CS 561 Assignment #1

Due date: 31st March 2024 (Sunday)

Exercise 1: The aim of this exercise is to understand the Bayesian Networks and the inference techniques.

Consider the task of credit card fraud detection system, where the objective is to detect fraudulent credit card transactions. Typically a credit card holder's transactions follow a certain pattern. If there is a disparity in the patterns then fraud is likely to happen. In this assignment, you will be implementing a simple credit card fraud detection system using Python.

Part A: Designing the Bayesian Network [20 Marks]

The following information is provided to you:

When a customer is travelling abroad it is more likely for a transaction to be fraudulent due to various reasons such as card theft or lost card. On average, a customer is travelling abroad 5% of the time. When the customer is travelling, then 88% of the transactions are foreign purchases regardless of the legitimacy of the transactions. On the other hand, when the customer is not travelling then mere 0.01% of transactions are foreign transactions. Also, online transactions are more likely to be fraudulent when a customer does not own his/her laptop or smartphone. Currently, 70% of the customers own a laptop or smartphone and for those customers, 40% of their transactions are done over the Internet. While for customers without their own device only 5% of transactions are online. Typically, 0.5 % of transactions are legitimate when the purchase is made online abroad, 20% of legitimate transactions are online purchases made within the country, 15% of legitimate transactions are purchases made abroad at POS, 25% of transactions are legitimate when a purchase is made within the country at POS.

Construct a Bayesian network using the following binary random variables and the probabilities as described above.

Fraud (F): current transaction is fraudulent

OwnsDevice (OD): customer owns a laptop or smart phone

Travel (T): customer is currently travelling

ForeignPurchase (FP): current transaction is a foreign purchase

OnlinePurchase (OP): current purchase is made online

For this part, you can use the Pomegranate Python package
https://pomegranate.readthedocs.io/en/latest/index.html

What to hand in: Show the graph defining the network and the conditional probability tables associated with each node in the graph. Also, explain the dependencies that you have considered while designing your network.

Part B: Inference in Bayesian Network [30 marks]

Implement the variable elimination method and Gibbs sampling to infer in your Bayesian Network. The following queries can be given:

(a) What is the prior probability of a fraudulent transaction (i.e. before we have information that the customer is travelling or whether the purchase is an online purchase or not)?
(b) What is the probability that the transaction is fraudulent when you are being given the information that the customer owns a smartphone?
(c) Finally, you came to know that the customer is abroad. Now, what is the probability of the transaction being fraudulent?

You should be able to demonstrate the steps of the inference process. Also, compare the time and space requirements of the methods.

Exercise 2: The objective of this exercise is to understand the Metropolis-Hastings algorithm, a Markov chain Monte Carlo (MCMC) method for sampling. [**20 marks**]

Consider the following distribution

$$P(x) = \frac{\exp(-x^4)\ (2 + \sin(5x) + \sin(-2x^2))}{\int_{-\infty}^{\infty} \exp(-x'^4)\ \left(2 + \sin(5x') + \sin\left(-2x'^2\right)\right)dx'}$$

Assume that the integration is difficult to solve and you know that

$$P(x) \propto \exp(-x^4)\ (2 + \sin(5x) + \sin(-2x^2))$$

The distribution is shown in Figure 1. Generate samples from this distribution using Metropolis-Hastings algorithm with normal distribution as proposal distribution.

- Generate the candidate using normal distribution with the current state as mean of the distribution i.e. $x^*|x_n \sim Normal(x_n, \sigma^2)$.

- Set $x_0 = -1$, generate 1500 samples for three different values of σ ($low = .05$, $medium = 1$, and $high = 50$) . Plot the histogram of the generated samples and compare with actual distribution for each of the σ values, and also plot the generated sample versus iteration (the actual Markov chain-the sequence of generated values) for each of the σ values.

- Submit your code and a report (hard copy not more than one page –both sides printed) that should have the plots and the conclusions drawn from the plots.

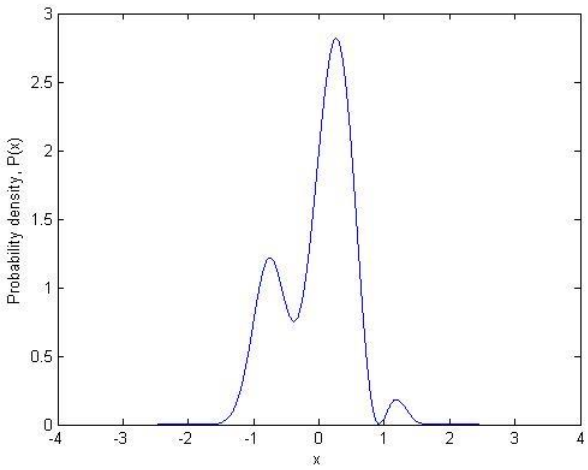

Figure 1. The distribution from which samples are required to be drawn