Project 1 Report

Topic: Student Performance Evaluation

Dataset: Turkiye Student Evaluation Dataset

Source

Ernest Fokoue
Center for Quality and Applied Statistics.

Necla Gunduz
Department of statistics Faculty of Science, Gazi University

By
Manisha Tadikonda

(50207628)

Puneeth Pepalla

(50206906)

Probabilistic Graphical Models
Bayesian Networks

1.Aim: The goal of this project is to implement and evaluate algorithms pertaining to inference with Probabilistic Graphical Models(Directed PGMs in particular) and use that model/algorithm to make meaningful deductions on a dataset.

2.Approach: The project is done in python by using a library called PGMPY. For this project we use Directed probabilistic graphical models (Bayesian Networks).Here we use Bayesian Network because the variables in our dataset exhibit some causality towards certain variables. For example, if the professor teaches the class by giving more real examples and creates interest to the student in that course then the students are more likely to attend the class daily.

3.Project Domain: The project domain is Student Performance Evaluation where the variables correspond to opinion of the student on different course specific questions like difficulty of the course, participation in the course etc.
Domain Knowledge: This is a very generic topic and no in-depth domain knowledge is required. This is domain which can be easily understood by everyone if they can understand the flow of probability among various random variables.

4.Dataset: The dataset is actually a survey that is conducted on the students of Gazi University in Ankara(Turkey). This data set contains a total 5820 evaluation scores(observations) provided by students from Gazi University in Ankara (Turkey). There is a total of 28 course specific questions(attributes) and additional 5 attributes. So based on the rest of the data we can classify any of the 5 additional Attributes.
Each attribute is a specific course related question and students were asked to rate the question or answer the question on a scale of 5.(for example, if the question is difficulty of the course, then if the course is not so difficult then the student gives 1 and if the student feels the course is difficult then he would give a score of 5 to the course. There are 28 such questions. Here the variables are discrete which take the values in the range 0,1,2,3,4,5 and class in the range 1-13. The details of the dataset are given below.

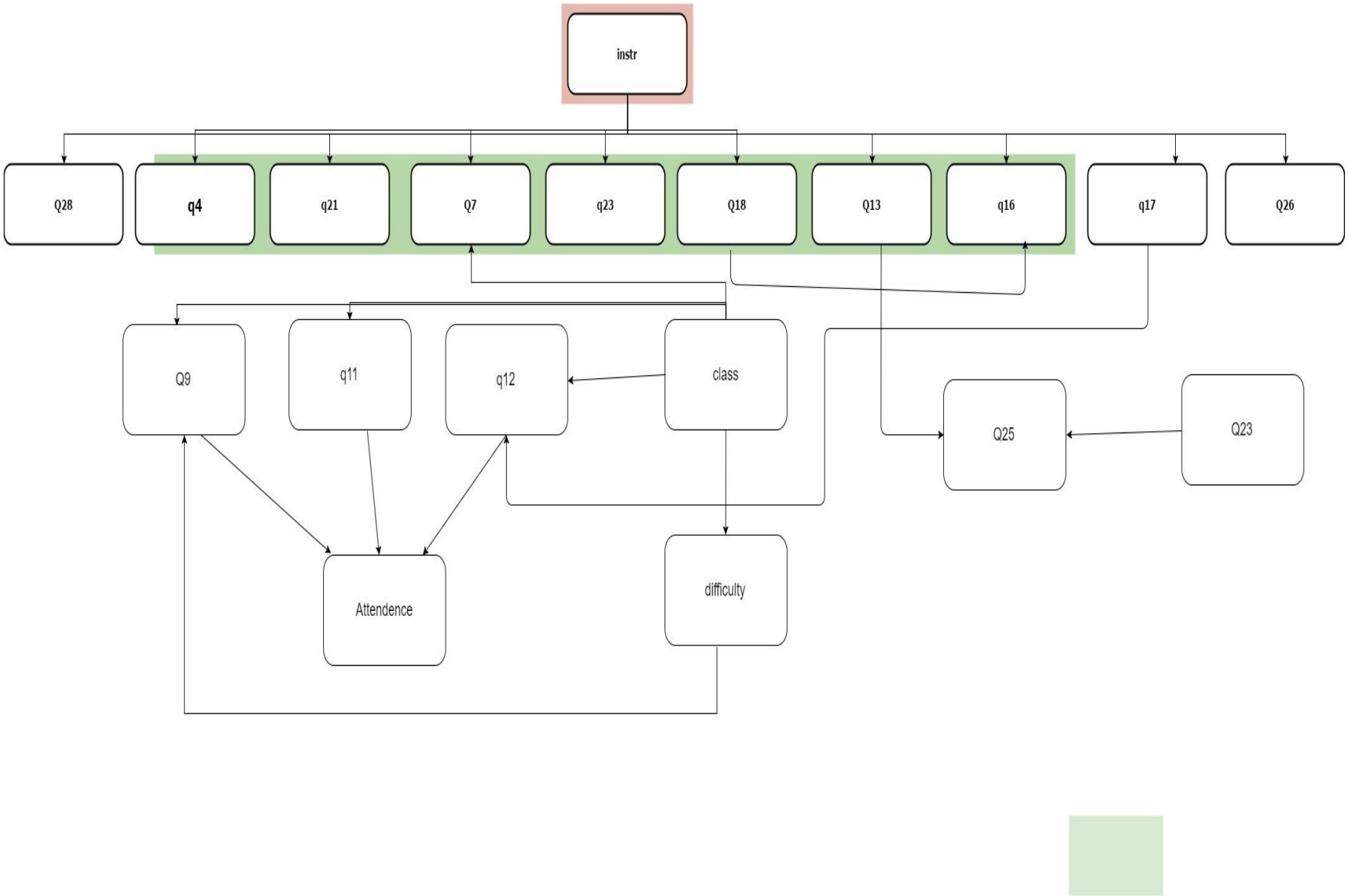
5.Attribute Information:
instr: Instructor's identifier; values taken from {1,2,3}
class: Course code (descriptor); values taken from {1-13}
repeat: Number of times the student is taking this course; values taken from {0,1,2,3,...}
attendance: Code of the level of attendance; values from {0, 1, 2, 3, 4}
difficulty: Level of difficulty of the course as perceived by the student; values taken from {1,2,3,4,5}

Q1: The semester course content, teaching method and evaluation system were provided at the start.
Q2: The course aims and objectives were clearly stated at the beginning of the period.
Q3: The course was worth the amount of credit assigned to it.
Q4: The course was taught according to the syllabus announced on the first day of class.
Q5:The class discussions, homework assignments, applications and studies were satisfactory.
Q6: The textbook and other courses resources were sufficient and up to date.
Q7: The course allowed field work, applications, laboratory, discussion and other studies.
Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
Q10: My initial expectations about the course were met at the end of the period or year.
Q11: The course was relevant and beneficial to my professional development.
Q12: The course helped me look at life and the world with a new perspective.
Q13: The Instructor's knowledge was relevant and up to date.
Q14: The Instructor came prepared for classes.
Q15: The Instructor taught in accordance with the announced lesson plan.
Q16: The Instructor was committed to the course and was understandable.
Q17: The Instructor arrived on time for classes.
Q18: The Instructor has a smooth and easy to follow delivery/speech.
Q19: The Instructor made effective use of class hours.
Q20: The Instructor explained the course and was eager to be helpful to students.
Q21: The Instructor demonstrated a positive approach to students.
Q22: The Instructor was open and respectful of the views of students about the course.
Q23: The Instructor encouraged participation in the course.
Q24: The Instructor gave relevant homework assignments/projects, and guided students.
Q25: The Instructor responded to questions about the course.
Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.
Q27: The Instructor provided solutions to exams and discussed them with students.
Q28: The Instructor treated all students in a right and objective manner.

Q1-Q28 are all Likert-type, meaning that the values are taken from {1,2,3,4,5}

6.Bayesian Network:

The Bayesian Network for our given dataset is as follows. Here the causality between variables is represented by a directed arrow between them. Some of the attributes in the dataset exhibit causality and some don’t. The variables that exhibit causality are mentioned below. The below Bayesian network is constructed by intuition on how the causalities exist between variables. The variable on which the attribute depends is called the parent node and the node is known as child node.
There are totally 21 variables in which not all the variables should exhibit causality
Some variables exhibit causality and some variables don’t.


instr
Q28
q4
q21
Q7
q23
Q18
Q13
q16
q17
Q26
Q9
q11
q12
class
Q25
Q23
Attendence
difficulty

7.Construction of model:

The second step in this project is construction of the model after obtaining the Bayesian network. For developing the model we used a predefined libraries called PGMPY and LIBPGM. The steps involved in developing the model are as follows:

1.For constructing the Bayesian model we use a single line code from PGMPY library called BayesianModel that can be found in pgmpy.models

Code -> bayesmodel = BayesianModel ([(variables that exhibit causality((parent1,child1), (parent2,child2) ,(parent3,child3)……)

2.And after obtaining the above model we have to fit that model by using Maximum Likelihood Estimator which can be imported from pgmpy.estimators

Code-> model = bayesmodel.fit(df(data from the original input file of dataset), estimator=MaximumLikelihoodEstimator)

2.And from this model we get conditional CPD's by using the get_cpds() function form BayesianModel.

code

```
for cpd in bayesmodel.get_cpds():
    print("CPD of {variable}:".format(variable=cpd.variable))
    print(cpd)
```

3.By using the above code we can get fit a Bayesian model depending on their causalities given its dependencies and obtain conditional probabilities from the model.

The CPDs for some of the attributes are as follows.

```
CPD of Q13:
+--------+----------------+---------------+---------------+
| instr  | instr(1)       | instr(2)      | instr(3)      |
+--------+----------------+---------------+---------------+
| Q13(1) | 0.14064516129  | 0.101108033241| 0.15940016662 |
+--------+----------------+---------------+---------------+
| Q13(2) | 0.0825806451613| 0.084487534626| 0.123576784227|
+--------+----------------+---------------+---------------+
| Q13(3) | 0.234838709677 | 0.262465373961| 0.30380449875 |
+--------+----------------+---------------+---------------+
| Q13(4) | 0.303225806452 | 0.353185595568| 0.267425715079|
+--------+----------------+---------------+---------------+
| Q13(5) | 0.238709677419 | 0.198753462604| 0.145792835324|
+--------+----------------+---------------+---------------+
```

CPD of Q28:

instr	instr(1)	instr(2)	instr(3)
Q28(1)	0.138064516129	0.102493074792	0.154679255762
Q28(2)	0.0903225806452	0.0720221606648	0.107192446543
Q28(3)	0.241290322581	0.264542936288	0.289641766176
Q28(4)	0.289032258065	0.33864265928	0.272702027215
Q28(5)	0.241290322581	0.222299168975	0.175784504304

CPD of class:

class(1)	0.0520619
class(2)	0.024055
class(3)	0.155326
class(4)	0.0321306
class(5)	0.112715
class(6)	0.0958763
class(7)	0.0321306
class(8)	0.0859107
class(9)	0.09811
class(10)	0.0769759
class(11)	0.0831615
class(12)	0.00704467
class(13)	0.144502

8.Queries:
We can answer different type of queries from the Bayesian model we obtained. These queries can be used to rate a professor, rate a class, test the intelligence of a student. This project would be helpful for those who have trouble in selecting the courses for the semester based on the workload and difficulty. Some people might select different courses without appropriate knowledge and end up getting bad grades. So, this could be helpful to many students seeking a way to help them in selecting courses and guide them in various course related problems. Some example queries that can be answered using this project/model are:

Types of queries that can be answered:

1. P(Attendance | Interest in course=4, difficulty=3) -> This query assigns the probability for all the 5 classes of attendance and finally outputs the greatest of all possibilities.
2. P(nb.repeat / attendance=1, difficulty=5, evaluation=1) -> This marginalizes the variables by using variable elimination with attendance=1,difficulty=5 and evaluation=1 and from the rest it outputs the probabilities for all values of nb.repeat and selects the value with maximum probability.
3. P(course/difficulty=3, instr=1, attendance=4, nb.repeat=1, interest=3) -> This outputs the probabilities for all the values of course and selects the course with the maximum probability.

9. Inference Algorithms Used: The inference algorithms used in the project are:
Approximate Inference algorithm:
Belief Propogation: Belief propagation is also known as sum-product message passing, which is a technique(algorithm) for performing inference on graphical models, such as Bayesian networks and Markov random fields. It calculates the marginal distribution for each unobserved node, conditional on any observed nodes. Belief propagation is commonly used in artificial intelligence and information theory and has demonstrated empirical success in numerous applications including low-density parity-check codes, turbo codes, free energy approximation, and satisfiability.
This is the code for Beliefpropogation using pgmpy library
belpro = BeliefPropagation(bayesmodel).
Since our data set doesn't have any continuous variables, we had to resort to approximate inference from variable elimination and exact inference and we use belief propogation to obtain the maximum value of variables given their evidences.

10.Results:
The below code is used to execute the queries. Given, the variable to be found and the evidence variables given:
print(belpro.map_query(variables=['attendance'],evidence={'difficulty':2,'Q9':3}))

(Examples) The results for the above queries:

1.print(belpro.map_query(variables=['attendance'],evidence={'difficulty':2,'Q9':3})
)
2.print(belpro.map_query(variables=['attendance','Q9','difficulty'],evidence={'class
':7}))

```
{'attendance': 3}
{'Q9': 2, 'difficulty': 2, 'attendance': 3}
```

11.Bayesian Model Sampling:

Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. This sequence can be used to approximate the joint distribution (generating a histogram of the distribution),to approximate the marginal distribution of one of the variables, or some subset of the variables (for example, the unknown parameters or latent variables); or to compute an integral. We tried gibbs sampling and ancestral sampling to sample our data set. After analyzing we found out that those methods of sampling were not appropriate for our datset because our data set consists of discrete variables.So, we used Bayesian Sampling model. The three methods present in Bayesian sampling model are

1.Likelihood weighted sample
2.Rejection Sample and
3.Forward Sample.

We used forward sampling to obtain samples from our data set.

	class	difficulty	Q11	Q9	instr	Q21	Q17	Q12	attendance	Q13	Q28	\
0	6	4	4	0	2	2	0	0	2	1	4	
1	5	2	3	2	1	4	3	3	2	0	2	
2	3	0	2	3	2	2	1	1	2	4	2	
3	9	2	0	2	2	1	4	4	3	2	2	
4	0	1	2	1	2	2	0	0	3	3	3	

	Q18	Q16	Q7	Q4	Q23	Q25	Q26
0	2	2	3	2	1	1	3
1	3	3	1	2	4	4	2
2	3	3	0	3	2	4	4
3	4	4	3	4	2	2	1
4	0	0	2	1	4	4	4

Mean: Mean of the columns of the given data set:

```
class          4.6
difficulty     1.8
Q11            2.2
Q9             1.6
instr          1.8
Q21            2.2
Q17            1.6
Q12            1.6
attendance     2.4
Q13            2.0
Q28            2.6
Q18            2.4
Q16            2.4
Q7             1.8
Q4             2.4
Q23            2.6
Q25            3.0
Q26            2.8
```

12.Entropy

Entropy, in general is defined as the randomness or noise in a distribution.
In order to fully specify the Bayesian network and thus fully represent the joint probability distribution, it is necessary to specify for each node X the probability distribution for X conditional upon X's parents. The distribution of X conditional upon its parents may have any form. It is common to work with discrete or Gaussian distributions since that simplifies calculations. Sometimes only constraints on a distribution are known; one can then use the principle of maximum entropy to determine a single distribution, the one with the greatest entropy given the constraints. (Analogously, in the specific context of a dynamic Bayesian network, one commonly specifies the conditional distribution for the hidden state's temporal evolution to maximize the entropy rate of the implied stochastic process.) Relative entropy can be found out by

- It can be estimated using N samples as:

$$\hat{KL}(p||q) = -\frac{1}{N}\sum_i \left[\ln q(\mathbf{x}_k) - \ln p(\mathbf{x}_k)\right].$$

```
[ 1.31512011  1.27302834  1.34211318  1.32088834  1.58109375  1.51570795
  0.97431475  0.97431475  1.58902692  1.27985423  1.56495725  1.35797785
  1.35797785  1.31078368  1.5171064   1.49856908  1.50659509  1.51245118]
```

13.Conclusion:

From this project we implemented various sophisticated algorithms like belief propogation and gibbs sampling and made some sense of raw data and the data is given as an input to the Bayesian model and this helps in transformation of the data to some useful results.

On the whole we tried to model the dataset using Bayesian models using the causality exhibited by the variables.We tried to apply approximate inference using belief propogation and sample the data and tried to answer the queries limited to our dataset that would comprehend the dataset in a better way.