Bayesian Machine Learning for Gen AI

Assignment - I

October 12, 2025

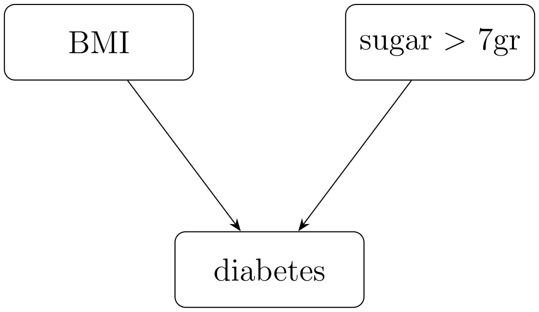
NOTE: If you can solve this assignment, you can confidently say that you have a good understanding and applied capacity for Directed Acyclic Graphs (DAGs), learning DAGs, inference and diagnosis, likelihood, and model comparison.

Problem 1

Download the simulated dataset `diabetes_data.csv`. This dataset involves records of 10,000 simulated patients who answered some questions about their lifestyle and underwent some medical tests. The dataset has the following columns:

- `sugar`: If the patient uses more than 7gr of sugar on a typical day. (yes, no)
- `exercise`: How much the patient exercises (high, medium, low)
- `BMI`: Body Mass Index of the patient (high, medium, low)
- `fatty liver`: Fat level on the patient's liver (high, low)
- `anxiety`: Anxiety level of the patient (high, low)
- `diabetes`: Whether the patient has diabetes (yes, no)

Clinicians at your institution decided that the most important factors for diabetes are `body mass index` and `daily sugar intake (whether it is greater than 7gr)`. Therefore, you establish the following Bayes Net:



(a) (20p) Build this model and learn the conditional probability tables (CPTs) from data.

(b) (10p) Given a patient is diagnosed with diabetes, what is the probability that their daily sugar intake is high?

Problem 2 (40p)

In this question, you will use the following 3 estimators from the `pgmpy` library:

- PC (Constraint-Based Estimator)
- Hill Climb Search
- Tree Search

(a) Split the data into 80% training and 20% validation sets.

(b) Using each of these structure learning methodologies, search for a better model using the training data. Draw the Bayes Nets (nodes and edges) for each model and calculate their conditional probability tables.

Problem 3 (10p)

Using the `structure_score` function from `pgmpy.metrics`, calculate the BIC scores of the three models (from the three search algorithms) over the validation data. Which model do you choose? Briefly comment on the score.

Reminder

Please remember our discussion in class about system evaluation by looking at the BIC score. In proper definition of BIC score, the probabilistic term is the negative log likelihood, making the smaller better. But, as you will see pgmpy will return the log likelihood (as you can suspect from the very large and negative number, which is the log of a very very small number - i.e. the joint probability).

Problem 4 (20p)

Using the best model from Question 3, answer the following:

(a) How much does increasing exercise levels, from "low" to "high", reduce the likelihood of experiencing anxiety?

(b) How does increasing sugar intake affect the likelihood of experiencing anxiety? Similarly, how does increasing sugar intake affect the likelihood of experiencing diabetes?

(Provide two QUANTITATIVE answers.)