ELE 364: Assignment #3

1. (10 pts) The table below presents a dataset containing details of policyholders at an insurance company. The descriptive features included in the table describe each policyholder's ID, occupation, gender, age, and type of insurance policy. The preferred contact channel is the target feature in this domain..

ID	Occupation	Gender	Age	Policy type	Pref. channel
1	lab tech	female	43	planC	email
2	farmhand	female	57	planA	phone
3	biophysicist	male	21	planA	email
4	sheriff	female	47	planB	phone
5	painter	male	55	planC	phone
6	manager	male	19	planA	email
7	geologist	male	49	planC	phone
8	messenger	male	51	planB	email
9	nurse	female	18	planC	phone

   (a) Using equal-frequency binning (see Section 3.6.2 in the book), transform the Age feature into a categorical feature with three levels: young, middle-aged, mature.

   (b) Examine the descriptive features in the dataset and list any feature you would exclude before you would use the dataset to build a predictive model. Explain why you made the decision.

   (c) Calculate the probabilities required by a naive Bayes model to represent this domain.

   (d) What target level will a naive Bayes model predict for the following query:
   Gender = female, Age = 30, Policy = planA

2. (10 pts) Imagine that you have been given a dataset of 1,000 douments that have been classified as being about entertainment or education. There are 700 entertainment documents and 300 education documents in the dataset. The tables below give the number of documents from each topic that a selection of words occurred in.

Table 1: Word-document counts for the entertainment dataset

fun	is	machine	christmas	family	learning
415	695	35	0	400	70

Table 2: Word-document counts for the education dataset

fun	is	machine	christmas	family	learning
200	295	120	0	10	105

What target level will a naive Bayes model predict for the query document: *christmas family fun*, if Laplace smoothing with $k = 10$ and vocabulary size of 6 is used?

3. (10 pts) A naive Bayes model is being used to predict whether patients have a high risk of stroke in the next five years (Stroke = true) or a low risk in the next five years (Stroke = false). The model

uses two continuous descriptive features: Age and Weight (in kilograms). Both of these descriptive features are represented by probability density functions, specifically normal distributions. The table below shows the representation of the domain used by the model.

$P(\mathit{Stroke} = \mathit{true}) = 0.25$	$P(\mathit{Stroke} = \mathit{false}) = 0.75$
$P(\mathit{Age} = x \mid \mathit{Stroke} = \mathit{true}) = N(x, \mu = 65, \sigma = 15)$	$P(\mathit{Age} = x \mid \mathit{Stroke} = \mathit{false}) = N(x, \mu = 20, \sigma = 15)$
$P(\mathit{Weight} = x \mid \mathit{Stroke} = \mathit{true}) = N(x, \mu = 88, \sigma = 8)$	$P(\mathit{Age} = x \mid \mathit{Stroke} = \mathit{false}) = N(x, \mu = 76, \sigma = 6)$

What target level will the naive Bayes model predict for the following query:

Age = 45, Weight = 80.

4. (10 pts) The following is a description of the causal relationship between financial crisis, behavior of companies A and B, and the change in our investment portfolio return.

   Financial crisis is a rare event. Companies A and B are two IPO companies that we are currently investing in as part of our investment portfolio. For either company, a bankruptcy would be a rare event. However, in a financial crisis, this likelihood can certainly increase. Our portfolio return depends on the performance of these two companies. The portfolio return typically becomes worse when either of the two companies goes bankrupt, but sometimes can perform better if we can capture the market signals beforehand and short-sell the stocks.

   (a) Define the topology of a three-level Bayesian network that encodes these causal relationships. Treat company A and company B as two separate entities in your model. Then, use the table below to create the conditional probility tables (CPTs) for your Bayesian network. This table captures the historical data for 1910-2019. Company A and company B both recovered from several rounds of bankruptcy due to external investments.

ID	Financial crisis	Company A bankrupt	Company B bankrupt	Portfolio Return Drop
1	F	F	F	F
2	F	F	F	F
3	T	F	T	T
4	F	F	F	T
5	F	T	T	T
6	T	F	T	F
7	F	F	T	F
8	F	T	F	F
9	F	T	F	T
10	T	T	F	T
11	F	F	F	F
12	F	F	F	F
13	T	F	T	F

   (b) What is probability that our portfolio return becomes worse when we find out that both companies went bankrupt, but there is no financial crisis?

   (c) What is probability that our portfolio return becomes worse when we only know that there is a financial crisis?

5. (20 pts) **Coding project**

   For this project you will train naive Bayes models and Bayesian networks to determine whether a patient has diabetes.

   The dataset consists of numerical descriptive features, including the number of pregnancies, glucose level, blood pressure, a skin-based measurement, insulin level, BMI, a pedigree-based measurement and age.

   The target feature is a binary variable, where 1 indicates the presence of diabetes.

   You will fit naive Bayes models and investigate the effect of different hyperparameters, including different feature distributions and strategies for deriving categorical features. You will also fit Bayesian networks and experiment with networks capturing different feature relationships.

   See the Jupyter notebook for more details.