Determining Probabilities of Handwriting Formations using PGMs

Abhishek Bhave
UBitName: abhave PersonNumber: 50289049
Department of Computer Science
University of Buffalo
Buffalo, NY 14214
abhave@buffalo.edu

Abstract

The goal of this project is to determine probabilities of observations which are described by several variables. We will work with handwriting patterns which are described by document examiners. They can be used to determine whether a particular handwriting sample is common (high probability) or rare (low probability) and which in turn can be useful to determine whether a sample was written by a certain individual. We will work on the th dataset and also test on the AND dataset.

1.1 Introduction

A probabilistic graphical model is a model which expresses the conditional dependence between random variables. Probabilistic graphical models are capable of solving inference and learning problems. Neural Networks are also capable of solving these problems but the major difference between them arises is to how they incorporate prior knowledge in the existing model. This incorporation of prior knowledge is what make PGM different from neural networks. When there is dependence between variables, graphical models can help to reduce the computation required to infer something. Probabilistic graphical models use a graph-based representation as the foundation for encoding a distribution over a multi-dimensional space. Two branches of graphical representations of distributions are commonly used, namely, Bayesian networks and Markov random fields. Bayesian models are directed graphs whereas Markov random fields are undirected graphical models.

1.2 Dataset

We are going to create the Bayesian model on the th dataset. For this purpose, we have Table2 which is the marginal probability distribution table. We also have Table 3 to table 8 which is the conditional probability distribution. Here Table 3 corresponds to the x1 random variable, Table 4 corresponds to the x2 random variable and so on till Table 8 which corresponds to the x6 random variable. These conditional probability distribution (CPD) consists of conditional probabilities of parent with respect to child. The feature definition is as follows:

A characterization of the structure of th as given by document examiners (human experts) as shown in the table below. In this characterization there are six random variables x1-x6. Variable xi can take one of a set of discrete values, denoted as xji .

Table 1: Six features of *th* and their possible values. As provided by document examiners.

x_1 **(Height Relationship of t to h)**	x_2 **(Shape of Loop of h)**	x_3 **(Shape of Arch of h)**	x_4 **(Height of Cross on t staff)**	x_5 **(Baseline of h)**	x_6 **(Shape of t)**
x_1^0: t shorter than h	x_2^0: retraced	x_3^0: rounded arch	x_4^0: upper half of staff	x_5^0: slanting upward	x_6^0: tented
x_1^1: t even with h	x_2^1: curved right side and straight left side	x_3^1: pointed	x_4^1: lower half of staff	x_5^1: slanting downward	x_6^1: single stroke
x_1^2: t taller than h	x_2^2: curved left side and straight right side	x_3^2: no set pattern	x_4^2: above staff	x_5^2: baseline even	x_6^2: looped
x_1^3: no set pattern	x_2^3: both sides curved		x_4^3: no fixed pattern	x_5^3: no set pattern	x_6^3: closed
	x_2^4: no fixed pattern				x_6^4: mixture of shapes

Figure 1 Structure of th

1.3 Probabilities in PGM

The main aim of Probabilistic Graphical Models is to provide an intuitive understanding of joint probability among random variables. We have marginal and conditional probabilities in the dataset. Below we will describe the marginal probability and the conditional probability distribution. In simple words marginal probability applies to those random variables that do not have a dependency with any other random variables. In the below example marginal probability is applied to c as it has no dependency i.e. it does not depend on any other random variable i.e. it does not have any parent nodes.

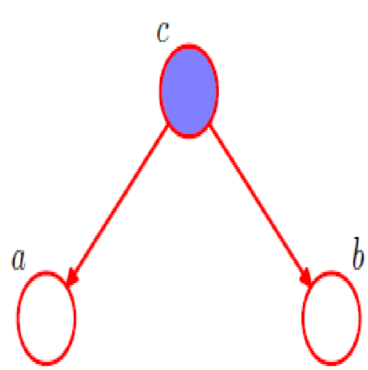

Figure 2

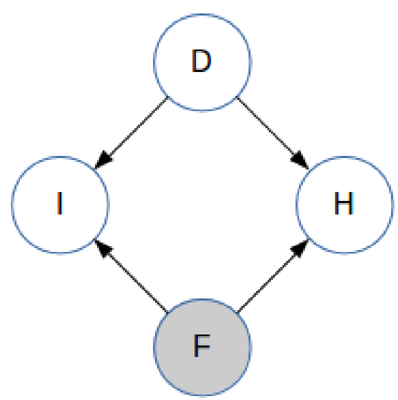

Figure 3

The **marginal probability** is the probability of occurrence of a single event. In calculating marginal probabilities, we disregard any secondary variable calculation. As we can see from the above figures, in Figure 1 c is has marginal probability and in Figure 2 nodes D and F have marginal probability.

A **conditional probability** is the probability that an event will occur given that another specific event has already occurred. We say that we are placing a condition on the larger distribution of data, or that the calculation for one variable is dependent on another variable. In the above 2 figures, in figure 1 a and b will have conditional probability as they are dependent on node a. The probability of a is p(a/c) and that of b is p(b/c).

For each random variable we have a set number of conditional distributions. In our project we have a total of 17different conditional probability distributions. Some of them are as follows: p(x2/x3), where x2 is the shape of the h loop, p(x4/x1), where x4 is the height of cross on t staff.

We would like to create a probabilistic graphical model (PGM) so that we can evaluate the probability of any given combination of the six feature values of th. This process of evaluation is called the process of inference. We will follow this process of inference in probabilistic graphical models.

2 Task 1 Data Preprocessing

2.1 Determining Correlations and Independencies

Evaluate pairwise correlations and independences that exist in the data. Note that we can determine whether xi and xj are independent by testing if p(xi, xj) = p(xi)p(xj), where the joint probability between a pair of variables can be determined from the tables as p(xi,xj) = p(xi|xj)p(xj)

In this task we have to calculate the closeness of P(x,y) and P(x)P(y) by using entropy. Below is the method I have used to calculate closeness of P(x,y) i.e. dependencies between them.

$$\sum abs\left(\left(P(x,y) - P(x)P(y)\right)\right)$$

To calculate the entropy or closeness I have first used regular expressions or regex to clean the datasets. One of the examples of this is 20.5%(32) which is one of the values in the Table3. To process this data, I have removed everything after % which gave me 20.5 and then divided all the values by 100 which gave me the value as 0.205 which I added to the CPD table of x1.

This completed the data cleaning part of the project. Next I calculated the marginal probability P(x,y). Here I used the np.outer() function to get the marginal probabilities. Next, I calculated the conditional probablilty and I got the P(x)P(y) then I subtracted the marginal and the conditional probability to get a temporary result. I used the absolute function and then calculated the sum of all the values to generate 17 values which is the final result of Task 1.

The 17 CPDs are as follow:

```
{'x2/x1': 0.15977, 'x4/x1': 0.11943000000000004, 'x6/x1': 0.16015500000000005
'x3/x2': 0.21852500000000002, 'x5/x2': 0.12926000000000004, 'x2/x3':
0.21875800000000006, 'x5/x3': 0.11551999999999997, 'x6/x3':
0.11324000000000001, 'x1/x4': 0.11957000000000002, 'x2/x4':
0.11569999999999997, 'x6/x4': 0.14346999999999996, 'x2/x5':
0.8561449999999998, 'x3/x5': 0.11670000000000004, 'x1/x6':
0.17684499999999995, 'x2/x6': 0.17531500000000003, 'x3/x6':
0.13903000000000004, 'x4/x6': 0.14307000000000003}
```

Figure 4 CPD

3 Task 2 Bayesian network construction and 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:

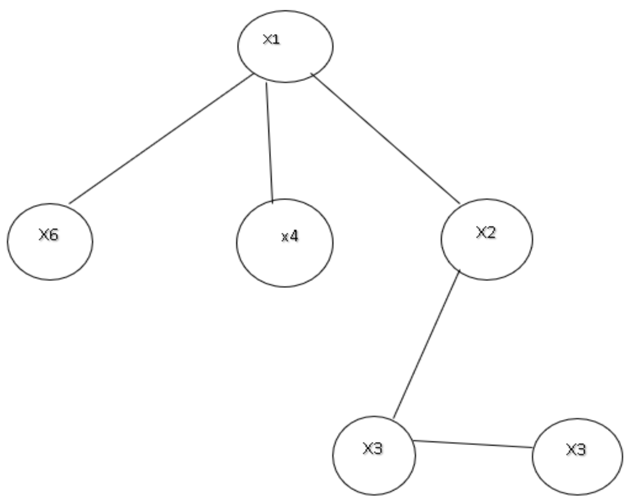

Figure 5

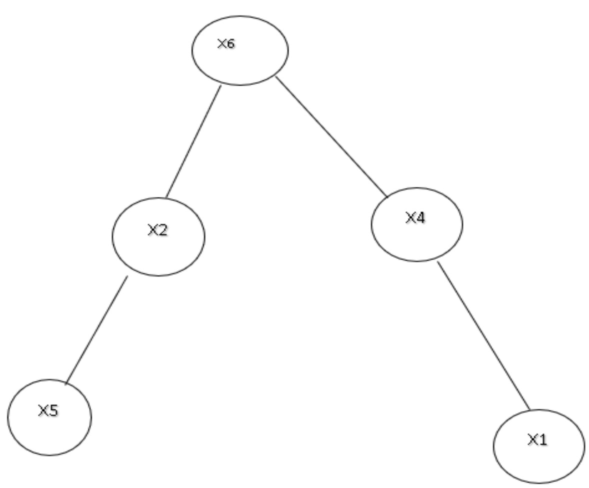

Figure 6

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

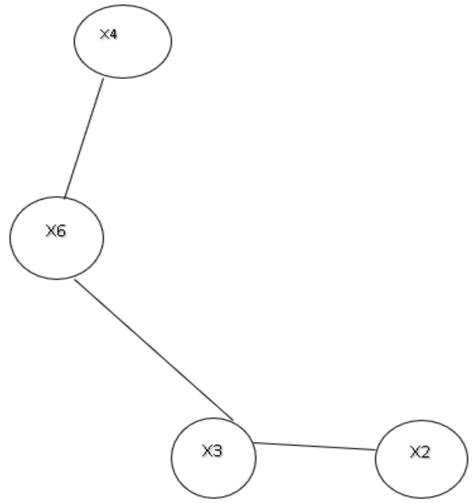

Figure 7 Best Model

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:

```
Bayesian Model Inference
Model Query Time: 0.00624
+------+-----------+
| x2   |   phi(x2) |
+======+===========+
| x2_0 |    0.2620 |
+------+-----------+
| x2_1 |    0.3090 |
+------+-----------+
| x2_2 |    0.0000 |
+------+-----------+
| x2_3 |    0.1670 |
+------+-----------+
| x2_4 |    0.2620 |
+------+-----------+
```

```
Markov Model Inference
Model Query Time: 0.005169
+------+-----------+
| x2   |   phi(x2) |
+======+===========+
| x2_0 |    0.2620 |
+------+-----------+
| x2_1 |    0.3090 |
+------+-----------+
| x2_2 |    0.0000 |
+------+-----------+
| x2_3 |    0.1670 |
+------+-----------+
| x2_4 |    0.2620 |
+------+-----------+
```

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.com/@neerajsharma_28983/intuitive-guide-to-probability-graphical-models-be81150da7a

[2]- https://blog.statsbot.co/probabilistic-graphical-models-tutorial-and-solutions-e4f1d72af189

[3]- https://en.wikipedia.org/wiki/Graphical_model