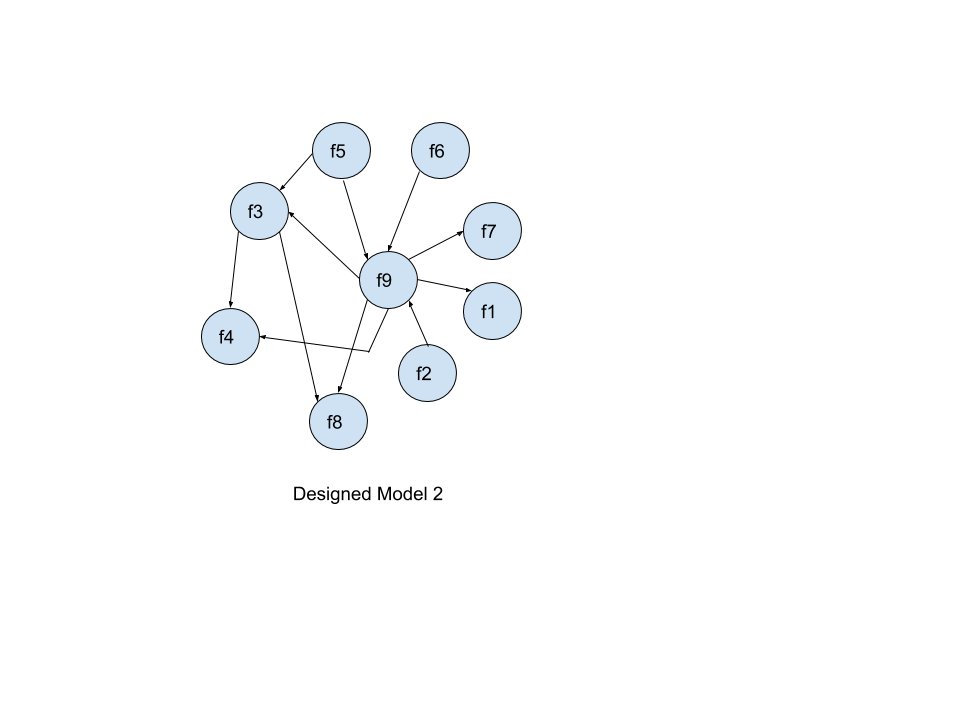
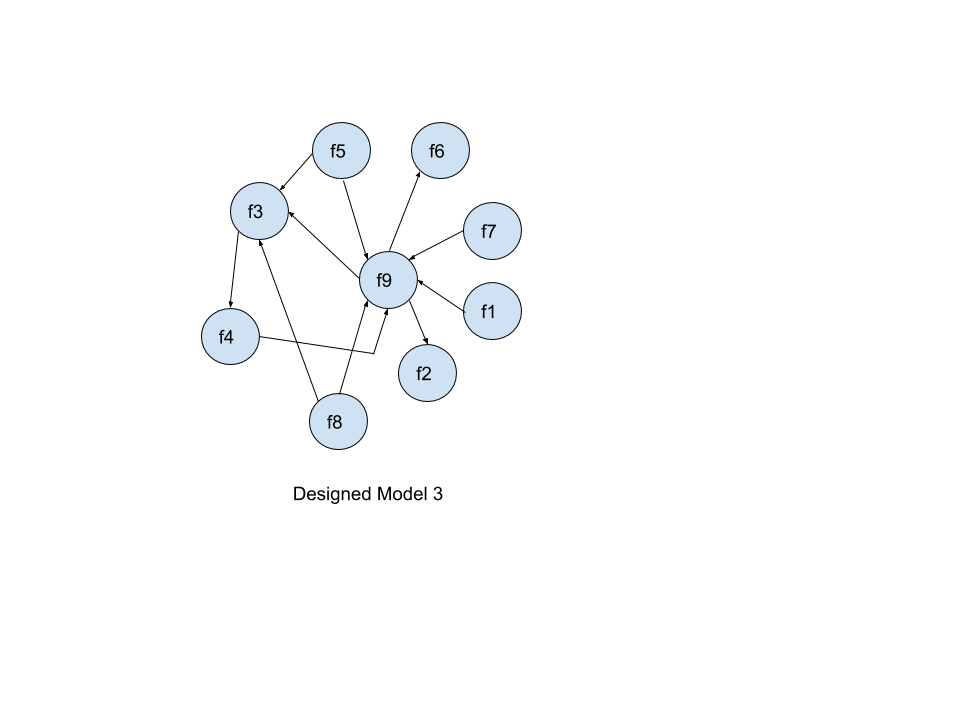
**By using the independences from Task-1 we obtain the nodes and construct the following Bayesian Models. We threshold the values to 0.14 ≥ and include all the values and nodes above it in our models to find the best model.  
  
  
Model 1 : x2 -> x5 -> x3 -> x6 -> x1 -> x4  
  
  
  
  
  
Model 2 : x5 -> x2 -> x3 -> x6 -> x1 -> x4  
  
  
  
  
  
  
  
  
  
  
  
  
  
Model 3 : x1 -> x6 (-> x4) -> x2 (-> x5) -> x3  
  
  
  
  
  
Model 4 : x4 -> x6 -> x1 -> x2 (-> x3) -> x5   
  
  
  
Model 5 : x4 -> x2 -> x5 -> x3 -> x6 -> x1   
Model 6 : x6 <- x1 (-> x4) -> x2 (-> x5) -> x3  
  
  
  
  
Comparison of K2 Score of Above Models :**

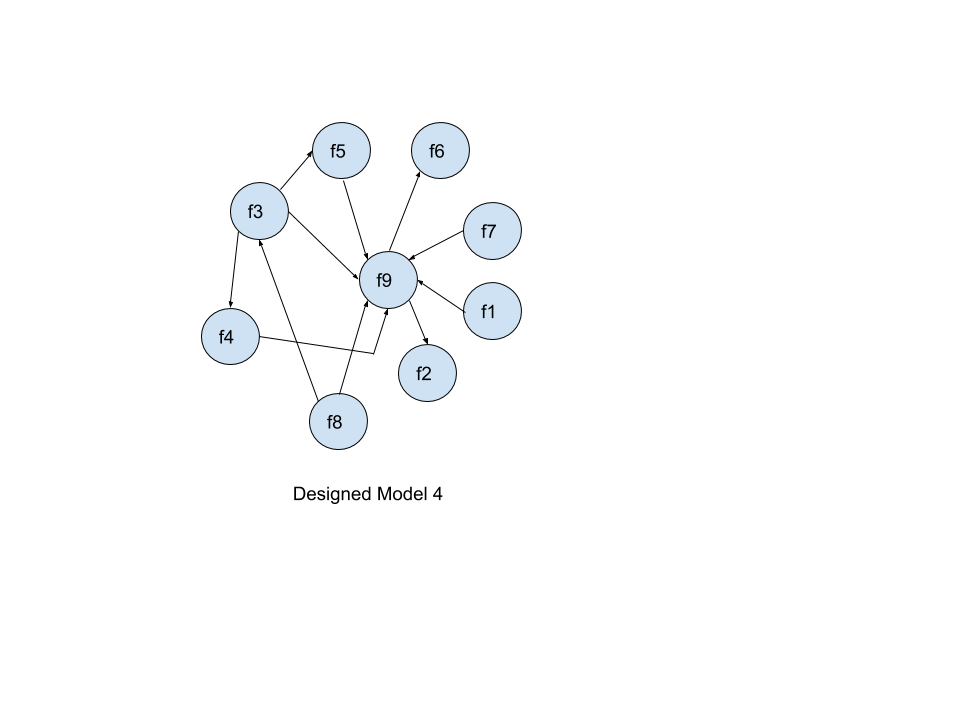
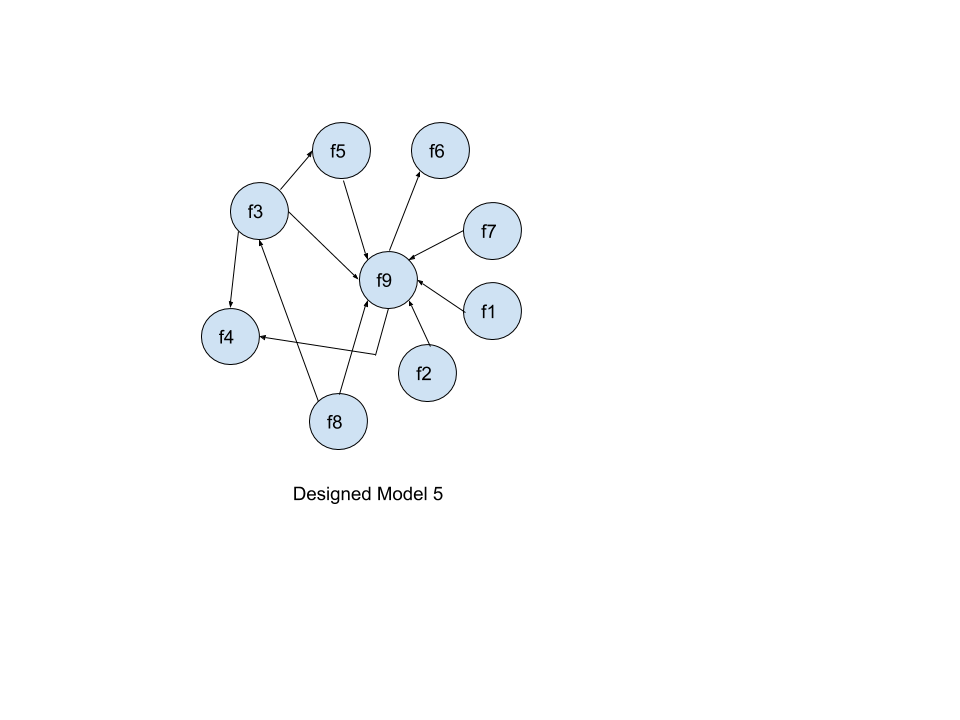
|  |  |  |
| --- | --- | --- |
| **Model** | **K2 Score Values** | **Model Status** |
| Model 1 K2 Score | -63823.400819165974 | Worst Model |
| Model 2 K2 Score | -63233.1436414088 | -- |
| Model 3 K2 Score | -62736.41002359058 | Best Model |
| Model 4 K2 Score | -62935.39000819125 | -- |
| Model 5 K2 Score | -63761.71947430618 | -- |
| Model 6 K2 Score | -63060.49203702826 | -- |

**5.3 Task-4 : Use the ‘and’ image dataset to construct a Bayesian network and evaluate the goodness score of several Bayesian Networks**

For this task we use the and dataset and the Hill Climb Algorithm to Predict the Best Mode with Max K2 Score or Least Negative K2 Score in this case. **Best Model Predicted By Hill Climb Search Algorithm :**

**  
Model 1 :**

**  
  
  
  
  
  
  
  
  
  
Model 2 :  
  
  
  
  
  
Model 3 :  
**

**Model 4 :   
  
  
  
  
Model 5 :  
  
  
  
K2 Score Comparisons for ‘and’ dataset designed Models :**

|  |  |  |
| --- | --- | --- |
| **Model** | **K2 Score Values** | **Model Status** |
| Best Model K2 Score | -9462.704892371388 | Best Model |
| Model 1 K2 Score | -9499.509888377794 | -- |
| Model 2 K2 Score | -9489.8756494027 | -- |
| Model 3 K2 Score | -9758.300190005033 | Worst Model |
| Model 4 K2 Score | -9741.224097239161 | -- |
| Model 5 K2 Score | -9745.741509234966 | -- |

**5.4 Task-3 : Convert the Best Model Bayesian Network into Markov Network using moralization. Also Compare the inferences of Bayesian and Markov Networks**

**For Bayesian Model :**

The Bayesian Model with Max K2 score and Max Probability is **Model 3 :**

K2 Score for Best Bayesian Model :

The Time Required for Best Bayesian Network is **0.09885811805725098 seconds**

Model 3 K2 Score: **-62736.41002359058**

Model 3 Edges : **('x1', 'x6'), ('x6', 'x4'), ('x6', 'x2'), ('x2', 'x5'), ('x2', 'x3')**

The Edges with Max Probability are :

The Probability of Occurrence is : **1.9300000000000002% or 0.193**

**For Markov Model :**

Time required for Markov Netwrok : **0.06509017944335938 seconds**

Nodes : **{'x2': 1, 'x5': 3, 'x3': 1, 'x6': 1, 'x1': 0, 'x4': 0}**

The Edges for best Markov Model are : ('x1', 'x6'), ('x6', 'x4'), ('x6', 'x2'), ('x2', 'x5'), ('x2', 'x3')

The Probability of Occurrence is : **1.9300000000000002% or 0.193**

**6 Conclusion :**

All the observations are noted as above for all the model both for the ‘th’ & ‘and’ datasets. The Bayesian Networks were built from the obtained nodes and also converted to Markov Networks. The Models for both ‘th’ & ‘and’ datasets were constructed and based on the K2 scores the Best Models were acquired and the probabilities of these models were given. Thus we conclude that we can determine probabilities of observations which are described by several variables using various methods and can implement other analysis on the same for time, accuracy and determine whether a model is cyclic or acyclic based on its network models to further evaluate that they belong to particular handwriting sample which is common or rare and which in turn can be useful to determine whether a particular handwriting sample was written by a certain individual.

**References**

[1] A Statistical Examination of Selected Handwriting Characteristics - *R. J. Muehlberzer et al.*

[2] <https://en.wikipedia.org/wiki/Marginal_distribution/>.

[3] <https://en.wikipedia.org/wiki/Conditional_probability_distribution/>

[4] <https://en.wikipedia.org/wiki/Graphical_model/>

[5] <https://medium.com/@neerajsharma_28983/intuitive-guide-to-probability-graphical-models-be81150da7a/>