ELE 364: Assignment #3 Solutions

1. (10 pts)

   (a) There are three instances in each bin. Ages 18, 19, 21 go into the young bin, ages 43, 47, 49 into the middle-aged bin, and ages 51, 55, 57 into the mature bin. The threshold between young and middle-aged is the average of 21 and 43, that is 32. Similarly, the threshold between middle-aged and mature is the average of 49 and 51, that is 50. Thus, we map Age = 30 in the query to the young bin.

   (b) As is always the case, the ID feature should not be used as a descriptive feature during training. However, in this example, there is another feature that should be removed from the dataset prior to training. The Occupation feature has different and unique levels for each instance in the dataset. In other words, the Occupation feature is equivalent to the ID for each data instance. Consequently, it should also be removed from the dataset prior to training a model.

   (c) The probabilities for $P(Channel = phone|\mathbf{q})$ are

$$
\begin{aligned}
P(phone) &= 0.56 \\
P(Gender = female|phone) &= 0.60 \\
P(Gender = male|phone) &= 0.40 \\
P(Age = young|phone) &= 0.20 \\
P(Age = middle - aged|phone) &= 0.40 \\
P(Age = mature|phone) &= 0.40 \\
P(Policy = planA|phone) &= 0.20 \\
P(Policy = planB|phone) &= 0.20 \\
P(Policy = planC|phone) &= 0.60
\end{aligned}
$$

   The probabilities for $P(Channel = email|\mathbf{q})$ are

$$
\begin{aligned}
P(email) &= 0.44 \\
P(Gender = female|email) &= 0.25 \\
P(Gender = male|email) &= 0.75 \\
P(Age = young|email) &= 0.50 \\
P(Age = middle - aged|email) &= 0.25 \\
P(Age = mature|email) &= 0.25 \\
P(Policy = planA|email) &= 0.50 \\
P(Policy = planB|email) &= 0.25 \\
P(Policy = planC|email) &= 0.25
\end{aligned}
$$

   (d) The query is converted to Gender = female, Age = young, Policy = planA. Hence, the calculations are:

   $(\prod_{k=1}^{m} P(\mathbf{q}[k]|phone)) \times P(phone) = 0.56 \times 0.6 \times 0.2 \times 0.2 = 0.01344$

   $(\prod_{k=1}^{m} P(\mathbf{q}[k]|email)) \times P(email) = 0.44 \times 0.25 \times 0.5 \times 0.5 = 0.0275$

   Thus, the prediction is email.

2. (10 pts) The table illustrates the smoothing of the posterior probabilities for $P(word|entertainment)$:

Raw Probabilities	$P(christmas\|entertainment)$	$=$	0
	$P(family\|entertainment)$	$=$	0.5714
	$P(fun\|entertainment)$	$=$	0.5929
Smoothing Parameters	k	$=$	10
	$count(entertainment)$	$=$	700
	$count(christmas\|entertainment)$	$=$	0
	$count(family\|entertainment)$	$=$	400
	$count(fun\|entertainment)$	$=$	415
	$\|Domain(vocabulary)\|$	$=$	6
Smoothed Probabilities	$P(christmas\|entertainment) = \frac{0+10}{700+(10\times6)}$	$=$	0.0132
	$P(family\|entertainment) = \frac{400+10}{700+(10\times6)}$	$=$	0.5395
	$P(fun\|entertainment) = \frac{415+10}{700+(10\times6)}$	$=$	0.5592

The table illustrates the smoothing of the posterior probabilities for $P(word|education)$:

Raw Probabilities	$P(christmas\|education)$	$=$	0
	$P(family\|education)$	$=$	0.0333
	$P(fun\|education)$	$=$	0.6667
Smoothing Parameters	k	$=$	10
	$count(education)$	$=$	300
	$count(christmas\|education)$	$=$	0
	$count(family\|education)$	$=$	10
	$count(fun\|education)$	$=$	200
	$\|Domain(vocabulary)\|$	$=$	6
Smoothed Probabilities	$P(christmas\|education) = \frac{0+10}{300+(10\times6)}$	$=$	0.0278
	$P(family\|education) = \frac{10+10}{300+(10\times6)}$	$=$	0.0556
	$P(fun\|education) = \frac{200+10}{300+(10\times6)}$	$=$	0.5833

We can now compute the probabilities (unnormalized) of each target level:

$P(entertainment|q) = P(entertainment) \times P(christmas|entertainment) \times P(family|entertainment) \times P(fun|entertainment) = 0.7 \times 0.0132 \times 0.5395 \times 0.5592 = .0028.$

$P(education|q) = P(education) \times P(christmas|education) \times P(family|education) \times P(fun|education) = 0.3 \times 0.0278 \times 0.0556 \times 0.5833 = .0003.$

Hence, the model will predict entertainment for this query.

3. (10 pts) First, we need the following probability density function values:

$P(Age = 45|Stroke = true) = N(45, 65, 15) = 0.0109$

$P(Weight = 80|Stroke = true) = N(80, 88, 8) = 0.0302$

$P(Age = 45|Stroke = false) = N(45, 20, 15) = 0.0066$

$P(Weight = 80|Stroke = false) = N(80, 76, 6) = 0.0532$

Hence, $P(Stroke = true|q) = P(Stroke = true) \times P(Age = 45|Stroke = true) \times P(Weight = 80|Stroke = true) = 0.25 \times 0.0109 \times 0.0302 = 0.0000822$.

$P(Stroke = false|q) = P(Stroke = false) \times P(Age = 45|Stroke = false) \times P(Weight = 80|Stroke = false) = 0.75 \times 0.0066 \times 0.0532 = 0.0002633$. Hence, prediction is Stroke = false.

4. (10 pts)

(a) The Bayesian network is as follows.

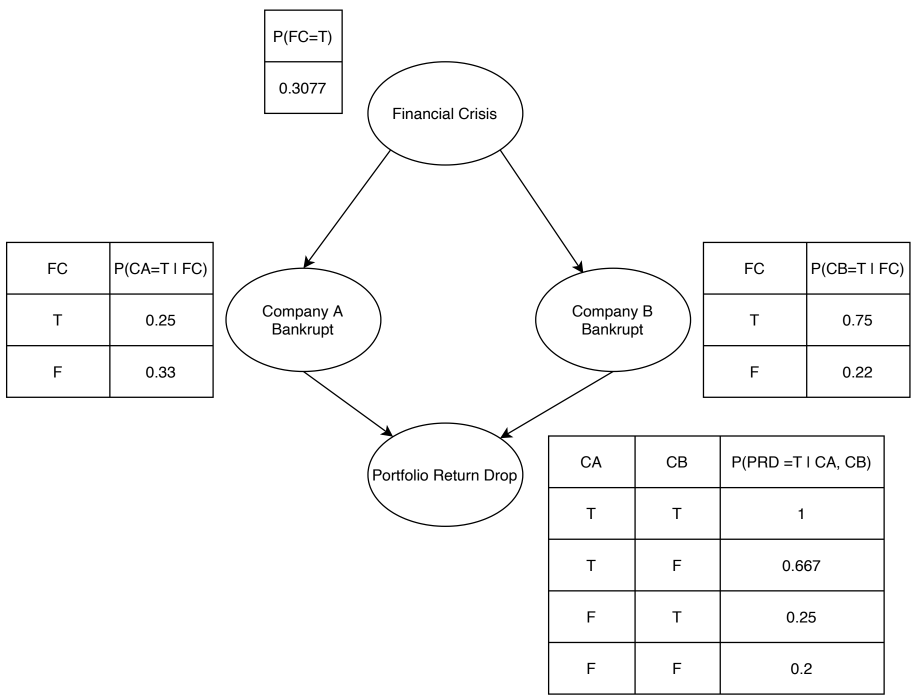

Figure 1: Bayesian network for Q4.

(b) As can be seen from the Bayesian network, when both companies go bankrupt, the predicted probability of a portfolio return drop is 100%.

(c) Only the root node is known. In this case, the prediction is based on:
$P(prd|fc) = \frac{P(prd,fc)}{P(fc)} = \frac{\sum_{i,j} P(prd,CA_i,CB_j,fc)}{P(fc)}$
where we have
$P(prd, fc) = \sum_{i,j} P(prd, CA_i, CB_j, fc)$
$= P(prd|ca, cb) \times P(ca|fc) \times P(cb|fc) \times P(fc)$
$+P(prd|ca', cb) \times P(ca'|fc) \times P(cb|fc) \times P(fc)$
$+P(prd|ca, cb') \times P(ca|fc) \times P(cb'|fc) \times P(fc)$
$+P(prd|ca', cb') \times P(ca'|fc) \times P(cb'|fc) \times P(fc)$
$= 0.1253$

and $P(fc) = 0.3077$

Thus, the final result is 0.1253/0.3077=0.4072, which indicates a 40.72% probability that our portfolio return becomes worse.