PROJECT 1:DETERMINING PROBABILITIES OF HANDWRITING FORMATIONS USING PGMS

Abstract

The aim is to develop probabilistic graphical models (PGMs) to determine probabilities of observations which are described by several variables. We use handwriting patterns which are described by document examiners. They are used to check whether a particular handwriting sample is common (high probability) or rare (low probability) and which in turn is used to determine whether a sample was written by a certain individual. We consider only the letter pair 'th' , since it is the most commonly encountered pair of letters called bigram in English .

1. Task 1:

1.1 Determining independence and correlations between variables:

- Given six variables x1-x6 we determine whether every pair of variables is independent or not using the following formulas.
- For two variables to be independent , P(xi,xj)=P(xi)P(xj).The joint probability between the pairs of variables is given by, P(xi,xj)=P(xj|xi)P(xi).
- We subtract the independent probability from the joint probability and check whether the subtracted value is lesser than a particular threshold value while constructing bayesian network.
- The results obtained on applying the independent and the joint probabiltiy formulas are as follows:

Table 1: Result of independent and joint probability distributions

Pairs of variables	Value
X1,X2	0.15977
X1,X4	0.11943
X1,X6	0.160155
X2,X3	0.218525
X2,X5	0.13246
X3,X2	0.218758
X3,X5	0.11552
X3,X6	0.11324
X4,X1	0.11957
X4,X2	0.1157

X4,X6	0.14347
X5,X2	0.8561
X5,X3	0.1167
X6,X1	0.2181249999
X6,X2	0.1755315
X6,X3	0.13903
X6,X4	0.14307

2. Task 2:

2.1 Constructing bayesian network:

Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian networks aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors.

2.2 Determining variables which are independent:

- We set a threshold value on the previously calculated result to determine if two variables are independent or not.
- The threshold value ranges between 0 and 1.
- If the value calculated is less than the threshold value then the pairs of variables are said to be independent.
- Otherwise the pairs of variables are said to be dependent.

Case1 : On setting the threshold value to 0.14 :

Result:

- **Dependent pairs** :(x1, x2) (x2, x5) (x4, x1) (x4 , x6) (x6, x3).
- The remaining pairs of variables with values less than 0.14 were found to be independent.

Case 2 : On setting the threshold value to 0.13 :

Result:

- **Dependent pairs** :(x1, x2) (x1, x6) (x2, x5) (x2 , x3) (x6, x4)
- The remaining pairs of variables with values less than 0.14 were found to be independent.

2.3 Creating models:

2.3.1 Model 1:

- From case 1, we construct a bayesian model with the dependent pairs of variables as the edge set.
- Add_cpds() is used to add the cpds for variables in the edge set.

- If there is no loop and the graph constructed is a directed acyclic graph (DAG) on checking the model using check_model() function the result will be true.
- Otherwise the result will be false.

2.3.2 Model 2:

- From case 2, we construct a bayesian model with the dependent pairs of variables as the edge set.
- Add_cpds() is used to add the cpds for variables in the edge set.
- If there is no loop and the graph constructed is a directed acyclic graph (DAG) on checking the model using check_model() function the result will be true, otherwise the result will be false.

2.3.3 Comparing models:

- K2 score is used to compare the different models created.
- Since we do not have the dataset needed to calculate the k2 score we generate the dataset.
- Sampling method , in particular forward or ancestral sampling is used to generate the dataset.
- Using ancestral sampling , we start from parent in order of topological sort which samples out the parent first and then the children.
- The number of variables to be sampled out in our case is six, after which we get one data.
- We create a dataset of size 3000 with the return type as dataframe.
- The same steps are repeated for models 1 and 2 ,where each model creates a different dataset.

2.3.4 K2 Score:

Score method measures how well a model is able to describe a given dataset. Therefore the model with highest K2 score is the best model. It measures how well a variable is influenced by a set of potential parents. K2.score() method is used to calculate the score.

Result:

Model 1 K2 Score : -19176.12997380593
Model2 K2 Score: -18924.581645560553 [BEST MODEL]

Conclusion:

- Model 2 has the highest K2 score.
- Model 2 is able to describe well the dataset generated more than model 1, making it a better model.

2.3.5 Inference:

Variable elimination is used to infer from the model created.

1. **Model 1:**

Table 2: Result of Inference (X3)

X3	**Phi(X3)**
X3_0	0.1757
X3_1	0.6479
X3_2	0.1763

Table 3: Result of Inference (X5|X2_1)

X5	**Phi(X5)**
X5_0	0.3374
X5_1	0.1101
X5_2	0.1263
X5_3	0.4263

2. Model 2:

Table 4: Result of Inference (X4 in Bayesian Model)

X4	**Phi(X4)**
X4_0	0.7153
X4_1	0.1049
X4_2	0.0100
X4_3	0.1698

Table 5: Result of Inference (X4|X6_2 in Bayesian Model)

X4	**Phi(X4)**
X4_0	0.7860
X4_1	0.1072
X4_2	0.0000
X4_3	0.1068

3. Task 3:

- From the result obtained in Task 2 we know that Model 2 is the best model with highest K2 Score.
- This model is converted into a markov model using moralization.
- It converts the given directed graph into a undirected graph.
- Edges : (x1, x2), (x1, x6), (x2, x5), (x2, x3), (x6, x4)
- We infer the same values as in the inference of Bayesian version of model 2 to compare.

3.1 Result of Markov version of Best Model (Model 2):

Table 6: Result of Inference (X4 in Markov Model)

X4	**Phi(X4)**
X4_0	0.7152
X4_1	0.1048
X4_2	0.0101
X4_3	0.1699

Table 7: Result of Inference (X4|X6 _2 in Markov Model)

X4	**Phi(X4)**
X4_0	0.7860
X4_1	0.1072
X4_2	0.0000
X4_3	0.1070

3.2 Comparison between inferences :

The best bayesian model (Model 2) was converted into Markov model. We infered the same values X4 and X4|X6_2 from the created markov model. The result seems to differ by a negligible amount in the markov model. Most of the values either increased or decreased by 0.0001 or 0.0002. Some results were the same as the bayesain model values.

3.3 'And' dataset conversion:

- The bayesian network is constructed for the "and" dataset.
- Using hillclimb search we get the best bayesian model with the highest K2 score.
- This best model is converted into markov model using moralization.

Edges for markov model using 'And ' dataset :

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

4. Task 4:

- Using the 'And' dataset we find the best model using HillClimb Search.
- Even though Exhaustive search can be used , Hill Climb search is more efficient.
- Two more models are created with different set of edges.
- Both of these models have a lesser K2 score than the Best Model.

Results:

4.1 Best Model:

Edges: (f 3, f 4), (f 3, f 9), (f 3, f 8), (f 5, f 9), (f 5, f 3), (f 9, f 8), (f 9, f 7), (f 9, f 1), (f 9, f 6), (f 9, f 2), (f 9, f 4)
K2 score: -9462.704892371388

4.2 Model 1:

Edges: (f 1, f 2), (f 3, f 4),(f 4, f 6), (f 6, f 7),(f 2,f 5),(f 1, f 9),(f 9, f 3),(f 9, f 8)
K2 score : -9660.314507402692

4.3 Model 2:

Edges : (f 4, f 8), (f 2, f 8),(f 3, f 1), (f 9, f 1),(f 7, f 9),(f 4,f 1),(f 6,f 7),(f 5,f 4)
K2 Score: -9801.832643103877