ST301 / ST413 - Bayesian Statistics & Decision Theory
2010/11

Murray Pollock & Prof. Jim Smith
May 14, 2011

1 Exercise Sheet 1

1.1 Question 1

A doctor needs to diagnose which disease a patient has \((I = \{1, 2, 3\})\). She knows the patient will have one (and only one) of these illnesses. She also knows that of the types of patients she will be asked to diagnose 80% will have illness 1, 10% will have illness 2 and 10% will have illness 3.

To help her in her diagnosis she will observe the existence or otherwise (1 or 0) of four symptoms \((X_1, X_2, X_3, X_4)\) representing (respectively), high pulse rate, high blood pressure, low blood sugar and chest pains.

The probabilities of exhibiting various symptoms \((X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1)\) given various diseases are shown in Table 1:-

<table>
<thead>
<tr>
<th>Symptom - (X_j)</th>
<th>Disease - (I = i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(X_1 = 1</td>
<td>I = i))</td>
</tr>
<tr>
<td>(P(X_2 = 1</td>
<td>I = i))</td>
</tr>
<tr>
<td>(P(X_3 = 1</td>
<td>I = i))</td>
</tr>
<tr>
<td>(P(X_4 = 1</td>
<td>I = i))</td>
</tr>
</tbody>
</table>

Table 1: Conditional probabilities of symptoms given disease.

You may assume the naive Bayes hypothesis\(^1\) throughout.

1.1.1 Question 1 i)

Given the doctor observes a patients’ symptoms \((X_1, X_2, X_3, X_4)\) to be \((1, 1, 0, 1)\) calculate her posterior probabilities that the patient has illness 1, 2 and 3.

Using log-odds\(^2\) and noting from the question that the patient has precisely one disease (with \(P(I = 1) = 0.8, P(I = 2) = 0.1 & P(I = 3) = 0.1\)) and furthermore noting that the patient either does or does not exhibit a particular symptom (ie \(- P(X_i = 1 | I = j) = 1 - P(X_i = 0 | I = j) \forall i, j\) then it is sufficient calculate for illnesses \(I = \{2, 3\}\) the posterior log-odds ratio with respect to illness \(I = 1\).

---

\(^1\)Naive Bayes hypothesis (see [1] §1.2.2 Def\(^a\) 1.8): Assume that all symptoms are independent given the illness class \(I = i\) for all possible illness classes \(i, 1 \leq i \leq n\).

\(^2\)Log-odds (see [1] §1.2.3): Denoting \(I\) as an illness and \(Y\) as a vector of symptoms then given \(P(Y | I = k)Y = y) > 0\) we have,

\[
P(I = i | Y = y) = \left(\frac{P(Y = y | I = i) \cdot P(I = i)}{P(Y = y)}\right) = \left(\frac{\prod_{j=1}^{n} P(Y_j = y | I = i) \cdot P(I = i)}{P(I = k)}\right)\]

(1)
In particular for $I = 2$ we have,

\[
\sum_{j=1}^{4} A_j(2, 1, x_j) = \log \frac{\mathbb{P}(X_1 = 1|I = 2)}{\mathbb{P}(X_1 = 1|I = 1)} + \log \frac{\mathbb{P}(X_2 = 1|I = 2)}{\mathbb{P}(X_2 = 1|I = 1)} + \log \frac{\mathbb{P}(X_3 = 0|I = 2)}{\mathbb{P}(X_3 = 0|I = 1)} + \log \frac{\mathbb{P}(X_4 = 1|I = 2)}{\mathbb{P}(X_4 = 1|I = 1)}
\]

\[
= \log \frac{0.2}{0.3} + \log \frac{0.9}{0.1} + \log \frac{1 - 0.1}{1 - 0.8} + \log \frac{0.9}{0.2} = 4.8 \text{ to 1 decimal place.} \tag{3}
\]

\[
O(2, 1) = \log \frac{\mathbb{P}(I = 2)}{\mathbb{P}(I = 1)}
\]

\[
= \log \frac{0.1}{0.8} = -2.1 \text{ to 1 decimal place.} \tag{4}
\]

\[
O(2, 1|x) = \sum_{j=1}^{4} A_j(2, 1, x_j) + O(2, 1)
\]

\[
= 4.8 - 2.1 = 2.7 \text{ to 1 decimal place.} \tag{5}
\]

Similarly we have for $I = 3 \sum_{j=1}^{4} A_j(3, 1, x_j) = 4.3$ (to 1 dp) and $O(3, 1) = -2.1$ (to 1 dp) implying $O(3, 1|x) = 2.2$ (to 1 dp).

Now noting that,

\[
\sum_{i=1}^{3} \mathbb{P}(I = i|X = x) = \mathbb{P}(I = 1|X = x) + e^{2.7} \cdot \mathbb{P}(I = 1|X = x) + e^{2.2} \cdot \mathbb{P}(I = 1|X = x) = 1 \tag{6}
\]

we have,

\[
\mathbb{P}(I = 1|X = x) = \frac{1}{1 + e^{2.7} + e^{2.2}} = 0.04 \text{ to 2 decimal places.} \tag{7}
\]

\[
\mathbb{P}(I = 2|X = x) = e^{2.7} \cdot \mathbb{P}(I = 1|X = x) = 0.60 \text{ to 2 decimal places.} \tag{8}
\]

\[
\mathbb{P}(I = 3|X = x) = e^{2.2} \cdot \mathbb{P}(I = 1|X = x) = 0.36 \text{ to 2 decimal places.} \tag{9}
\]

### 1.1.2 Question 1 ii)

Let $R(D, I)$ denote the reduction in the patients’ life expectancy when he has disease $I$ and is first given treatment $D$. The values of $R(D, I)$ are given in Table 2 (note here that $D(i)$ is the best treatment for $I = i$ where $i = \{1, 2, 3\}$). What action minimizes the doctors expectation of $R$?

Using the law of total probability and noting that knowledge of the symptoms is irrelevant given knowledge

Now taking logs we have,

\[
O(i, k|x) := \log \mathbb{P}(I = i|Y = y) = \sum_{j=1}^{m} \log \left( \frac{\mathbb{P}(Y_j = y|I = i)}{\mathbb{P}(Y_j = y|I = k)} \right) + \log \frac{\mathbb{P}(I = i)}{\mathbb{P}(I = k)} =: \sum_{j=1}^{m} A_j(i, k, y_j) + O(i, k) \tag{2}
\]
of the illness then we have,

\[ R(D(k)|X = x) = \sum_{i=1}^{3} R(D(k), I = i|X = x) \]

\[ = \sum_{i=1}^{3} R(D(k)|I = i, X = x) \cdot P(I = i|X = x) \]

\[ = \sum_{i=1}^{3} R(D(k)|I = i) \cdot P(I = i|X = x) \]

which using the results from Question 1 i), and considering \( I = 1, 2, 3 \) we find treatment \( D(3) \) is the best in terms of lowest expected reduction in life expectancy (for some intuition consider the importance of prescribing \( D(3) \) in the case where \( I = 3 \)),

\[ R(D(1)|X = \{1, 1, 0, 1\}) = 1 \cdot 0.04 + 9 \cdot 0.6 + 9 \cdot 0.36 = 8.68 \]  \hspace{1cm} (10)

\[ R(D(2)|X = \{1, 1, 0, 1\}) = 1 \cdot 0.04 + 8 \cdot 0.6 + 7 \cdot 0.36 = 7.36 \]  \hspace{1cm} (11)

\[ R(D(3)|X = \{1, 1, 0, 1\}) = 2 \cdot 0.04 + 9 \cdot 0.6 + 3 \cdot 0.36 = 6.56 \]  \hspace{1cm} (12)

1.1.3 Question 1 iii)

Does the doctors decision differ if the patients symptoms measurement \( X_3 \) was omitted? What difference would such an omission cause and why?

Noting the form of the log-likelihood ratio in (2) and the results of Question 1 i), the posterior probabilities can be easily modified. In particular denoting \( \hat{O}(j, i|x) \) as the modified log-likelihood ratio (omitting \( X_3 \)) we have,

\[ \hat{O}(2, 1|x) = O(2, 1|x) - \lambda_3(2, 1|x_3) = 2.7 - \log \frac{1 - 0.1}{1 - 0.8} = 1.2 \text{ to 1 decimal place.} \]  \hspace{1cm} (13)

\[ \hat{O}(3, 1|x) = O(3, 1|x) - \lambda_3(3, 1|x_3) = 2.2 - \log \frac{1 - 0.1}{1 - 0.8} = 0.7 \text{ to 1 decimal place.} \]  \hspace{1cm} (14)

As before note that,

\[ \sum_{i=1}^{3} \mathbb{P}(I = i|X = x) = \mathbb{P}(I = 1|X = x) + e^{1.2} \cdot \mathbb{P}(I = 1|X = x) + e^{0.7} \cdot \mathbb{P}(I = 1|X = x) = 1 \]  \hspace{1cm} (15)

we have,

\[ \mathbb{P}(I = 1|X = x) = \frac{1}{1 + e^{1.2} + e^{0.7}} = 0.16 \text{ to 2 decimal places.} \]  \hspace{1cm} (16)

\[ \mathbb{P}(I = 2|X = x) = e^{1.2} \cdot \mathbb{P}(I = 1|X = x) = 0.52 \text{ to 2 decimal places.} \]  \hspace{1cm} (17)

\[ \mathbb{P}(I = 3|X = x) = e^{2.2} \cdot \mathbb{P}(I = 1|X = x) = 0.32 \text{ to 2 decimal places.} \]  \hspace{1cm} (18)
Although treatment $D(3)$ is still the preferred treatment note that the cost of mistreatment in terms of life expectancy is now lower. In particular note that a priori it is assumed the patient has disease $I = 1$ which is less threatening than the other two diseases, however the symptom $X_3 = 0$ is inconsistent with this disease and as such moves the assessment away from this prior (the addition of more information has provided a more informative posterior).

2 Exercise Sheet 2

2.1 Question 1

You work for a catalogue sales agency. They have recently had designed a children’s jogging suit which they intend to market as sportswear and as such expect to make a profit of £30000 per month over the next year. However the promotion team believe with probability 0.2 it could “take-off” (X = 1) as day-wear as well, if (and only if) it is marketed as such. If X = 1 and it is marketed as day-wear then the company expect to receive an extra £100000 per month profit. However if it does not “take-off” (X = 0) and they market it as day-wear they expect to lose £50000 per month in profits wasted by ineffective use of advertising space in the day-wear section of the catalogue to subtract from their sportswear profits.

The company’s options are:-
- $d_1$ - market the suit as sportswear only for 12 months.
- $d_2$ - market the suit as both sportswear and day-wear for 12 months.
- $d_3$ - survey the market quickly at the cost of £20000 and on the basis of this choose $d_1$ or $d_2$. This result is not however completely reliable giving a positive indication ($Y = 1$) given $X = 1$ with $P(Y = 1|X = 1) = 0.8$ and a negative result with $P(Y = 0|X = 0) = 0.6$.

Draw the decision tree of this problem and determine your optimal decision rule when you are trying to maximise your expected profit over the year. What is this expected maximum profit?

From the information contained in the question the decision tree can be constructed as per Figure 1.

The unknown probabilities from Figure 1 can now be explicitly found. In particular,

\[
p_1 := P(Y = 1) = P(Y = 1|X = 0) \cdot P(X = 0) + P(Y = 1|X = 1) \cdot P(X = 1) \]
\[
= (1 - 0.6) \cdot (1 - 0.2) + 0.8 \cdot 0.2 = 0.48 = 1 - p_2
\]
\[
p_3 := P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1) \cdot P(X = 1)}{P(Y = 1)} = \frac{0.8 \cdot 0.2}{0.48} = \frac{1}{3} =: 1 - p_4
\]
\[
p_5 := P(X = 1|Y = 0) = \frac{P(Y = 0|X = 1) \cdot P(X = 1)}{P(Y = 0)} = \frac{(1 - 0.8) \cdot 0.2}{0.52} = \frac{1}{13} =: 1 - p_6
\]

Using the probabilities calculated and making decisions based on maximizing expected pay-off then the optimal decision can be found by backwards evaluation of the decision tree. In particular we have,

\[
E_a = £1560000 \cdot 0.2 - £240000 \cdot 0.8 = £120000
\]
\[
E_b = \max\{£340000, E_c \cdot p_1 + \max\{£340000, E_d \cdot p_2 \}
\]
\[
= \max\left\{£340000, £1540000 \cdot \frac{1}{3} - £260000 \cdot \frac{2}{3} \right\} \cdot 0.48 + \max\left\{£340000, £1540000 \cdot \frac{1}{13} - £260000 \cdot \frac{12}{13} \right\} \cdot 0.52
\]
\[
= £340000
\]
Figure 1: Ex2 Q1 decision tree
As such we find the expected pay-offs for \(d_1, d_2 \& d_3\) to be £360000, £120000 \& £340000 respectively and as such \(d_1\) is the best option.

### 2.2 Question 2

Define the expected value of perfect information (EVPI). How can it be used to simplify a decision tree?

EVPI is the expected pay-off given perfect information about the current uncertainties minus the expected pay-off with current uncertainties,

\[
EVPI = \mathbb{E}_d \left( \max_d R(d, \theta) \right) - \max_d \mathbb{E}_d \left( R(d, \theta) \right)
\]  

(27)

EVPI is useful in decision trees as it provides an upper bound on the price worth paying for any experimentation, limiting the extent of the decision tree.

Your customer insists you give him a guarantee that a piece of machinery will not be faulty for one year. As the salesman you have the option of overhauling your machinery before delivering it to the customer (action \(a_2\)) or not (action \(a_1\)). The pay-offs you will receive by taking these actions when the machine is, or is not, faulty are given in Table 3.

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>(a_1)</th>
<th>(a_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Faulty</td>
<td></td>
<td>1000</td>
<td>800</td>
</tr>
<tr>
<td>Faulty</td>
<td></td>
<td>0</td>
<td>700</td>
</tr>
</tbody>
</table>

Table 3: Machinery guarantee pay-offs (£’s)

The machinery will work if a certain plate of metal is flat enough. You have scanning devices which will sound an alarm if a small region of this plate is bumpy. The probabilities that any one of these scanning devices sound an alarm given that the machine is, or is not, faulty, are respectively 0.9 and 0.4.

You may decide to scan with scanning devices \(d_n\) \((n = 0, 1, \ldots)\) at the cost of £50 \cdot n and overhaul the machine or not depending on the number of alarms that ring. These devices will give independent readings conditional on whether the machine is faulty or not. Prior to scanning you believe that the probability that your machine is faulty is 0.2.

Draw a decision tree of the necessary decision sequences to find the optimal number of scanning devices to use and how to use them to maximize your expected pay-off.

Letting \(F\) denote a fault (and \(F^c\) no fault), \(U\) denote some pay-off function and noting from the question that \(\mathbb{P}(F) = 0.2\) then applying (27) we have,

\[
EVPI = (\mathbb{P}(F^c)U(a_1, F^c) + \mathbb{P}(F)U(a_2, F)) - \max \{ [\mathbb{P}(F)U(a_1, F) + \mathbb{P}(F^c)U(a_1, F^c)), (\mathbb{P}(F)U(a_2, F) + \mathbb{P}(F^c)U(a_2, F^c))] \}
\]

\[
= ((1 - 0.2) \cdot 1000 + 0.2 \cdot 700) - \max \{ [0.2 \cdot 0 + (1 - 0.2) \cdot 1000), (0.2 \cdot 700 + (1 - 0.2) \cdot 800)] \}
\]

\[= 940 - 800 = 140 (28)\]

As a result it isn’t worth purchasing more than \([140/50] = 2\) alarms hence limiting the size of the decision tree that has to be considered as shown in Figure 2.

In particular considering the three cases of device number \((d_n, d_1, d_2)\) then we find the conditional probabilities from the decision tree to be as follows (noting the number of rings is bounded by 0 and the number of devices) using the notation whereby \(A\) denotes an alarm activation (and \(A^c\) denotes no alarm activation) we have,
- In the case of \( d_0 \) we have,

\[
\mathbb{P}_{F|0-0} = \mathbb{P}(F) = 0.2 \quad (29)
\]
\[
\mathbb{P}_{F'|0-0} = \mathbb{P}(F') = 0.8 \quad (30)
\]
\[
E_0 = \mathbb{P}(b = 0|a = 0) \cdot D_{0-0}
\]
\[
= 1 \cdot \max \{0.2 \cdot 0 + 0.8 \cdot 1000, 0.2 \cdot 700 + 0.8 \cdot 800 \} = 800 \quad (31)
\]

- In the case of \( d_1 \) we have either 1 or 0 rings. In particular,

\[
\mathbb{P}_{F|1-0} = \mathbb{P}(F|A') = \frac{\mathbb{P}(A'|F) \cdot \mathbb{P}(F)}{\mathbb{P}(A')} = \frac{0.1 \cdot 0.2}{0.1 \cdot 0.2 + 0.6 \cdot 0.8} = 0.04 \quad (32)
\]
\[
\mathbb{P}_{F'|1-0} = \mathbb{P}(F'|A') = 0.96 \quad (33)
\]
\[
\mathbb{P}_{F|1-1} = \mathbb{P}(F|A) = 0.36 \quad (34)
\]
\[
\mathbb{P}_{F'|1-1} = \mathbb{P}(F'|A) = 0.64 \quad (35)
\]
\[
E_1 = \mathbb{P}(b = 0|a = 1) \cdot D_{1-0} + \mathbb{P}(b = 1|a = 1) \cdot D_{1-1}
\]
\[
= (0.2 \cdot 0.1 + 0.8 \cdot 0.6) \cdot D_{1-0} + (0.2 \cdot 0.9 + 0.8 \cdot 0.4) \cdot D_{1-1}
\]
\[
= 0.5 \cdot \max \{0.04 \cdot (0 - 50) + 0.96 \cdot (1000 - 50), 0.04 \cdot (700 - 50) + 0.96 \cdot (800 - 50)\}
\]
\[
+ 0.5 \cdot \max \{0.36 \cdot (0 - 50) + 0.64 \cdot (1000 - 50), 0.36 \cdot (700 - 50) + 0.64 \cdot (800 - 50)\}
\]
\[
= 0.5 \cdot \max(910, 746) + 0.5 \cdot \max(590, 714) = 812 \quad (36)
\]
- In the case of \( d_5 \) we have either 2, 1 or 0 rings. In particular,
\[
\mathbb{P}_{F|\bar{F}} = \frac{\mathbb{P}(F|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(F)} = \frac{0.1 \cdot 0.1 \cdot 0.2}{0.1 \cdot 0.1 \cdot 2 + 0.6 \cdot 0.6 \cdot 0.8} = 0.0069
\] (37)
\[
\mathbb{P}_{F|\bar{F}} = \frac{\mathbb{P}(F|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(F)} = 0.9931
\] (38)
\[
\mathbb{P}_{F|\bar{F}} = \frac{\mathbb{P}(F|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(F)} = 0.086
\] (39)
\[
\mathbb{P}_{F|\bar{F}} = \frac{\mathbb{P}(F|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(F)} = 0.914
\] (40)
\[
\mathbb{P}_{F|\bar{F}} = \frac{\mathbb{P}(F|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(F)} = 0.56
\] (41)
\[
\mathbb{P}_{F|\bar{F}} = \frac{\mathbb{P}(F|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(F)} = 0.44
\] (42)
\[
\mathbb{P}(AA) = 0.2 \cdot 0.9 \cdot 0.9 + 0.8 \cdot 0.4 \cdot 0.4 = 0.29
\] (43)
\[
\mathbb{P}(A\bar{A}) = 0.42
\] (44)
\[
\mathbb{P}(A\bar{A}) = 0.29
\] (45)
\[
D_2 = \max \{0.0069 \cdot (0 - 100) + 0.9931 \cdot (1000 - 100), 0.0069 \cdot (700 - 100) + 0.9931 \cdot (800 - 100)\}
= \max(893.1, 699.31) = 893.1
\] (46)
\[
D_2 = \max \{0.0086 \cdot (0 - 100) + 0.914 \cdot (1000 - 100), 0.0086 \cdot (700 - 100) + 0.914 \cdot (800 - 100)\}
= \max(814, 645) = 815
\] (47)
\[
D_2 = \max \{0.56 \cdot (0 - 100) + 0.44 \cdot (1000 - 100), 0.56 \cdot (700 - 100) + 0.44 \cdot (800 - 100)\}
= \max(340, 644) = 644
\] (48)
\[
E_2 = \mathbb{P}(b = 0|a = 2) \cdot D_2 + \mathbb{P}(b = 1|a = 2) \cdot D_2 + \mathbb{P}(b = 2|a = 2) \cdot D_2
= 0.29 \cdot 893.1 + 0.42 \cdot 815 + 0.29 \cdot 644 = 788
\] (49)
As a result the optimal choice of the number of scans is the choice which maximizes \( E_0, E_1, E_2 \) which with reference to (31), (36) and (49) we find to be £812 using a single scan (choice \( d_1 \)).

3 Exercise Sheet 3

3.1 Question 1

Given a minimum possible reward \( s \) and a maximum possible reward \( t \) describe how you might construct your utility \( U(r) \) for a reward \( s \leq r \leq t \). Given any two decisions \( d_1 \) and \( d_2 \) giving reward probability mass functions \( p_1(r) \) and \( p_2(r) \) respectively and your utility function \( U(r) \) how do you decide whether to take decision \( d_1 \) or \( d_2 \)?

Prefer \( d_1 \) to \( d_2 \) if,
\[
\tilde{U}(d_1) := \sum_r U(r) \cdot p_1(r) \geq \sum_r U(r) \cdot p_2(r) =: \tilde{U}(d_2)
\] (50)

Your utility function \( U(r) \) is given by,
\[
U(r) = 1 - \exp(-cr)
\] (51)

You must decide between two decisions. Decision \( d_1 \) will give you reward £1 with probability \( 3/4 \) and £0 with probability \( 1/4 \). Decision \( d_2 \) will give you a reward of £\( r \) with probability \( (1/2)^{r+1} \) (where \( r = 0, 1, \ldots \)). Prove that you will prefer \( d_1 \) to \( d_2 \) if and only if the value of \( c \) in your utility function satisfies,
\[
c \geq \log 3 - \log 2
\] (52)

Re-arranging (50) we have \( d_1 > d_2 \) if,
\[
1 - \sum_r U(r) \cdot p_1(r) \leq 1 - \sum_r U(r) \cdot p_2(r)
\]
\[
1 - \frac{1 - e^0}{4} - \frac{3(1 - e^{-c})}{4} \leq \sum_{r=2}^\infty \frac{e^{-cr}}{2^r}
\] (53)
Letting \( x := e^{-c} \) then \( d_1 > d_2 \) if,

\[
4(2 - x)^{-1} \geq 1 + 3x \\
(3x - 2)(x - 1) \geq 0
\]

(54)

As a result \( e^{-c} \leq 2/3 \) (we are given \( c > 0 \) so \( e^{-c} \geq 1 \) isn’t possible), so re-arranging we have,

\[
c \geq \log 3 - \log 2
\]

(55)

3.2 Question 2

Three investors \((I_1, I_2, I_3)\) each wish to invest $10000 in shareholdings. Four shares \((S_1, S_2, S_3, S_4)\) are available to each person. \( S_1 \) will guarantee the shareholder exactly 8% interest over the next year. Share \( S_2 \) will pay nothing with probability 0.1, 8% interest with probability 0.5 and 16% interest with probability 0.4. Share \( S_3 \) will pay 4% interest with probability 0.2 and 12% with probability 0.8. Share \( S_4 \) will pay 16% interest with probability 0.8 and nothing with probability 0.2 over the next year.

Each investor must rank five portfolios \((P_1, P_2, P_3, P_4, P_5)\) in order of preference. Portfolio \( P_i \) (for \( 1 \leq i \leq 4 \)) represents the investment of $10000 in share \( S_i \). Portfolio \( P_5 \) represents the investment of $5000 in share \( S_1 \) and $5000 in share \( S_4 \).

Investor \( I_1 \) prefers \( P_1 \) and \( P_3 \) to \( P_5 \). Investor \( I_2 \) prefers \( P_1 \) to \( P_3 \) and \( P_3 \) to \( P_4 \). Investor \( I_3 \) prefers \( P_1 \) to \( P_4 \) and \( P_2 \) to \( P_1 \).

Find the pay-off distribution over the next year associated with each portfolio. Hence, or otherwise, identify those investors whose preferences given above violate Axioms 8–11\(^3\) and explain why. You may assume that each investor’s utility is an increasing function of his pay-off in the next year.

\(^3\)As given in the course notes we have,

- **Axiom 8**: Bets giving the same distribution of rewards should be considered equivalent.
- **Axiom 9**: If \( P_1, P_2 \) and \( P \) are any three reward distributions, then for all \( \alpha \) such that \( 0 < \alpha < 1 \) then,
  \[
P_1 \preceq P_2 \iff \alpha P_1 + (1 - \alpha)P \preceq \alpha P_2 + (1 - \alpha)P
  \]
  (56)
- **Axiom 10**: If \( P_1, P_2 \) and \( P \) are any three reward distributions such that \( P_1 < P < P_2 \) then there exist \((\alpha, \beta)\) where \( 0 < \alpha, \beta < 1 \) such that,
  \[
P \prec \alpha P_2 + (1 - \alpha)P_1
  \]
  \[
P \succ \beta P_2 + (1 - \beta)P_1
  \]
  (57)
  (58)
- **Axiom 11**: Let \( P(r) \) be any distribution on the reward random vector \( R \) for which \( P(r_1 < R < r_2) \) for some rewards \( r_1, r_2 \) and \( p(r) \) be its probability mass function. For each \( r \) let \( a(r) := U(r) - U(r_1) \), then \( P \prec \beta r_2 + (1 - \beta)r_1 \) where,
  \[
  \beta = \sum_{r \in R} a(r) \cdot p(r)
  \]
  (59)
From the question we find the distribution of the portfolios to be as follows,

\[
P_1 \sim \begin{cases} 
    \$800 & \text{w.p. 1} \end{cases} 
\]

\[
P_2 \sim \begin{cases} 
    \$1600 & \text{w.p. 0.4} \\
    \$800 & \text{w.p. 0.5} \\
    0 & \text{w.p. 0.1} 
\end{cases} 
\]

\[
P_3 \sim \begin{cases} 
    \$1200 & \text{w.p. 0.8} \\
    400 & \text{w.p. 0.2} 
\end{cases} 
\]

\[
P_4 \sim \begin{cases} 
    \$1600 & \text{w.p. 0.8} \\
    0 & \text{w.p. 0.2} 
\end{cases} 
\]

\[
P_5 \sim \begin{cases} 
    \$400 + \$800 = \$1200 & \text{w.p. 0.8} \\
    \$400 + 0 = \$400 & \text{w.p. 0.2} 
\end{cases} 
\]

Now considering the investors we find as follows,
- \( I_1 (P_1 \succ P_3 \& P_1 \succ P_5) \) violates Axiom 8 (see footnote 3) as \( P_3 \succ P_5 \) yet they have the same pay-off.
- \( I_2 (P_1 \succ P_3 \& P_3 \succ P_4) \) doesn’t violate Axioms 8-11 (see footnote 3) as, for instance, considering the increasing utility function such that \( U(0) = 0, U(400) = 1, U(800) = 10, U(1200) = 11, U(1600) = 11.1 \) then,
  \[
  U(P_1) = 10 > U(P_3) = 11 \cdot 0.8 + 1 \cdot 0.2 = 9 > U(P_4) = 11.1 \cdot 0.8 + 0 \cdot 0.2 = 8.88 
  \]
- \( I_3 (P_1 \succ P_4 \& P_2 \succ P_1) \) violates Axiom 9 (see footnote 3) as \( P_2 = 0.5 \cdot P_1 + 0.5 \cdot P_4 > P_1 \), however \( P_1 = 0.5 \cdot P_1 + 0.5 \cdot P_1 \) and yet \( P_1 \succ P_4 \).

### 3.3 Question 3

Table 4 gives the pay-offs in $ associated with six possible decisions \( (d_1, d_2, \ldots, d_6) \) where the outcome \( \theta \) can take one of two possible values (\( \theta_1 \) or \( \theta_2 \)).

<table>
<thead>
<tr>
<th>Decision</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( d_2 )</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>( d_3 )</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>( d_4 )</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>( d_5 )</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>( d_6 )</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4: Decision pay-offs (\$’s)

#### 3.3.1 Question 3 i)

Find those decisions which can be Bayes decisions under a linear utility function for some values of the probability \( p(\theta_1) \).

Specify the corresponding possible values of \( p(\theta_1) \) for each of these decisions.

Plotting the decision pay-offs (Table 3) and considering linear utility functions of the form \( G(d, \theta_1) = m G(d, \theta_2) + b \) we find that only \( d_2 \) and \( d_6 \) can possibly maximise expected utility. In particular the red line in Table 3 connecting \( d_2 \) and \( d_6 \) has a gradient \( m = -3/2 \).

In general some decision \( a \) will be chosen in preference to \( b \) if the expected utility of \( a \) is greater than \( b \). Further-
more in the case of a linear utility function with two outcomes we have that \( a \) is preferred to \( b \) if,

\[
\mathbb{E}(R(a)) \geq \mathbb{E}(R(b)) \\
R(a, \theta_1) \cdot \mathbb{P}(\theta_1) + R(a, \theta_2) \cdot \mathbb{P}(1 - \theta_1) \geq R(b, \theta_1) \cdot \mathbb{P}(\theta_1) + R(b, \theta_2) \cdot \mathbb{P}(1 - \theta_1) \\
(R(a, \theta_1) - R(b, \theta_1)) \cdot \mathbb{P}(\theta_1) \geq -(R(a, \theta_2) - R(b, \theta_2)) \cdot \mathbb{P}(1 - \theta_1)
\]

\[
\frac{\mathbb{P}(\theta_1)}{1 - \mathbb{P}(\theta_1)} \geq -\frac{(R(a, \theta_2) - R(b, \theta_2))}{(R(a, \theta_1) - R(b, \theta_1))} = \frac{1}{m}
\]

For our example we have that \( d_2 \) will be preferred to \( d_6 \) whenever,

\[
\mathbb{P}(\theta_1) \geq -\frac{1}{3/2} = \frac{2}{3} \\
\mathbb{P}(\theta_1) \geq 2/5
\]

Figure 3: Ex3 Q3 plot of decision pay-offs

### 3.3.2 Question 3 ii)

Explain carefully which of the decisions \((d_1, d_2, \ldots, d_6)\) can be Bayes decisions under some utility function (assumed strictly increasing in pay-off) and probability value \(0 < p(\theta) < 1\).

A Bayes’ decision \( d^* \) is one that maximizes expected utility (\( \max_d \mathbb{E}(U(d)) \)). Decisions can only be Bayes decisions if there does not exist any other decision which can with certainty provide a pay-off at least as great as the chosen decision for all outcomes and greater than the chosen decision for at least one outcome. More formally for some decision \( d^* \) if \( d \neq d^* \) such that \( G(d^*, \theta_k) \geq G(d_i, \theta_k) \) \( \forall k \) and \( G(d_i, \theta_k) > G(d^*, \theta_k) \) for some \( k \).

Possible Bayes decisions can be found by comparing each pair of possible decisions. More intuitively, with reference to Table 3 decisions can be eliminated by considering each decision and removing any decisions falling in the lower left hand corner of that decision. The possible Bayes decisions are those left after this elimination procedure. For instance, in this case we have \( d_2 \) (which eliminates \( d_1 \) and \( d_4 \)), \( d_5 \) (which eliminates \( d_3 \)) and \( d_6 \) as our possible Bayes decisions.

The Bayes decision can only be found once the utility function is known, but will be one of the possible Bayes decision as above. As in Question 3 ii) knowledge about the form of the utility function may further reduce the number of possible Bayes decisions.
4 Exercise Sheet 4

4.1 Question 1

You are a member of a jury and you want to be rational (you want to maximize your expected utility). There are essentially 4 combinations of outcomes and decisions on which your utility can depend.

- \( u(0, 0) \) - Find the suspect is innocent and they are actually innocent.
- \( u(0, 1) \) - Find the suspect is innocent and they are actually guilty.
- \( u(1, 0) \) - Find the suspect is guilty and they are actually innocent.
- \( u(1, 1) \) - Find the suspect is guilty and they are actually guilty.

Assume that

\[ u(0, 1) \leq u(0, 0) < u(0, 0) = u(1, 1) \] (68)

Write down your expected utility as a function of your posterior probability of guilt and the utilities given above. If your utilities satisfy this what does this mean?

Let \( d_0 \) and \( d_1 \) be respectively the decision to find the suspect innocent and guilty. Furthermore set the boundary utilities as follows - \( u(0, 1) = 0 \) and \( u(0, 0) = u(1, 1) = 1 \) - and further let \( u(1, 0) = u \in [0, 1] \).

For shorthand let \( G \) and \( G^c \) be guilty and innocent respectively, \( E \) be evidence given in the trial and denote \( p = \mathbb{P}(G) \) and \( p^* = \mathbb{P}(G | E) \).

Now we find the expected utilities associated with our decisions \( (d_0, d_1) \) to be,

\[
\mathbb{E}(U(d_0)) = \mathbb{P}(G | E) \cdot u(0, 1) + \mathbb{P}(G^c | E) \cdot u(0, 0) = 0 + (1 - p^*) \\
\mathbb{E}(U(d_1)) = \mathbb{P}(G | E) \cdot u(1, 1) + \mathbb{P}(G^c | E) \cdot u(1, 0) = p^* + (1 - p^*) \cdot u
\] (69) (70)

Assuming rationality then the suspect should be found guilty iff \( \mathbb{E}(U(d_1)) > \mathbb{E}(U(d_0)) \),

\[ p^* + (1 - p^*) \cdot u > 1 - p^* \]

\[ \frac{p^*}{1 - p^*} > 1 - u \] (71)

Prove that in this case you should decide to say guilty only if the following equality holds for some \( A \) (some function of your prior odds of innocence and the utility values) and evaluate \( A \) explicitly,

\[ \frac{\mathbb{P}(\text{Evidence}|\text{Guilty})}{\mathbb{P}(\text{Evidence}|\text{Innocent})} \geq A \] (72)

This follows directly applying Bayes’ Theorem,

\[ \frac{\mathbb{P}(E|G)}{\mathbb{P}(E|G^c)} = \frac{\mathbb{P}(G|E) \cdot \mathbb{P}(G^c)}{\mathbb{P}(G^c|E) \cdot \mathbb{P}(G)} = \frac{p^* \cdot (1 - p)}{(1 - p^*) \cdot p} > (1 - u) \cdot \left( \frac{p}{1 - p} \right)^{-1} \] (73)

\[ \text{This can be interpreted as the worst outcome is that someone is found innocent when actually guilty and the best outcome is you determine someones guilt or innocence correctly.} \]
The decision made depends on two components - the strength of your preferences (u(0, 1)) and your assessment of the likelihood someone is guilty as opposed to innocent.

4.2 Question 2

A client’s utility function has three mutually utility independent attributes (m.u.i.a.) where each attribute (X_i) can take one of only two values, where U_i(X_i) = 1 denotes the successful outcome of the i^{th} attribute and U_i(X_i) = 0 the failed outcome where U_i denotes the clients’ marginal utilities on X_i.

Letting \( \theta_{x_1,x_2,x_3}(d) \) denote the probability that \( \{X_1, X_2, X_3\} = \{x_1, x_2, x_3\} \) under decision \( d \in D \) prove that when \( \sum_{i=1}^{3} k_i = 1 \) then the expected utility of decision \( d \) is,

\[
\mathbb{E}(U(d)) = k_1 \cdot \theta_{1,0,0}(d) + k_2 \cdot \theta_{0,1,0}(d) + k_3 \cdot \theta_{0,0,1}(d) \\
+ (k_1 + k_2) \cdot \theta_{1,1,0}(d) + (k_1 + k_3) \cdot \theta_{1,0,1}(d) + (k_2 + k_3) \cdot \theta_{0,1,1}(d) + \theta_{1,1,1}(d) \quad (74)
\]

As it is stated we have m.u.i.a. we can apply multi-attribute utility theory as developed in class (see [1] §6.3 Thm 6.9 for more detail). As we have \( \sum_{i=1}^{3} k_i = 1 \) then,

\[
U(d) = \sum_{i=1}^{3} k_i U_i(x_i) \quad (75)
\]

Noting that \( U_i(x_i) \) has binary outcomes (it’s an indicator variable!), so \( \mathbb{E}(1) = \mathbb{P}(X) \) then consider,

\[
\mathbb{E}(U(d)) = \mathbb{E}\left( \sum_{i=1}^{3} k_i U_i(x_i) \right) \\
= \sum_{k=1}^{3} k_i \cdot \mathbb{E}(U_i(x_i)) = \sum_{k=1}^{3} k_i \cdot \mathbb{P}(X_i = i) \\
= k_1 \cdot (\theta_{1,0,0}(d) + \theta_{1,1,0}(d) + \theta_{1,0,1}(d) + \theta_{1,1,1}(d)) \\
+ k_2 \cdot (\theta_{0,1,0}(d) + \theta_{1,1,0}(d) + \theta_{0,1,1}(d) + \theta_{1,1,1}(d)) \\
+ k_3 \cdot (\theta_{0,0,1}(d) + \theta_{1,0,1}(d) + \theta_{0,1,1}(d) + \theta_{1,1,1}(d)) \\
= k_1 \cdot \theta_{1,0,0}(d) + k_2 \cdot \theta_{0,1,0}(d) + k_3 \cdot \theta_{0,0,1}(d) \\
+ (k_1 + k_2) \cdot \theta_{1,1,0}(d) + (k_1 + k_3) \cdot \theta_{1,0,1}(d) + (k_2 + k_3) \cdot \theta_{0,1,1}(d) + \theta_{1,1,1}(d) \quad (76)
\]

and when \( \sum_{i=1}^{3} k_i \neq 1 \) (letting \( s_{i,j} = k_i + k_j + K \cdot k_i \cdot k_j \) for \( i \neq j \)),

\[
\mathbb{E}(U(d)) = k_1 \cdot \theta_{1,0,0}(d) + k_2 \cdot \theta_{0,1,0}(d) + k_3 \cdot \theta_{0,0,1}(d) + s_{1,2} \cdot \theta_{1,1,0}(d) + s_{1,3} \cdot \theta_{1,0,1}(d) + s_{2,3} \cdot \theta_{0,1,1}(d) + \theta_{1,1,1}(d) \quad (77)
\]

Referring to the case in [1] §6.3 Thm 6.9 where \( \sum_{i=1}^{3} k_i \neq 1 \), constructing some \( K = \prod_{i} (1 + k_i) - 1 \) (the normalizing constant which scales \( U(x) \) to the range [0, 1]) and denoting \( r_i = K \cdot k_i \) then,

\[
U(d) = \frac{1}{K} \left[ \prod_{i} \left( r_i \cdot U_i(x_i) + 1 \right) - 1 \right] \quad (78)
\]
Again as \( U_i(X_i) \) has binary outcomes it’s an indicator variable and so,

\[
K \cdot \mathbb{E}(U(d)) = \mathbb{E}\left( \prod_{i=1}^{3} (r_i \cdot U_i(x_i) + 1) - 1 \right)
\]

\[
= \sum_{i=1}^{3} r_i \cdot \mathbb{E}[U_i(x_i)] + \sum_{i \neq j=1}^{3} r_i \cdot r_j \cdot \mathbb{E}[U_i(x_i) \cdot U_j(x_j)] + r_1 \cdot r_2 \cdot r_3 \cdot \mathbb{E}[U_1(X_1) \cdot U_2(X_2) \cdot U_3(X_3)]
\]

\[
= \sum_{i=1}^{3} r_i \cdot \mathbb{P}(X_i = 1) + \sum_{i \neq j=1}^{3} r_i \cdot r_j \cdot \mathbb{P}(X_i = 1, X_j = 1) + r_1 \cdot r_2 \cdot r_3 \cdot \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1)
\]

\[
= r_1 \cdot (\theta_{1,0,0}(d) + \theta_{1,1,0}(d) + \theta_{1,0,1}(d) + \theta_{1,1,1}(d)) + \ldots + r_1 \cdot r_2 \cdot (\theta_{1,1,0}(d) + \theta_{1,1,1}(d)) + \ldots + r_1 \cdot r_2 \cdot r_3 \cdot \theta_{1,1,1}(d)
\]

(79)

so by dividing through by \( K \) and simple re-arrangement we find our result as desired,

\[
\mathbb{E}(U(d)) = k_1 \cdot \theta_{1,0,0} + k_2 \cdot \theta_{0,1,0} + k_3 \cdot \theta_{0,0,1} + s_{1,2} \cdot \theta_{1,1,0} + s_{1,3} \cdot \theta_{1,0,1} + s_{2,3} \cdot \theta_{0,1,1} + \theta_{1,1,1}
\]

(80)

Discuss how this responds to different values of criterion weights.

Noting that \( s_{i,j} = k_i + k_j + K \cdot k_i \cdot k_j \) for \( i \neq j \) we see that as \( K \to 0 \) we obtain a neutral position whereas \( K < 0 \) down-weights and \( K > 0 \) up-weights the gain from having two successes.

### 4.3 Question 3

Let \( D_x U(x) \) and \( D^2_x U(x) \) represent respectively the first and second derivatives of \( U(x) \) and assume \( U(x) \) is strictly increasing in \( x \) \( \forall x \).

Suppose your client tells you the following - If she is offered a choice to add to her current fortune an amount \( x \) for certain and she finds this equivalent to a gamble which gives a reward \( x + h \) with probability \( \alpha \) and \( x - h \) with probability \( 1 - \alpha \) (for some \( h \neq 0 \)) then her value of \( \alpha(h) \) may depend on \( h \) but does not depend on \( x \).

Letting \( \lambda(h) = \frac{1 - 2 \cdot \alpha(h)}{h \cdot \alpha(h)} \) prove that,

\[
\lambda(h) \cdot \left( \frac{U(x) - U(x - h)}{h} \right) = \left( \frac{U(x + h) - 2 \cdot U(x) + U(x - h)}{h^2} \right)
\]

(81)

Equivalence of the gambles implies their expected utilities are the same,

\[
U(x) = \alpha(h) \cdot U(x + h) + (1 - \alpha(h)) \cdot U(x - h)
\]

(82)

Hence,

\[
U(x) - \alpha(h) \cdot U(x) - (1 - \alpha(h)) \cdot U(x - h) = \alpha(h) \cdot U(x + h) - \alpha(h) \cdot U(x)
\]

\[
(1 - \alpha(h)) \cdot (U(x) - U(x - h)) = \alpha(h) \cdot (U(x + h) - U(x))
\]

\[
(1 - 2\alpha(h)) \cdot (U(x) - U(x - h)) = \alpha(h) \cdot (U(x + h) - 2U(x) + U(x - h))
\]

\[
\frac{1 - 2 \cdot \alpha(h)}{h \cdot \alpha(h)} \cdot \frac{U(x) - U(x - h)}{h} = \frac{U(x + h) - 2U(x) + U(x - h)}{h^2}
\]

\[
\lambda(h) \cdot \left( \frac{U(x) - U(x - h)}{h} \right) = \left( \frac{U(x + h) - 2 \cdot U(x) + U(x - h)}{h^2} \right)
\]

(83)

Prove that her utility function \( U(x) \) must either: be linear in the case where \( \lambda = \lim_{h \to 0} \lambda(h) = 0 \); or; otherwise
satisfy \( AD_sU(x) = D_s^2U(x) \) and take the form \( U(x) = A + B \exp(\lambda x) \lambda \neq 0 \) (by noting that for any continuous differentiable function \( f(x) \) that \( D_s \log f(x) = \frac{D_s f(x)}{f(x)} \).

In the case of linearity simply consider (83). Now if \( \lambda = \lim_{h \to 0} \lambda(h) = 0 \) we have for all \( x \) that,

\[
D_s^2U(x) = \lim_{h \to 0} \left[ \frac{U(x + h) - 2U(x) + U(x - h)}{h^2} \right] = 0
\]

(84)

Implying that \( U(x) \) is linear.

Considering (83) in the case where \( \lim_{h \to 0} \lambda(h) \neq 0, \) recalling \( AD_sU(x) = D_s^2U(x) \) and applying the identity in the question by letting \( f(x) = D_sU(x) \) we have that if \( D_s^2U(x) \neq 0 \) then,

\[
D_s [\log (D_s U(x))] = \frac{D_s [D_s U(x)]}{D_s U_s} = \frac{D_s^2 U(x)}{D_s U_s} = \lambda
\]

\[
\log (D_s U(x)) = \beta + \lambda x
\]

\[
D_s U(x) = \exp(\beta + \lambda x)
\]

\[
U(x) = A + B \exp(\lambda x)
\]

(86)

where \( A \) and \( B \) are some constants and noting that to ensure an increasing utility function \( B \) must be of the same sign as \( \lambda \). Note that this characterization is is the same for small and large values of \( h \).

5 Exercise Sheet 5

5.1 Question 1

For the purposes of assessing probabilities given by weather forecasters you choose to use the Brier scoring rule \( S_1(a, q) \) where,

\[
S_1(a, q) = (a - q)^2
\]

(87)

You suspect however that for some \( \lambda > 0 \) that your clients’ utility function is not linear but is instead of the form,

\[
U(S_1) = 1 - S_1^4
\]

(88)

Prove that if this assertion is true that your client will quote \( q = p \) for all values of \( p \) \( \iff \lambda = 1 \).

Prove that if \( \lambda > 1/2 \) and you are able to elicit the value of \( \lambda \) then you are able to express the clients’ true probability \( p \) as a function of their quoted probability \( q \). Write down this function explicitly and explain in what sense when \( 1/2 < \lambda < 1 \) your client will appear overconfident and if \( \lambda > 1 \) under-confident in her probability predictions.

Prove that if \( 0 \leq \lambda \leq 1/2 \) it will only be possible to determine from her quoted probability whether or not she believes that \( p \leq 1/2 \).

Note: \( a \) is some indicator variable representing the binary outcome of rain or no rain.

We are interested in the question - What \( q \) will the client quote when \( p \) is the true probability?

If the client is rational they will attempt to maximize their expected utility,

\[
\mathbb{E}(U(S_1)) = \mathbb{P}(a = 0) \cdot U(S_1(0, q)) + \mathbb{P}(a = 1) \cdot U(S_1(1, q))
\]

\[
= (1 - p) \cdot U(q^2) + p \cdot U((1 - q)^2)
\]

\[
= (1 - p) \cdot (1 - q^2) + p \cdot (1 - (1 - q)^2)
\]

(89)
Now, finding the \( q \) which maximizes (89) we have,

\[
\frac{d}{dq} \mathbb{E}(U(S_1)) = (1 - p) \cdot (-2\lambda q^{2\lambda-1}) + p \cdot (2\lambda(1-q)^{2\lambda-1}) = 0
\]  

(90)

which holds whenever \( \lambda \neq 1/2 \) and \( p \neq 0, 1 \) iff,

\[
p = \frac{1}{1-q} \left[ \frac{q}{1-q} \right]^{2\lambda-1}
\]  

(91)

\[
\log(p) - \log(1-p) = (2\lambda - 1) \cdot [\log(q) - \log(1-q)]
\]  

(92)

\[
p = \frac{\left( \frac{q}{1-q} \right)^{2\lambda-1}}{1 + \left( \frac{q}{1-q} \right)^{2\lambda-1}}
\]  

(93)

Checking the second derivative we have,

\[
-2\lambda \cdot (2\lambda - 1) [(1-q)^{2\lambda+2} \cdot p + q^{2\lambda+2} \cdot (1-p)]
\]  

(94)

which clearly shows the stationary point is a maximum iff \( \lambda > \frac{1}{2} \). Furthermore considering (93) note that \( q = p \implies \lambda = 1 \).

Considering (91) note that for \( \frac{1}{2} < \lambda < 1 \) then\(^5\),

\[
\left| q - \frac{1}{2} \right| > \left| p - \frac{1}{2} \right|
\]  

(95)

which can be interpreted as the client will make their forecasts closer to 0 or 1 than they should be. On the other hand if \( \lambda > 1 \) then,

\[
\left| q - \frac{1}{2} \right| < \left| p - \frac{1}{2} \right|
\]  

(96)

which can be interpreted as the statements will be more conservative.

Considering the case where \( 0 \leq \lambda \leq \frac{1}{2} \), note that (94) will imply the stationary points will occur at either 0 or 1, and as such if \( p > \frac{1}{2} \) then to maximize utility then \( q = 1 \) will be chosen and if \( q < \frac{1}{2} \) then \( q = 0 \) will be chosen.

5.2 Question 2

On each of 1000 consecutive days, two probability forecasters \( F_1 \) and \( F_2 \) state one of 6 possible outcomes which represent the probabilities that rain will occur the next day \( \{q(1) = 0, q(2) = 0.2, q(3) = 0.4, q(4) = 0.6, q(5) = 0.8, q(6) = 1\} \). The results of their selection and actual occurrence of rain is given in Table 5.

5.2.1 Question 2 i)

Demonstrate which, if either, of the two forecasters are well calibrated and calculate each of their empirical Brier scores. Who has the better score?

\(^5\)To see this note that if \( \lambda \in \left( \frac{1}{2}, 1 \right) \), then \( 2\lambda - 1 \in (0, 1) \). Now fixing \( \lambda \), if \( p < 1/2 \) then \( \frac{p}{1-p} < 1 \), which in turn implies that as \( \frac{q}{1-q} \) has been risen to a power less than 1 that \( \frac{q}{1-q} < \frac{p}{1-p} \), meaning that \( q \) and \( 1-q \) are closer to one another, implying \( q - \frac{1}{2} > p - \frac{1}{2} \). On the other hand if \( p > 1/2 \) then \( \frac{p}{1-p} > 1 \), which in turn implies that as \( \frac{q}{1-q} \) has been risen to a power less than 1 that \( \frac{q}{1-q} > \frac{p}{1-p} \), meaning that \( q \) and \( 1-q \) are closer to one another, implying \( q - \frac{1}{2} > p - \frac{1}{2} \).
With reference to Table 6 which contains marginal counts note that the second forecaster \((F_2)\) is well calibrated as for every quote \(q = \hat{q}\). The first forecaster \((F_1)\) is not well calibrated as \(0.4 = 0.6 = 50/100 = 0.5\)

Denoting for the \(i^{th}\) prediction \(a_i\) as the actual outcome (binary), \(f_i\) as the forecasted probability and \(N\) as the total number of predictions then the Brier score is simply,

\[
\frac{1}{N} \sum_{i=1}^{N} (a_i - f_i)^2
\]

\((97)\)

\(F_1\)’s Brier score is

\[
\frac{1}{1000} \sum_{i=1}^{1000} (a_i - f_i)^2 = \left\{ 200((0/200) \cdot (1 - 0)^2 + (200/200) \cdot (0 - 0)^2) \\
+ 200((40/200) \cdot (1 - 0.2)^2 + (160/200) \cdot (0 - 0.2)^2) \\
+ 100((50/100) \cdot (1 - 0.4)^2 + (50/100) \cdot (0 - 0.4)^2) \\
+ 100((50/100) \cdot (1 - 0.6)^2 + (50/100) \cdot (0 - 0.6)^2) \\
+ 200((160/200) \cdot (1 - 0.8)^2 + (40/200) \cdot (0 - 0.8)^2) \\
+ 200((200/200) \cdot (1 - 1)^2 + (0/200) \cdot (0 - 1)^2) \right\}/1000
= 0.116
\]

\((98)\)
F2’s Brier score is

\[
\frac{1}{1000} \sum_{i=1}^{1000} (a_i - f_i)^2 = \left\{ 100\left(\frac{0}{100}\right) \cdot (1 - 0)^2 + \frac{100}{100} \cdot (0 - 0)^2 \right\} 
+ 100\left(\frac{20}{100}\right) \cdot (1 - 0.2)^2 + \left(\frac{80}{100}\right) \cdot (0 - 0.2)^2 
+ 300\left(\frac{120}{300}\right) \cdot (1 - 0.4)^2 + \left(\frac{180}{300}\right) \cdot (0 - 0.4)^2 
+ 300\left(\frac{180}{300}\right) \cdot (1 - 0.6)^2 + \left(\frac{120}{300}\right) \cdot (0 - 0.6)^2 
+ 100\left(\frac{80}{100}\right) \cdot (1 - 0.8)^2 + \left(\frac{20}{100}\right) \cdot (0 - 0.8)^2 
+ 100\left(\frac{100}{100}\right) \cdot (1 - 1)^2 + \left(\frac{0}{100}\right) \cdot (0 - 1)^2 \right\} / 1000 
= 0.176 \quad (99)
\]

F1’s score is better as it is lower, despite F1 being uncalibrated.

Adapt the forecasts of any uncalibrated forecaster so that their empirical Brier score improves and calculate the extent of this improvement.

Merging \( q_1(i) = 0.4 \) and \( q_1(i) = 0.6 \) into \( q_1(i) = 0.5 \) in Table 6 will result in F1 being well calibrated and will reduce the Brier score by 0.01 \( [2 \cdot 50 \cdot 0.6^2 + 2 \cdot 50 \cdot 0.4^2 - 200 \cdot 0.5^2] / 200 \).

5.2.2 Question 2 ii)

It is suggested to you that improvements to the forecast could be achieved by combining in some way the two forecasts. Find a probability formula which gives an empirical Brier score of zero based on the data above.

Notice that in Table 6 jointly F1 and F2 can predict the rain with certainty. Simplifying Table 6 to Table 7 we see that if jointly F1 and F2 predict rain then with probability of 1 predict rain, otherwise predict rain with probability of 0. The joint prediction is well calibrated and has a Brier score of 0.

<table>
<thead>
<tr>
<th>( q_1(i) )</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
</tr>
<tr>
<td>0.2 No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>Rain</td>
<td>Rain</td>
</tr>
<tr>
<td>0.4 No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
</tr>
<tr>
<td>0.6 No Rain</td>
<td>No Rain</td>
<td>No Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
</tr>
<tr>
<td>0.8 No Rain</td>
<td>No Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
</tr>
<tr>
<td>1.0 No Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
<td>Rain</td>
</tr>
</tbody>
</table>

Table 7: Simplified Joint Forecast.

6 Exercise Sheet 6

6.1 Recap - Relevance

- \( X \perp \!\!\!\!\perp Y \iff p_{X,Y}(x,y) = p_X(x)p_Y(y) \)
- \( X \perp \!\!\!\!\perp Y \iff Y \perp \!\!\!\!\perp X \)
- \( X \perp \!\!\!\!\perp Y|Z \iff p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \)
For some \( X, Y, Z \) then \( X \perp Y|Z \iff Y \perp X|Z \)

For some \( X, Y, Z, W \) then \( (X \perp Y|Z, W), X \perp (Y, Z)|W \iff X \perp (Y, Z)|W \)

6.2 Question 1

One alternative to the usual definition of irrelevance might be to say \( Y \) is irrelevant for forecasting \( X \) given the information in \( Z \) if the expectation of \( X \) depended on \( (Y, Z) \) only through it’s dependence on \( Z \). Check whether properties P1 and P2 hold under this definition.

Noting that as \( X \perp Y|Z \iff p_{X|YZ}(y|z) = p_{XY}(x|z)p_{YZ}(y|z) \), we have that Bayes rule gives \( p_{X|YZ}(x|y, z) = p_{XZ}(x|z) \) so as a result,

\[
\mathbb{E}(X|Y, Z) := \int x \cdot p_{X|YZ}(x|y, z) \, dx = \int x \cdot p_{XZ}(x|z) \, dx =: \mathbb{E}(X|Z) \tag{100}
\]

In addition we have \( X \perp Y|Z \iff Y \perp X|Z \), however considering Table 8 we have that \( \mathbb{E}(Z|X, Y) = \mathbb{E}(Z|Y) \) however \( \mathbb{E}(X|Z, Y) \neq \mathbb{E}(X|Z) \).

Table 8: Outcome Table

<table>
<thead>
<tr>
<th>( X = Y )</th>
<th>( Z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0/1</td>
</tr>
<tr>
<td>1/4</td>
<td>0/1</td>
</tr>
</tbody>
</table>

6.3 Question 2 - Simpson’s Paradox

As the statistical meaning of “independence” is not the same as its colloquial meaning this leads to an elementary but plausible error made by scientists is to believe that if \( X \) is independent of \( Y \) given \( W \) then \( X \) must be independent of \( Y \). Construct a simple joint distribution over three binary random variables \( \{X, Y, W\} \) to show that this can’t be deduced in general.

Consider the following joint mass function,

\[
\mathbb{P}(w = 0) = \mathbb{P}(w = 1) = 1/2 \\
\mathbb{P}(x = 0|w = 0) = 1 - \mathbb{P}(x = 0|w = 1) = 2/3 \\
\mathbb{P}(x = 1|w = 0) = 1 - \mathbb{P}(x = 1|w = 1) = 1/3 \\
\mathbb{P}(y = 0|w = 0) = 1 - \mathbb{P}(y = 0|w = 1) = 1/5 \\
\mathbb{P}(y = 1|w = 0) = 1 - \mathbb{P}(y = 1|w = 1) = 4/5
\]

As a result we have the following joint mass function,

\[
\mathbb{P}(x = 0, y = 0) = \mathbb{P}(x = 0, y = 0, w = 0) + \mathbb{P}(x = 0, y = 0, w = 1) \\
= \mathbb{P}(x = 0|y = 0, w = 0)\mathbb{P}(y = 0|w = 0)\mathbb{P}(w = 0) + \mathbb{P}(x = 0|y = 0, w = 1)\mathbb{P}(y = 0|w = 1)\mathbb{P}(w = 1) \\
= \mathbb{P}(x = 0|w = 0)\mathbb{P}(y = 0|w = 0)\mathbb{P}(w = 0) + \mathbb{P}(x = 0|w = 1)\mathbb{P}(y = 0|w = 1)\mathbb{P}(w = 1) \\
= 2/3 \cdot 4/5 \cdot 1/2 + 1/3 \cdot 1/5 \cdot 1/2 = 3/10
\]

\[
\mathbb{P}(x = 0, y = 1) = \mathbb{P}(x = 0, y = 1, w = 0) + \mathbb{P}(x = 0, y = 1, w = 1) \\
= \mathbb{P}(x = 0|y = 1, w = 0)\mathbb{P}(y = 1|w = 0)\mathbb{P}(w = 0) + \mathbb{P}(x = 0|y = 1, w = 1)\mathbb{P}(y = 1|w = 1)\mathbb{P}(w = 1) \\
= \mathbb{P}(x = 0|w = 0)\mathbb{P}(y = 1|w = 0)\mathbb{P}(w = 0) + \mathbb{P}(x = 0|w = 1)\mathbb{P}(y = 1|w = 1)\mathbb{P}(w = 1) \\
= 2/3 \cdot 1/5 \cdot 1/2 + 1/3 \cdot 4/5 \cdot 1/2 = 1/5
\]

\[
\mathbb{P}(x = 1, y = 0) = \ldots = 1/5 \\
\mathbb{P}(x = 1, y = 1) = \ldots = 3/10
\]
Together (104) and (105) imply $X \perp W$ and $X \perp Y|W = 1$ but we don’t have $X \perp Y$.

Why must it be true if, in addition, either $X$ or $Y$ is independent of $W$?

We have,

$$X \perp W \iff p_{X,W}(x, w) = p_X(x)p_W(w)$$

$$X \perp Y|W \iff p_{X,Y|W}(x, y|w) = p_{X|W}(x|w)p_{Y|W}(y|w)$$

Now, $p_{X,Y}(x, y) = \sum_W p_{X,Y|W}(x, y|w)p_W(w)$

$$= \sum_W p_{X|W}(x|w)p_{Y|W}(y|w)p_W(w)$$

$$= \sum_W p_X(x)p_{Y|W}(y|w)p_W(w)$$

$$= p_X(x)p_{Y|W}(y)p_W(w)$$

$$= p_X(x)p_Y(y) \iff X \perp Y$$

6.4 Question 3 - Positivity

Another axiom of irrelevance which is sometimes used is to demand that if $Y$ is irrelevant about $X$ once $Z$ is known and $X$ is irrelevant about $X$ once $Y$ is known then the pair $[Y, Z]$ must be irrelevant to $X$.

By considering the case $X = Y = Z$ show that this can’t always be deduced for the usual conditional independence definition of irrelevance. Show however that when $[X, Y, Z]$ have a discrete distribution and that all combinations of these variables are possible (i.e. $\Pr(X = x, Y = y, Z = z) > 0 \forall$ possible $x, y, z$) then this property will hold for conditional independence.

Considering $X = Y = Z$ we have $X \perp X|X$, however we clearly don’t have $X \perp (X, X)$ as they are equal! As a result this can’t always be deduced.

Now considering the case where $X, Y, Z$ are discrete and the joint density is positive for all $x, y, z$ recall that, $X \perp (Y, Z) \iff p_{X,Y|Z}(x, y, z) = p_X(x)p_{Y|Z}(y, z)$ (101)

Noting that $X \perp Y|Z \iff p_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$ we have,

$$p_{X,Y,Z}(x, y, z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z)$$

$$= p_{X|Z}(x|z)p_{Y|Z}(y, z)$$

Similarly $X \perp Z|Y \iff p_{X,Z|Y}(x, z|y) = p_{X|Y}(x|y)p_{Z|Y}(z|y)$ means we have,

$$p_{X,Y,Z}(x, y, z) = p_{X|Y}(x|y)p_{Z|Y}(z|y)p_Y(y)$$

$$= p_{X|Y}(x|y)p_{Z|Y}(y, z)$$

Equating (101), (102) and (103) we have $p_X(x) = p_{X|Z}(x|z) = p_{X|Y}(x|y)$ and so,

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \iff X \perp Y$$

$$p_{X,Z}(x, z) = p_X(x)p_Z(z) \iff X \perp Z$$

Together (104) and (105) imply $X \perp (Y, Z)$. 

20
7 Exercise Sheet 7

7.1 Question 1

A Markov chain \( \{X_i\}_{i=1}^{n} \) has the property that \( \forall i \in [3, \ldots, n] \ X_i | X_{i-1} \perp \perp \{X_j\}_{j<i} \). Representing this as an influence diagram, use d-Separation\(^6\) to prove the following.

The DAG of this process takes the form of that in Figure 4 below,

\[
\circ X_1 \rightarrow \circ X_2 \rightarrow \cdots \circ X_{i-2} \rightarrow \circ X_{i-1} \rightarrow \circ X_i \rightarrow \circ X_{i+1} \rightarrow \circ X_{i+2} \rightarrow \cdots \rightarrow \circ X_{n-1} \rightarrow \circ X_n
\]

Figure 4: Ex7 Q1 - DAG

The DAG in Figure 4 is clearly decomposable so we can check for irrelevance by considering separation on the moralized and undirected version of this graph as in Figure 5.

\[
\circ X_1 \rightarrow \circ X_2 \rightarrow \cdots \circ X_{i-2} \rightarrow \circ X_{i-1} \rightarrow \circ X_i \rightarrow \circ X_{i+1} \rightarrow \circ X_{i+2} \rightarrow \cdots \rightarrow \circ X_{n-1} \rightarrow \circ X_n
\]

Figure 5: Ex7 Q1 - Moralized & Undirected

7.1.1 Question 1a)

\( R_1(X_{i-1}) \perp \perp R_2(X_{i-1}) | X_{i-1} \), where \( R_1(X_{i-1}) = \{X_1, \ldots, X_{i-2}\} \) and \( R_2(X_{i-1}) = \{X_i, \ldots, X_n\} \) (i.e. our predictions about the future of the process depends only on the last observation).

As shown in Figure 6 \( R_1 \) is clearly separated from \( R_2 \) by \( X_{i-1} \) and so \( R_1 \perp \perp R_2 | X_{i-1} \).

\[
\circ X_1 \rightarrow \circ X_2 \rightarrow \cdots \circ X_{i-2} \bullet X_{i-1} \circ X_i \rightarrow \circ X_{i+1} \rightarrow \circ X_{i+2} \rightarrow \cdots \circ X_{n-1} \circ X_n
\]

Figure 6: Ex7 Q1a

7.1.2 Question 1b)

\( P \perp \perp X_i | (X_{i-1}, X_{i+1}) \) for \( 2 \leq i \leq n - 1 \) and \( P = \{X_1, X_2, \ldots, X_{i-2}, X_{i+2}, \ldots, X_n\} \) (i.e. given the variables preceding and succeeding \( X_i \) no other observation gives further relevant information about \( X_i \)).

As shown in Figure 7 \( P \) is separated from \( X_i \) given \( X_{i-1} \) and \( X_{i+1} \) so \( P \perp \perp X_i | (X_{i-1}, X_{i+1}) \).

\[
\circ X_1 \rightarrow \circ X_2 \rightarrow \cdots \circ X_{i-2} \bullet X_{i-1} \circ X_i \bullet X_{i+1} \circ X_{i+2} \rightarrow \cdots \circ X_{n-1} \circ X_n
\]

Figure 7: Ex7 Q1b

7.2 Question 2

The DAG of a valid influence diagram is given in Figure 8. Use the d-Separation theorem to check the following irrelevance statements,

To begin let's find the corresponding moralized undirected graph as shown in Figure 9.

---

\(^6\) d-Separation Theorem: Let \( X, Y, Z \) be disjoint sets of nodes in a valid DAG - \( G \). Then \( Y \) is irrelevant for predictions about \( X \mid Z \) is a valid deduction from \( G \) iff in the undirected version of the moralized graph \( G[S \mid m] \) of the ancestral DAG on the set \( S = \{X, Y, Z\} \), \( Z \) separates \( X \) from \( Y \).
Figure 8: Ex7 Q2 - DAG

Figure 9: Ex7 Q2 - Undirected moralized graph
7.2.1 Question 2a)

\((X_3, X_4) \perp X_7 | (X_5, X_6)\)

As shown in Figure 10 removal of \(X_5\) and \(X_6\) results in \(X_3\) and \(X_4\) being seperated from \(X_7\), so we can conclude that \((X_3, X_4) \perp X_7 | (X_5, X_6)\)

![Figure 10: Ex7 Q2a](image)

7.2.2 Question 2b)

\(X_2 \perp X_3 | (X_1, X_4)\)

As shown in Figure 11 removal of \(X_1\) and \(X_4\) still leaves an edge between \(X_2\) and \(X_3\), so we can conclude that \(X_2 \not\perp X_3 | (X_1, X_4)\)

![Figure 11: Ex7 Q2b & Q2c](image)

7.2.3 Question 2c)

\(X_2 \perp (X_5, X_8) | (X_1, X_4)\)

As shown in Figure 11 removal of \(X_1\) and \(X_4\) leaves \(X_1\) connected to \(X_5\) and \(X_8\) via \(X_3\), so we can conclude that \(X_2 \not\perp (X_5, X_8) | (X_1, X_4)\)
7.3 Question 3

Let \( W = (X, Y, Z) \) have a multivariate normal distribution with covariance matrix \( V \) of full rank and write (noting that \( P_{12} = P_{21}, P_{13} = P_{31} \) and \( P_{23} = P_{32} \)),

\[
V^{-1} = P = \begin{pmatrix}
P_{11} & P_{12} & P_{13} \\
P_{21} & P_{22} & P_{23} \\
P_{31} & P_{32} & P_{33}
\end{pmatrix}
\] (106)

By considering the form of the log joint density of \( W \) show that \( X \perp Y | Z \iff P_{12} \) is the matrix of all zeros.

Denoting \( n \) as the dimension of \( w \) we have,

\[
f(w) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp \left\{ -\frac{1}{2} (w - \mu)^T V^{-1} (w - \mu) \right\}
\]

\[
\log(f(w)) = \frac{1}{2} \log(2\pi) - \frac{1}{2} \log(V) - \frac{1}{2} (w - \mu)^T V^{-1} (w - \mu)
\]

\[
-2 \log(f(x, y, z)) = n \log(2\pi) - \log(P) + (x - \mu_x)^T P_{11}(x - \mu_x) + (x - \mu_x)^T P_{12}(y - \mu_y) + (x - \mu_x)^T P_{13}(z - \mu_z)
\]

\[
+ (y - \mu_y)^T P_{21}(x - \mu_x) + (y - \mu_y)^T P_{22}(y - \mu_y) + (y - \mu_y)^T P_{23}(z - \mu_z)
\]

\[
+ (z - \mu_z)^T P_{31}(x - \mu_x) + (z - \mu_z)^T P_{32}(y - \mu_y) + (z - \mu_z)^T P_{33}(z - \mu_z)
\]

Now as we have \( f_{X,Y,Z}(x, y, z) = \frac{f_{X,Y,Z}(x, y, z)}{f_Z(z)} \), we have \(-2 \log(f_{X,Y,Z}(x, y, z)) = -2 \log(f_{X,Y,Z}(x, y, z)) + 2 \log(f_Z(z))\) and as a result we can simply split our conditional density into components which depend on \( z \) and those which depend only on \( x \) and \( y \),

\[
-2 \log(f(x, y | z)) = \tau(z) + \tau(x, z) + \tau(y, z) + \tau(x, y)
\] (107)

where,

\[
\tau(z) = (z - \mu_z)^T P_{33}(z - \mu_z) + 2 \log(f_Z(z))
\]

\[
\tau(x, z) = (x - \mu_x)^T P_{11}(x - \mu_x) + (x - \mu_x)^T P_{13}(z - \mu_z) + (z - \mu_z)^T P_{33}(x - \mu_x)
\]

\[
\tau(y, z) = (y - \mu_y)^T P_{22}(y - \mu_y) + (y - \mu_y)^T P_{23}(z - \mu_z) + (z - \mu_z)^T P_{33}(y - \mu_y)
\]

\[
\tau(x, y) = (x - \mu_x)^T P_{12}(y - \mu_y) + (y - \mu_y)^T P_{23}(x - \mu_x)
\] (108)

Recalling that \( P_{12} = P_{21} \) we find that \( \tau(x, y) = 0 \iff P_{12} = 0 \) - i.e. \( X \perp Y | Z \iff P_{12} = 0 \).

8 Exercise Sheet 8

8.1 Question 1

Elementary particles are emitted independently from a nuclear source. If \( X_i \) denotes the time before the first emission, in minutes, and \( X_i \) denotes the time between the emission of the \((i - 1)^{th}\) and the \(i^{th}\) particle \((i = 1, 2, 3, 4, \ldots)\) we know that the density \( f(x|\theta) \) of each \( X_i \) is given by,

\[
f_1(x|\theta) = \begin{cases} 
\theta \exp(-\theta x) & x > 0, \quad y > 0 \\
0 & \text{otherwise}
\end{cases}
\] (109)

When you first obtain the data you find that you have only been given the number \( y_j \) of emissions in the interval \((j - 1, j]\). You know, however, that from the above \( Y_j (j = 1, 2, \ldots) \) are independent with the following mass function,

\[
f_2(y|\theta) = \begin{cases} 
\frac{\theta^y}{y!} \exp(-\theta) & y = 0, 1, 2, \ldots \quad \theta > 0 \\
0 & \text{otherwise}
\end{cases}
\] (110)
You take in observations and your last observation arrives exactly on the $n^{th}$ minute. Your prior distribution for $\theta$ was a Gamma distribution $G(\alpha^*, \beta^*)$, with density $\pi(\theta)$ where,

$$
\pi(\theta) \propto \begin{cases} 
\theta^{\alpha^* - 1} \exp[-\beta^* \theta] & \text{if } \theta, \alpha, \beta > 0 \\
0 & \text{otherwise} 
\end{cases}
$$

Find your posterior distribution for $\theta$,

i) using $y_1, \ldots, y_n$

ii) using $x_1, \ldots, x_n$

and show that the two analyses give identical inferences on $\theta$.

Calculating the likelihoods of i) and ii) and noting that the sum of the particle emissions in each period will produce the total number emissions ($\sum_{i=1}^{n} y_i = m$) and that the sum of time between emissions to some time $n$ where an emission occurs will produce the length of the total interval ($\sum_{i=1}^{n} x_i = n$), we have (for the parameters conforming to the above restrictions),

$$
L(\theta|y) = \prod_{i=1}^{n} f_2(y_i|\theta) \\
\propto \prod_{i=1}^{n} \theta^{y_i} \exp[-\theta] = \theta^{\sum_{i=1}^{n} y_i} \exp[-n\theta] = \theta^m \exp[-n\theta] 
$$

$$
L(\theta|x) = \prod_{i=1}^{m} f_2(x_i|\theta) \\
= \prod_{i=1}^{m} \theta \exp[-\theta x_i] = \theta^m \exp\left\{-\sum_{i=1}^{m} x_i \theta\right\} = \theta^m \exp[-n\theta] 
$$

Noting that the prior doesn’t contain any data then,

$$
\pi(\theta|x) = \pi(\theta|y) \propto L(\theta|y)\pi(\theta) \\
= \theta^m \theta^{\alpha^* - 1} \exp[-\beta^* \theta] \\
= \theta^{m+\alpha^* - 1} \exp[-\theta(n + \beta^*)] 
$$

Noting that as we have a density our posterior will integrate to 1 and further noting that we have the same form as the Gamma distribution earlier we conclude that our posterior has a $G(m + \alpha^*, n + \beta^*)$ distribution.

If $\alpha^* > 1$ express the posterior mode of $\theta$ as a weighted average of $y$ and the prior mode (where the weight $\rho$ depends only on $\beta^*/n$).

Denoting $m_1$ as the posterior mode and $m_0$ as the prior mode, and further noting that for a $G(\alpha, \beta)$ distribution the mode $m = \frac{\alpha - 1}{\beta}$, we have,

$$
m_1 = \frac{m + \alpha^* - 1}{n + \beta^*} = \frac{m}{n + \beta^*} + \frac{\alpha^* - 1}{n + \beta^*} = \bar{y} \rho + m_0(1 - \rho) \\ 
\text{where } \rho = \frac{n}{n + \beta^*} 
$$
8.2 Question 2

8.2.1 Question 2a)

Show that if \( \mathcal{P} = \{p(\theta, \alpha) : \alpha \in A \subseteq \mathbb{R}^n \} \) is a family closed under sampling\(^7\) to the likelihood \( \mathcal{L}(\theta|x) \) then so is the family.

\[
\mathcal{P}^k = \left\{ p(\theta, \alpha) = \sum_{i=1}^{k} \gamma_i p_i(\theta, \alpha) : \sum_{i=1}^{k} \gamma_i = 1 \quad \gamma_i > 0 \quad p_i \in \mathcal{P} \quad 1 \leq i \leq k \right\}
\] (116)

Let \( p(\theta) \in \mathcal{P}^k \) we have \( p(\theta) = \sum_{i=1}^{k} \gamma_i p_i(\theta) \) where \( p_i(\theta) \in \mathcal{P} \) and denoting \( \tau(x) \) as some normalizing constant then note that by applying Bayes rule we have,

\[
p(\theta|x) = p(\theta) \mathcal{L}(\theta|x) \tau^{-1}(x) = \sum_{i=1}^{k} \gamma_i p_i(\theta) \mathcal{L}(\theta|x) \tau^{-1}(x)
\]

\[
= \sum_{i=1}^{k} \gamma_i \tau_i(x)p_i(\theta|x) \tau^{-1}(x)
\]

\[
= \sum_{i=1}^{k} \beta_i p_i(\theta|x) \quad \text{where} \quad \beta_i = \frac{\gamma_i \tau_i(x)}{\tau(x)}
\] (117)

Now as \( p_i(\theta|x) \in \mathcal{P} (\mathcal{P} \text{ is closed under sampling}) \) we have \( \mathcal{P}^k \) is closed under sampling.

8.2.2 Question 2b)

Suppose random variables \( X_1, \ldots, X_n \) constitute a random sample from a rectangular distribution (continuous uniform distribution) with density \( f(x|\theta) \) given by,

\[
f(x|\theta) = \begin{cases} \theta^{-1} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}
\] (118)

Letting

\[
p_i(\theta) = \begin{cases} \alpha \theta_i \theta^{-(\alpha+1)} & \theta > \theta_i \quad \text{where} \quad \alpha, \theta_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2
\] (119)

If a prior density of the form \( p(\theta) \) of the form

\[
p(\theta) = \begin{cases} \beta p_1(\theta) + (1-\beta)p_2(\theta) & 0 < \beta < 1 \\ 0 & \text{otherwise} \end{cases} \quad \theta_1 < \theta_2
\] (120)

is used, find the posterior density \( p(\theta|x) \) of \( \theta \).

Defining \( x^* = \max_{i} x_i \) then

\[
\mathcal{L}(\theta|x) = \begin{cases} \theta^{-n} & \theta > x^* \\ 0 & \text{otherwise} \end{cases}
\] (121)

Letting \( \theta_i^* = \max_{i} (x^*, \theta_i) \) (recall the \( i \) subscript relates to the \( i^{th} \) prior!) Then we find our \( i^{th} \) posterior,

\[
p_i(\theta|x) \propto \begin{cases} \alpha \theta_i^* \theta^{-(\alpha+1)} \theta^{-n} & \theta > \theta_i^* \\ 0 & \text{otherwise} \end{cases}
\] (122)

\(^7\text{Closed Under Sampling (See [1] \S 5.4 Thm 5.3): A family of prior distributions } \mathcal{P} = \{p(\theta, \alpha) : \alpha \in A \subseteq \mathbb{R}^n \} \text{ is closed under sampling for a likelihood } \mathcal{L}(\theta|x) \text{ if for any prior } p(\theta) \in \mathcal{P} \text{ the corresponding posterior distribution } p(\theta|x) \in \mathcal{P}. \text{ Intuition: It is convenient if the prior and posterior have the same distribution form as it typically results in more tractibility, the normalizing constant 'for free' and some understanding on the contributory weighting effect of the prior and the data. Intuition in this question: Think of } \mathcal{P}^k \text{ as a richer form built through mixing.} \)
Noting the form of \( p_i(\theta) \) above we can find the normalizing constant,

\[
p_i(\theta|x) = \begin{cases} 
(\alpha + n)\theta_0^{(\alpha+n)}\theta^{-(\alpha+n+1)} & \text{for } \theta > \theta_i^* \\
0 & \text{otherwise}
\end{cases}
\]  

(123)

Explicitly our normalizing constant is,

\[
\tau_i(x) = \frac{\alpha\theta_i^*}{(\alpha + n)\theta_0^{(\alpha+n)}},
\]

(124)

Now from Question 1 we have \( p(\theta|x) = \beta_1^* p_1(\theta|x) + (1 - \beta_1^*)p_2(\theta|x) \) where \( \beta_1^* = \beta_1 \tau_1(x)\tau^{-1}(x) \) and \( \beta_2^* = (1 - \beta_1^*) = \beta_2 \tau_2(x)\tau^{-1}(x) \). Cancellation gives,

\[
\beta_i^* = \frac{\theta_i^* \left(\theta_0^{(\alpha+n)}\right)^{-1} \beta_1}{\theta_1^* \left(\theta_0^{(\alpha+n)}\right)^{-1} \beta_1 + \theta_2^* \left(\theta_0^{(\alpha+n)}\right)^{-1} (1 - \beta)},
\]

(125)

Sketch the different samples \( \{x_1, x_2, \ldots, x_n\} \).

Letting \( \theta_1 < \theta_2 \) and noting that \( \theta_1^* \) and \( \theta_2^* \) both depend on \( x^* \), we find that we have 3 cases to consider: (i) \( x^* \leq \theta_1 < \theta_2 \); (ii) \( \theta_1 < x^* \leq \theta_2 \); (iii) \( \theta_1 < \theta_2 < x^* \).

Case (i) - \( x^* \leq \theta_1 < \theta_2 \) - Here we have \( \theta_1^* = \theta_1 \) and \( \theta_2^* = \theta_2 \) and as such,

\[
\beta_i^* = \frac{\theta_1^* \beta_1}{\theta_1 \left(\theta_0^{(\alpha+n)}\right)^{-1} \beta_1 + \theta_2 \left(\theta_0^{(\alpha+n)}\right)^{-1} (1 - \beta)} \rightarrow 1 \text{ as } n \rightarrow \infty
\]

(126)

Figure 12: Ex8 Q2 - Case (i)

Case (ii) - \( \theta_1 < x^* \leq \theta_2 \) - Here we have \( \theta_1^* = x^* \) and \( \theta_2^* = \theta_2 \) and as such,

\[
\beta_i^* = \frac{\theta_1 \left(\theta_0^{(\alpha+n)}\right)^{-1} \beta_1}{\theta_1 \left(\theta_0^{(\alpha+n)}\right)^{-1} \beta_1 + \theta_2 \left(\theta_0^{(\alpha+n)}\right)^{-1} (1 - \beta)} \rightarrow 1 \text{ as } n \rightarrow \infty
\]

(127)
Case (iii) - $\theta_1 < \theta_2 < x^*$ - Here we have $\theta_1^* = \theta_2^* = x^*$. As a result $\beta^* = 1/2$ and $p_1(\theta|x) = p_2(\theta|x) = p(\theta|x)$.

As indicated in Cases (i)-(iii) we find in Figure 15 that as $n$ increases our posterior density $p(\theta|x) \rightarrow p_1(\theta|x)$. Notice that this change is noticeable for even a small number of samples.

Considering the effect of $\beta$ on the posterior, note that in effect we are simply weighting two densities (in both the prior and posterior case). Figure 16 demonstrates this effect, in particular we see that as $\beta$ is increased $p_2(\theta)$ contributes more to the prior $p(\theta)$ which in turn changes the posterior $p(\theta|x)$.

### 8.3 Question 3

For integer $(\alpha, \beta)$ the Beta distribution $Be(\alpha, \beta)$ has density $f(\theta|\alpha, \beta)$ given by

$$f(\theta|\alpha, \beta) = \begin{cases} \frac{(\alpha + \beta - 1)!}{(\alpha - 1)! (\beta - 1)!} \theta^{\alpha-1} (1 - \theta)^{\beta - 1} & 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (128)

A coin is tossed $n$ times and the number of heads $x$ is counted. You choose a prior of the form,

$$\pi(\theta) = \gamma f(\theta|k, k) + (1 - \gamma) f(\theta|1, 1)$$  \hspace{1cm} (129)
Denoting \( n \to \infty \)

\[
\gamma \text{ is the mixture of two Beta distributions with posterior weights } \gamma^* \text{ and } (1 - \gamma^*) \text{ where } \gamma^* = (1 + t(n))^{-1} \text{ and } t(n) = (n + 1)^{-1}[n + 2k - 1](n + 2k - 1)\ldots(n + k)]/[2k - 1)(2k - 1)\ldots(k)]^{-1}.
\]

Show that for \( k \geq 2 \) then \( \gamma^* \to 0 \) as \( n \to \infty \) and interpret this result.

Denoting \( f_1(\theta) = f(\theta) \) and \( f_2(\theta) = f(\theta|1,1) \) then,

\[
f(\theta|x = 0) = \gamma f_1(\theta)L_1(\theta|x = 0) + (1 - \gamma) f_2(\theta)L_2(\theta|x = 0)
\]

\[
= \tau_1(x)f_1(\theta|x = 0) + \tau_2(x)f_2(\theta|x = 0)
\]

(130)

Now,

\[
f_1(\theta|x = 0) \propto \frac{(2k - 1)!}{(k - 1)!((k - 1)!} \theta^{k-1}(1-\theta)^{n-k-1} \frac{n!}{0!(n-0)!} \theta^n(1-\theta)^n
\]

\[
\sim \text{Be}(k, n + k)
\]

(131)

\[
f_2(\theta|x = 0) \propto \frac{(2 - 1)!}{(1 - 1)!((1 - 1)!} \theta^{k-1}(1-\theta)^{n-k-1} \frac{n!}{0!(n-0)!} \theta^n(1-\theta)^n
\]

\[
\sim \theta^{k-1}(1 - \theta)^{(n+1)-1}
\]

\[
\sim \text{Be}(1, 1+n)
\]

(132)

Estimating the normalizing constants for \( f_1(\theta|x = 0) \) and \( f_2(\theta|x = 0) \) is now trivial given their analytic density,

\[
\tau_1(x) = \frac{k + (n+k-1)!}{(k-1)!((n+k-1)!}
\]

\[
\tau_1(x) = \frac{(2k - 1)!}{(k - 1)!(k + n - 1)!}
\]

(133)

\[
\tau_2(x) = \frac{n!}{(n+1)!} = \frac{1}{n+1}
\]

(134)

Noting that \( \text{Be}(\alpha, \beta) \) has pdf \( \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} \theta^{\alpha-1}(1-\theta)^{\beta-1} \)

Figure 15: Ex8 Q2 - Effect of varying n on posterior

where \( \theta \) represents the probability of heads and where \( f(\theta|\alpha,\beta) \) is defined above and \( k \) is a large integer. What sort of prior belief does this represent?

\( f(\theta|1, 1) \) is essentially a flat prior on \([0, 1]\) whereas \( f(\theta, k) \) has a density which tightens around \( 1/2 \) as \( k \to \infty \). In essence this prior for large \( k \) essentially says with probability \( \gamma \) the coin will be fair and with probability \( (1 - \gamma) \) heads could form any proportion of the total tosses.

If \( \gamma = 1/2 \) and the number of head observed \( x = 0 \), show that the posterior distribution of \( \theta \) is also a mixture of two Beta distributions with posterior weights \( \gamma^* \) and \( (1 - \gamma^*) \) where \( \gamma^* = (1 + t(n))^{-1} \) and \( t(n) = (n + 1)^{-1}[n + 2k - 1](n + 2k - 1)\ldots(n + k)]/[2k - 1)(2k - 1)\ldots(k)]^{-1} \). Show that for \( k \geq 2 \) then \( \gamma^* \to 0 \) as \( n \to \infty \) and interpret this result.
We have $\gamma^* = (1 + t)^{-1}$ therefore,

$$t = (1 - \gamma^*)/\gamma^*$$

$$= (n + 1)^{-1} \frac{(2k + n - 1)! (k - 1)!}{(k + n - 1)! (2k - 1)!}$$

$$= (n + 1)^{-1} \frac{[(2k + n - 1)(2k - 1)] \ldots (k + n)]}{[(2k - 1)(2k - 1) \ldots (k)]]}$$

$$= (n + 1)^{-1} \left[ \left( \frac{n}{2k - 1} \right) + 1 \right] \left[ \left( \frac{n}{2k - 2} \right) + 1 \right] \ldots \left( \frac{n}{k} \right) + 1 \right]$$

$$= (n + 1)^{-1} \prod_{r=2}^{2k-1} \left(1 + \frac{n}{r} \right) \to \infty \text{ as } n \to \infty$$

Hence $\gamma^* \to 0$ as $n \to \infty$ and as such as the sequence of tails becomes larger more weight is given to the coin being unfair.

Figure 17 illustrates that for even a moderate sequence of tails the contribution to the posterior of the informative prior ($f_2(\theta)$) is greatly diminished leaving only a slight ‘hump’ in the density. Note that considering the
posterior $f_2(\theta|x = 0)$ given the prior $f_2(\theta)$ we see that it moves only slightly and has no density for $\theta = 0$.

**Figure 17:** Ex8 Q3 Illustration of posterior for large n and large k

Investigating the effect on the posterior of an increase in the parameter $k$ of the informative prior $f_2(\theta) = f(\theta | k, k)$ in Figure 18 we see that it has the effect of changing the density and emphasising the ‘hump.’ As way of explaining this phenomenon note that as $k$ increases the posterior $f_2(\theta|x = 0)$ given the prior $f_2(\theta)$ becomes more concentrated around a mode which moves progressively more modestly.

**Figure 18:** Ex8 Q3 Comparison of posterior for varying k

Figure 19 illustrates the changes in the posterior whenever the sequence of tails becomes larger. We find that density is drawn from the ‘hump’ caused by the informative prior and resulting in it becoming closer to that of the posterior $f_1(\theta|x = 0)$ given the prior $f_1(\theta)$ as indicated in the results shown in (132).
Figure 19: Ex8 Q3 Comparison of posterior for varying n

References