# Week 5 Exercise Questions: Maximum Likelihood Estimation and Logistic Regression

March 16, 2023

1. (2021 AI2 Exam question) As a data scientist in a telecommunication company, your task is to analyse a customer dataset to predict whether a customer will terminate his/her contract. The dataset consists of around 8000 customer records, each consisting of one binary dependent variable $Y$, indicating whether the customer terminates the contract ($Y = 1$) or not ($Y = 0$), and 19 independent variables, which include the customer's information, e.g., age, subscription plan, extra data plan, etc., and the consumer behaviour such as average numbers of calls and hours per week. Since your boss needs some actionable insights to retain customers, you decided to use interpretable machine learning methods. Design your interpretable machine learning method by answering the following questions:

   (a) You have implemented a feature selection algorithm based on mutual information to select the most informative features from the 19 independent variables. To validate the implementation of your mutual information calculation function, you use a small subset of the data to calculate mutual information manually. You select one independent variable 'subscription plan', denoted as $S$, which takes two values, $S \in \{1, 2\}$. Please use the following Probability Mass Function table

| $p(S, Y)$ | $S = 1$ | $S = 2$ |
|---|---|---|
| $Y = 0$ | $\frac{2}{12}$ | $\frac{5}{12}$ |
| $Y = 1$ | $\frac{2}{12}$ | $\frac{3}{12}$ |

   to calculate

   - Entropies $H(S)$ and $H(Y)$
   - Conditional entropies $H(S|Y)$ and $H(Y|S)$
   - Joint entropy $H(S, Y)$
   - Mutual information $I(S; Y)$

   Show all your working. Discuss what mutual information means and whether this feature will be selected or not. (**6 marks**)

   (b) After applying your algorithm you selected two variables: 1) extra data plan $E$, which is a binary random variable that indicates whether the customer subscribes to the extra data plan ($E = 1$) or not ($E = 0$); and 2) averaged hours used per week $H$, which is a continuous random variable. You then built a logistic regression model

to classify customers into 'low risk' or 'high risk' of terminating the contract. The fitted model is

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

- Given a customer who has the extra data plan ($E = 1$) and spent on average 0.5 hours per week, calculate the odds and the probability the customer will terminate the contract ($Y = 1$). (**4 marks**)
- Using this fitted model, explain to your boss what actions should be taken to retain customers. (**10 marks**)

2. We flipped a coin 100 times. Given that there were 55 heads, use the maximum likelihood estimation to find for the probability $p$ of heads on a single toss.

3. Let $X$ be independent and identically distributed (i.i.d.) Poisson ($\lambda$) distributed. Find the maximum likelihood estimator for $\lambda$, $\hat{\lambda}$. Calculate an estimate using this estimator when, $x_1 = 1$, $x_2 = 2$, $x_3 = 5$, $x_4 = 3$.

# Solutions to Week 5 Exercise Questions: Maximum Likelihood Estimation and Logistic Regression

March 16, 2023

1. (2021 AI2 Exam question) As a data scientist in a telecommunication company, your task is to analyse a customer dataset to predict whether a customer will terminate his/her contract. The dataset consists of around 8000 customer records, each consisting of one binary dependent variable $Y$, indicating whether the customer terminates the contract ($Y = 1$) or not ($Y = 0$), and 19 independent variables, which include the customer's information, e.g., age, subscription plan, extra data plan, etc., and the consumer behaviour such as average numbers of calls and hours per week. Since your boss needs some actionable insights to retain customers, you decided to use interpretable machine learning methods. Design your interpretable machine learning method by answering the following questions:

   (a) You have implemented a feature selection algorithm based on mutual information to select the most informative features from the 19 independent variables. To validate the implementation of your mutual information calculation function, you use a small subset of the data to calculate mutual information manually. You select one independent variable 'subscription plan', denoted as $S$, which takes two values, $S \in \{1, 2\}$. Please use the following Probability Mass Function table

| $p(S,Y)$ | $S=1$ | $S=2$ |
|---|---|---|
| $Y=0$ | $\frac{2}{12}$ | $\frac{5}{12}$ |
| $Y=1$ | $\frac{2}{12}$ | $\frac{3}{12}$ |

   to calculate

   - Entropies $H(S)$ and $H(Y)$
   - Conditional entropies $H(S|Y)$ and $H(Y|S)$
   - Joint entropy $H(S, Y)$
   - Mutual information $I(S; Y)$

   Show all your working. Discuss what mutual information means and whether this feature will be selected or not.

   (b) After applying your algorithm you selected two variables: 1) extra data plan $E$, which is a binary random variable that indicates whether the customer subscribes to the extra data plan ($E = 1$) or

not ($E = 0$); and 2) averaged hours used per week $H$, which is a continuous random variable. You then built a logistic regression model to classify customers into 'low risk' or 'high risk' of terminating the contract. The fitted model is

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

- Given a customer who has the extra data plan ($E = 1$) and spent on average 0.5 hours per week, calculate the odds and the probability the customer will terminate the contract ($Y = 1$). (**4 marks**)
- Using this fitted model, explain to your boss what actions should be taken to retain customers. (**10 marks**)

**Answer:**

(c) Question a: **We will solve this next week.**

Question b: **First question**:

- Odds:

$$\begin{aligned} odds &= \frac{p}{1-p} = \exp(-0.77 + 0.23H - 1.18E) = \exp(-0.77 + 0.115 - 1.18) \\ &= 0.1596 \end{aligned} \tag{1}$$

- Probability: (Note: $p = \frac{1}{1+\frac{1}{odds}}$)

$$P(Y = 1) = \frac{1}{1 + \frac{1}{0.1596}} = 0.137 \tag{2}$$

**Second question**: We need to extend the logistic model with one independent variable as learned in the lecture to two independent variables. We then investigate the effect or each variable by deriving the odd ratios by fixing the value of the other variable:

$$OR_E = \frac{\text{odds when } E = 1}{\text{odds when } E = 0} = \frac{\exp(-0.77 + 0.23 - 1.18)}{\exp(-0.77 + 0.23)} = \exp(-1.18) \approx 0.31$$

$$\begin{aligned} OR_H &= \frac{\text{odds when } H = h + \Delta}{\text{odds when } H = h} && (3) \\ &= \frac{\exp(-0.77 + 0.23(h + \Delta) - 1.18E)}{\exp(-0.77 + 0.23h - 1.18E))} && (4) \\ &= \frac{\exp(-0.77)\exp(0.23h)\exp(0.23\Delta)\exp(-1.18E)}{\exp(-0.77)\exp(0.23h)\exp(-1.18E)))} && (5) \\ &= (\exp(0.23))^{\Delta} && (6) \\ &\approx 1.25^{\Delta} && (7) \end{aligned}$$

The following key points should be mentioned:

- If a customer add the extra data plan, the odds of terminating the contract will increase is 0.31, which means that odds the customer terminating the contract will decrease by a factor of 3.
- If a customer increase the average time by one hour, the odds of terminating the contract increase by a factor of 1.25.

From the analysis, we can suggest to the boss that, the more hours the customers spent, the more likely the customers will terminate, which means the company should improve its telecommunication service/price. However, by simply persuade them to subscribe to the extra data plan, they are more likely to stay.

2. We flipped a coin 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability $p$ of heads on a single toss.

**Answer:**

**Step 1** Write down the likelihood function. First, we need to determine the probability distribution function. Since counting the number of heads in 100 tosses is an experiment with 100 trials, therefore, we should use binomial distribution. The likelihood function, or the probability of 55 heads given that the probability of heads on a single toss is $p$, can be written:

$$P(55 \text{ heads}|p) = \binom{100}{55} p^{55}(1-p)^{45}$$

Explanation:

- **Experiments**: flip the coin 100 times and count the number of heads
- **Data**: The data is the result of the experiments, i.e., '55 heads'
- **Parameter(s)**: we want to estimate the value of the unknown parameter $p$
- **Likelihood or likelihood function**: $P(\text{data} \mid \text{p})$
- $\binom{100}{55}$: the binomial coefficient, which is called "Combination" and can be calculated using

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

**Step 2** Write down the log likelihood function:

$$\log(P(55 \text{ heads}|p)) = \log\left(\binom{100}{55} p^{55}(1-p)^{45}\right) \tag{8}$$

$$= \log\left(\binom{100}{55}\right) + 55\log(p) + 45\log(1-p) \tag{9}$$

$$\tag{10}$$

**Step 3** Take the derivative of the log likelihood function with respect to the unknown parameter $p$ and set it to 0

$$\frac{d}{dp}(\text{log likelihood function}) = \frac{d}{dp}\left[\log\left(\binom{100}{55}\right) + 55\log(p) + 45\ln(1-p)\right] \tag{11}$$

$$= \frac{55}{p} - \frac{45}{1-p} = 0 \tag{12}$$

$$\Rightarrow 55(1-p) = 45p \Rightarrow \hat{p} = 0.55$$

3. Let $X$ be independent and identically distributed (i.i.d.) Poisson ($\lambda$) distributed. Find the maximum likelihood estimator for $\lambda$, $\hat{\lambda}$. Calculate an estimate using this estimator when $x_1 = 1$, $x_2 = 2$, $x_3 = 5$, $x_4 = 3$.

   **Answer** We can calculate the maximum likelihood estimator for $\lambda$, $\hat{\lambda}$ as follows

**Step 1** Poisson PMF: We first write the probability density function of the Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

**Step 2** Write the likelihood function:

$$L(\lambda|x_1, \ldots, x_n) = \prod_{j=1}^{n} \frac{\lambda^{x_j} e^{-\lambda}}{x_j!}$$

**Step 3** By taking the natural logarithm on both sides, we have

$$\log L(\lambda|\boldsymbol{x}) = \log\left(\prod_{j=1}^{n} \frac{\lambda^{x_j} e^{-\lambda}}{x_j!}\right) \tag{13}$$

$$= \sum_{j=1}^{n} \log\left(\frac{\lambda^{x_j} e^{-\lambda}}{x_j!}\right) \tag{14}$$

$$= \sum_{j=1}^{n} \left[\log(\lambda^{x_j} e^{-\lambda}) - \log(x_j!)\right] \tag{15}$$

$$= \sum_{j=1}^{n} \left[\log(\lambda^{x_j}) + \log(e^{-\lambda}) - \log(x_j!)\right] \tag{16}$$

$$= \sum_{j=1}^{n} \left[x_j \log(\lambda) - \lambda - \log(x_j!)\right] \tag{17}$$

$$= -n\lambda + \log(\lambda)\sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \log(x_j!) \tag{18}$$

**Step 4** Calculate the derivative of the log likelihood function with respect to

$\lambda$:

$$\frac{d}{d\lambda}\log L(\lambda|\boldsymbol{x}) = \frac{d}{d\lambda}\left(-n\lambda + \log(\lambda)\sum_{j=1}^{n} x_j - \sum_{j=1}^{n}\log(x_j!)\right) \tag{19}$$

$$= -n + \frac{1}{\lambda}\sum_{j=1}^{n} x_j \tag{20}$$

**Step 5** Set the derivative $\frac{d}{d\lambda}\log L(\lambda|\boldsymbol{x}) = 0$

$$-n + \frac{1}{\lambda}\sum_{j=1}^{n} x_j = 0 \tag{21}$$

which yield

$$\lambda = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{22}$$

For the given data, the estimate $\hat{\lambda}$ is

$$\hat{\lambda} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{4}(1+2+5+3) = 2.75$$

UNIVERSITY OF BIRMINGHAM

**School of Computer Science**

Week 8 Exercise Questions

**Artificial Intelligence 2**

2023

Artificial Intelligence 2

# Exam paper

## Question 1 (2021 AI2 Exam question)

As a data scientist in a telecommunication company, your task is to analyse a customer dataset to predict whether a customer will terminate his/her contract. The dataset consists of around 8000 customer records, each consisting of one binary dependent variable $Y$, indicating whether the customer terminates the contract ($Y = 1$) or not ($Y = 0$), and 19 independent variables, which include the customer's information, e.g., age, subscription plan, extra data plan, etc., and the consumer behaviour such as average numbers of calls and hours per week. Since your boss needs some actionable insights to retain customers, you decided to use interpretable machine learning methods. Design your interpretable machine learning method by answering the following questions:

(a) You have implemented a feature selection algorithm based on mutual information to select the most informative features from the 19 independent variables. To validate the implementation of your mutual information calculation function, you use a small subset of the data to calculate mutual information manually. You select one independent variable 'subscription plan', denoted as $S$, which takes two values, $S \in \{1, 2\}$. Please use the following Probability Mass Function table

| $p(S, Y)$ | $S = 1$ | $S = 2$ |
|---|---|---|
| $Y = 0$ | $\frac{2}{12}$ | $\frac{5}{12}$ |
| $Y = 1$ | $\frac{2}{12}$ | $\frac{3}{12}$ |

to calculate

- Entropies $H(S)$ and $H(Y)$
- Conditional entropies $H(S|Y)$ and $H(Y|S)$
- Joint entropy $H(S, Y)$
- Mutual information $I(S; Y)$

Show all your working. Discuss what mutual information means and whether this feature will be selected or not.

(b) After applying your algorithm you selected two variables: 1) extra data plan $E$, which is a binary random variable that indicates whether the customer subscribes to the extra data plan ($E = 1$) or not ($E = 0$); and 2) averaged hours used per week $H$, which is a continuous random variable. You then built a logistic regression model

to classify customers into 'low risk' or 'high risk' of terminating the contract. The fitted model is

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

- Given a customer who has the extra data plan ($E = 1$) and spent on average 0.5 hours per week, calculate the odds and the probability the customer will terminate the contract ($Y = 1$). (**4 marks**)
- Using this fitted model, explain to your boss what actions should be taken to retain customers. (**10 marks**)

**0 marks for question not valid - all questions must have the same number of marks**

## Question 2 (2022 AI2 Exam question)

As a machine learning expert for an AI cyber security company, your task is to design an automated network intrusion detection system. You have collected a large number of records of network activities. Each record includes the log information about network activity, such as protocol types, duration, number of failed logins, which are random variables, denoted as $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^\intercal$. Each record also includes a binary random variable $Y$ called label that was labelled by cyber security experts as intrusions ($Y = 1$) or normal connections ($Y = 0$).

(a) Consider feature selection based on mutual information to reduce the number of independent variables.

(i) Explain to your colleague, who knows nothing about information theory, the concept of mutual information. **[2 marks]**

(ii) Explain the loop, i.e., lines 4-7 of the pseudocode in Table 1. Note $I(Y; X_i)$ is the mutual information between $Y$ and $X_i$.

1: **Initialisation:** Set $F \leftarrow \boldsymbol{X}$ and $S \leftarrow \emptyset$
2: $f_{\max} = \text{argmax}_{X_i \in \boldsymbol{X}} I(Y; X_i)$
3: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max}$
4: **Repeat** until $|S| = K$:
5: $f_{\max} = \text{argmax}_{X_i \in F} I(Y; X_i) - \beta \sum_{X_s \in S} I(X_s; X_i)$
6: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max} \cup S$
7: **End**
8: **End**

Table 1: Pseudocode of Mutual Information based Feature Selection Algorithm.

**[3 marks]**

(b) After applying your feature selection algorithm, assume you selected four random variables as features, denoted as $F_1, F_2, F_3, F_4$. Based on these features, you now work with a cyber security expert to construct a Bayesian network to harness the domain knowledge of cyber security. The expert first divides intrusions into three cyber attacks, $A_1, A_2, A_3$, which are marginally independent from each other. The expert suggests the presence of the four features are used to find the most probable type of cyber attacks. The four features are conditionally dependent on the three types cyber attacks as follows: $F_1$ depends only on $A_1$, $F_2$ depends on $A_1$ and $A_2$. $F_3$ depends on $A_1$ and $A_3$, whereas $F_4$ depends only on $A_3$. We assume all these random variables are binary, i.e., they are either 1 (true) or 0 (false).

(i) Draw the Bayesian network according to the expert?s description. **[2 marks]**

(ii) Write down the joint probability distribution represented by this Bayesian network. **[3 marks]**

(iii) How many parameters are required to describe this joint probability distribution? Show your working. **[5 marks]**

(iv) Suppose in a record we observe $F_2$ is true, what does observing $F_4$ is true tell us? If we observe $F_3$ is true instead of $F_2$, what does observing $F_4$ is true tell us? **[5 marks]**

# Artificial Intelligence 2

# Solutions

2023

Artificial Intelligence 2

# Exam paper

## Question 1 (2021 AI2 Exam question)

As a data scientist in a telecommunication company, your task is to analyse a customer dataset to predict whether a customer will terminate his/her contract. The dataset consists of around 8000 customer records, each consisting of one binary dependent variable $Y$, indicating whether the customer terminates the contract ($Y = 1$) or not ($Y = 0$), and 19 independent variables, which include the customer's information, e.g., age, subscription plan, extra data plan, etc., and the consumer behaviour such as average numbers of calls and hours per week. Since your boss needs some actionable insights to retain customers, you decided to use interpretable machine learning methods. Design your interpretable machine learning method by answering the following questions:

(a) You have implemented a feature selection algorithm based on mutual information to select the most informative features from the 19 independent variables. To validate the implementation of your mutual information calculation function, you use a small subset of the data to calculate mutual information manually. You select one independent variable 'subscription plan', denoted as $S$, which takes two values, $S \in \{1, 2\}$. Please use the following Probability Mass Function table

| $p(S, Y)$ | $S = 1$ | $S = 2$ |
|---|---|---|
| $Y = 0$ | $\frac{2}{12}$ | $\frac{5}{12}$ |
| $Y = 1$ | $\frac{2}{12}$ | $\frac{3}{12}$ |

to calculate

- Entropies $H(S)$ and $H(Y)$
- Conditional entropies $H(S|Y)$ and $H(Y|S)$
- Joint entropy $H(S, Y)$
- Mutual information $I(S; Y)$

Show all your working. Discuss what mutual information means and whether this feature will be selected or not.

(b) After applying your algorithm you selected two variables: 1) extra data plan $E$, which is a binary random variable that indicates whether the customer subscribes to the extra data plan ($E = 1$) or not ($E = 0$); and 2) averaged hours used per week $H$, which is a continuous random variable. You then built a logistic regression model

to classify customers into 'low risk' or 'high risk' of terminating the contract. The fitted model is

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

- Given a customer who has the extra data plan ($E = 1$) and spent on average 0.5 hours per week, calculate the odds and the probability the customer will terminate the contract ($Y = 1$). (**4 marks**)
- Using this fitted model, explain to your boss what actions should be taken to retain customers. (**10 marks**)

**0 marks for question not valid - all questions must have the same number of marks**

**Model answer / LOs / Creativity:**

(a)

| $p(S,Y)$ | $S=1$ | $S=2$ | $p(Y)$ |
|---|---|---|---|
| $Y=0$ | $\frac{2}{12}$ | $\frac{5}{12}$ | $\frac{7}{12}$ |
| $Y=1$ | $\frac{2}{12}$ | $\frac{3}{12}$ | $\frac{5}{12}$ |
| $p(S)$ | $\frac{4}{12}$ | $\frac{8}{12}$ | 1 |

$$H(S) = -\sum_{i}^{n} p(s_i)\log_2 p(s_i) = -(\frac{4}{12}\log\frac{4}{12} + \frac{8}{12}\log\frac{8}{12}) = 0.92 \text{ bits}$$

$$H(Y) = -\sum_{i}^{n} p(y_i)\log_2 p(y_i) = -(\frac{7}{12}\log\frac{7}{12} + \frac{5}{12}\log\frac{5}{12}) = 0.97 \text{ bits}$$

$$H(S,Y) = -\sum_{x_i \in R_X}\sum_{y_j \in R_Y} p(s_i, y_j)\log p(s_i, y_j) = 1.89 \text{ bits}$$

$$H(S|Y) = H(S,Y) - H(Y) = 1.89 - 0.97 = 0.92 \text{ bits}$$

$$H(S|X) = H(S,Y) - H(S) = 1.89 - 0.92 = 0.97 \text{ bits}$$

$$I(S;Y) = H(S) - H(S|Y) = 0.92 - 0.92 = 0 \text{ bits}$$

Since mutual information measures the information that two random variables $S$ and $Y$ share. In other words, it measures how much knowing one of these variables reduces uncertainty about the other, the value of 0 means this feature is not useful at all.

(b) **First question**:

- Odds:

$$odds = \frac{p}{1-p} = \exp(-0.77 + 0.23H - 1.18E) = \exp(-0.77 + 0.115 - 1.18)$$
$$= 0.1596 \quad (1)$$

- Probability:

$$P(Y = 1) = 0.137 \quad (2)$$

**Second question**: The student should extend the logistic model with one independent variable as learned in the lecture to two independent variables. The student should be able to investigate the effect or each variable by deriving the odd ratios by fixing the value of the other variable:

$$OR_E = \frac{\text{odds when } E = 1}{\text{odds when } E = 0} = \frac{\exp(-0.77 + 0.23 - 1.18)}{\exp(-0.77 + 0.23)} = \exp(-1.18) \approx 0.31$$

$$OR_H = \frac{\text{odds when } H = h + \Delta}{\text{odds when } H = h} = \frac{\exp(-0.77 + 0.23(H + \Delta) - 1.18E)}{\exp(-0.77 + 0.24H - -1.18E))} = (\exp(0.23))^{\Delta} \approx 1.25^{\Delta}$$

The following key points should be mentioned:

- If a customer add the extra data plan, the odds of terminating the contract will increase is 0.3, which means that odds the customer terminating the contract will decrease by a factor of 3.
- If a customer increase the average time by one hour, the odds of terminating the contract increase by a factor of 1.25.

From the analysis, we can suggest to the boss that, the more hours the customers spent, the more likely the customers will terminate, which means the company should improve its telecommunication service/price. However, by simply persuade them to subscribe to the extra data plan, they are more likely to stay.

## Question 2 (2022 AI2 Exam question)

As a machine learning expert for an AI cyber security company, your task is to design an automated network intrusion detection system. You have collected a large number of records of network activities. Each record includes the log information about network activity, such as protocol types, duration, number of failed logins, which are random variables, denoted as $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^\intercal$. Each record also includes a binary random variable $Y$ called label that was labelled by cyber security experts as intrusions ($Y = 1$) or normal connections ($Y = 0$).

(a) Consider feature selection based on mutual information to reduce the number of independent variables.

(i) Explain to your colleague, who knows nothing about information theory, the concept of mutual information. **[2 marks]**

(ii) Explain the loop, i.e., lines 4-7 of the pseudocode in Table 1. Note $I(Y; X_i)$ is the mutual information between $Y$ and $X_i$.

1: **Initialisation:** Set $F \leftarrow \boldsymbol{X}$ and $S \leftarrow \emptyset$
2: $f_{\max} = \text{argmax}_{X_i \in \boldsymbol{X}} I(Y; X_i)$
3: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max}$
4: **Repeat** until $|S| = K$:
5: $f_{\max} = \text{argmax}_{X_i \in F} I(Y; X_i) - \beta \sum_{X_s \in S} I(X_s; X_i)$
6: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max} \cup S$
7: **End**
8: **End**

Table 1: Pseudocode of Mutual Information based Feature Selection Algorithm.

**[3 marks]**

(b) After applying your feature selection algorithm, assume you selected four random variables as features, denoted as $F_1, F_2, F_3, F_4$. Based on these features, you now work with a cyber security expert to construct a Bayesian network to harness the domain knowledge of cyber security. The expert first divides intrusions into three cyber attacks, $A_1, A_2, A_3$, which are marginally independent from each other. The expert suggests the presence of the four features are used to find the most probable type of cyber attacks. The four features are conditionally dependent on the three types cyber attacks as follows: $F_1$ depends only on $A_1$, $F_2$ depends on $A_1$ and $A_2$. $F_3$ depends on $A_1$ and $A_3$, whereas $F_4$ depends only on $A_3$. We assume all these random variables are binary, i.e., they are either 1 (true) or 0 (false).

(i) Draw the Bayesian network according to the expert?s description. **[2 marks]**

(ii) Write down the joint probability distribution represented by this Bayesian network. **[3 marks]**

(iii) How many parameters are required to describe this joint probability distribution? Show your working. **[5 marks]**

(iv) Suppose in a record we observe $F_2$ is true, what does observing $F_4$ is true tell us? If we observe $F_3$ is true instead of $F_2$, what does observing $F_4$ is true tell us? **[5 marks]**

**Model answer / LOs / Creativity:**

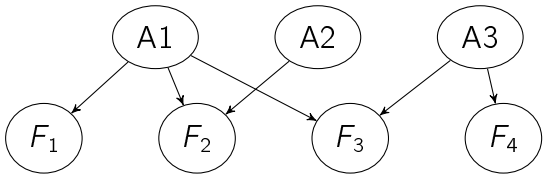
(a) Mutual Information Feature selection

(i) Basic motivation behind mutual information is to measure the information that two random variables $X$ and $Y$ share. In other words, it measures how much knowing one of these variables reduces uncertainty about the other.

(ii) The student should explain the two lines similar to the following: "The two lines in the loop are used to select $K$ features. In this loop, we also find the feature $f_{\max}$ which achieves the maximum mutual information $I$ among all the remaining independent variables in set $F$. However, because some features highly correlated with each other, selecting them will increase the number of features but does not improve the prediction. Therefore, we need to make sure there must be minimal redundancy between the candidate feature $X_i$ and the set of selected features $S$. That's exactly the second term of the equation ( $f_{\max}$) You then add this feature into $S$ and then subtract it from set $F$ and repeat until we got $K$ features".

(b) Bayesian Networks

(i) The following draw is acceptable:




(ii) The joint probability distribution represented by the Bayesian network:

$$\begin{aligned} &P(A_1, A_2, A_3, F_1, F_2, F_3, F_4) \\ &= P(A_1)P(A_2)P(A_3)P(F_1|A_1)P(F_2|A_1, A_2)P(F_3|A_1, A_3)P(F_4|A_3) \end{aligned} \tag{3}$$

(iii) Total number of parameter is 15, but the student should show the following working:

| Conditional Probability | number of parameters |
|---|---|
| $P(A_1)$ | 1 |
| $P(A_2)$ | 1 |
| $P(A_3)$ | 1 |
| $P(F_1\|A_1)$ | 2 |
| $P(F_2\|A_1, A_2)$ | 4 |
| $P(F_3\|A_1, A_3)$ | 4 |
| $P(F_4\|A_3)$ | 2 |

(iv) With $F_2 = 1$, observing $F_4 = 1$ still gives us information only about $A_3$. If we observe $F_3 = 1$ instead of $F_2$, then observing $F_4 = 1$ will give us information about $A_1$ and $A_3$ due to competing causes (aka. explaining away).

UNIVERSITY OF BIRMINGHAM

**School of Computer Science**

Week 9 Exercise Questions

**Artificial Intelligence 2**

2023

Artificial Intelligence 2

# Exercise questions Week 9

## Question 1

As a developer of a security equipment company, you are going to design an alarm that senses when an infra-red sensor gauge exceeds a given threshold. The infra-red sensor measures the infra-red temperature and the gauge measures the infra-red temperature obtained from the infra-red sensor. Consider the Boolean variables $A$ (alarm sounds), $F_a$ (alarm is faulty), $F_g$ (gauge is faulty) and the $G$ (gauge reading: normal and high) and $T$ (actual infra-red temperature: normal and high).

(a) Draw a Bayesian network for this problem.

(b) Write down the joint probability distribution represented by this Bayesian network.

(c) How many parameters are required to describe this joint probability distribution? Show your working.

**0 marks for question not valid - all questions must have the same number of marks**

## Question 2 (2022 AI2 Exam question)

As a machine learning expert for an AI cyber security company, your task is to design an automated network intrusion detection system. You have collected a large number of records of network activities. Each record includes the log information about network activity, such as protocol types, duration, number of failed logins, which are random variables, denoted as $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^\intercal$. Each record also includes a binary random variable $Y$ called label that was labelled by cyber security experts as intrusions ($Y = 1$) or normal connections ($Y = 0$).

(a) Consider feature selection based on mutual information to reduce the number of independent variables.

(i) Explain to your colleague, who knows nothing about information theory, the concept of mutual information. **[2 marks]**

(ii) Explain the loop, i.e., lines 4-7 of the pseudocode in Table 1. Note $I(Y; X_i)$ is the mutual information between $Y$ and $X_i$.

**[3 marks]**

1: **Initialisation:** Set $F \leftarrow \boldsymbol{X}$ and $S \leftarrow \emptyset$
2: $f_{\max} = \text{argmax}_{X_i \in \boldsymbol{X}} I(Y; X_i)$
3: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max}$
4: **Repeat** until $|S| = K$:
5: $f_{\max} = \text{argmax}_{X_i \in F} I(Y; X_i) - \beta \sum_{X_s \in S} I(X_s; X_i)$
6: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max} \cup S$
7: **End**
8: **End**

Table 1: Pseudocode of Mutual Information based Feature Selection Algorithm.

(b) After applying your feature selection algorithm, assume you selected four random variables as features, denoted as $F_1, F_2, F_3, F_4$. Based on these features, you now work with a cyber security expert to construct a Bayesian network to harness the domain knowledge of cyber security. The expert first divides intrusions into three cyber attacks, $A_1, A_2, A_3$, which are marginally independent from each other. The expert suggests the presence of the four features are used to find the most probable type of cyber attacks. The four features are conditionally dependent on the three types cyber attacks as follows: $F_1$ depends only on $A_1$, $F_2$ depends on $A_1$ and $A_2$. $F_3$ depends on $A_1$ and $A_3$, whereas $F_4$ depends only on $A_3$. We assume all these random variables are binary, i.e., they are either 1 (true) or 0 (false).

(i) Draw the Bayesian network according to the expert?s description. **[2 marks]**

(ii) Write down the joint probability distribution represented by this Bayesian network. **[3 marks]**

(iii) How many parameters are required to describe this joint probability distribution? Show your working. **[5 marks]**

(iv) Suppose in a record we observe $F_2$ is true, what does observing $F_4$ is true tell us? If we observe $F_3$ is true instead of $F_2$, what does observing $F_4$ is true tell us? **[5 marks]**

# Artificial Intelligence 2

# Solutions

2023

Artificial Intelligence 2

# Exercise questions Week 9

## Question 1

As a developer of a security equipment company, you are going to design an alarm that senses when an infra-red sensor gauge exceeds a given threshold. The infra-red sensor measures the infra-red temperature and the gauge measures the infra-red temperature obtained from the infra-red sensor. Consider the Boolean variables $A$ (alarm sounds), $F_a$ (alarm is faulty), $F_g$ (gauge is faulty) and the $G$ (gauge reading: normal and high) and $T$ (actual infra-red temperature: normal and high).

(a) Draw a Bayesian network for this problem.

(b) Write down the joint probability distribution represented by this Bayesian network.

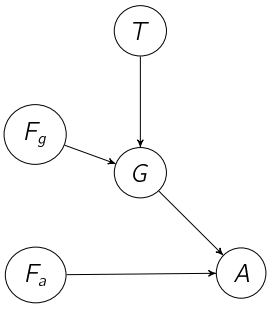
(c) How many parameters are required to describe this joint probability distribution? Show your working.

**0 marks for question not valid - all questions must have the same number of marks**

**Model answer / LOs / Creativity:**

(a) Draw a Bayesian network for this problem.

**Answer**:




(b) Write down the joint probability distribution represented by this Bayesian network.

**Answer**:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \quad (1)$$

$$= P(T)P(F_g)P(G|F_g, T)P(F_a)P(A|G, F_a) \quad (2)$$

(c) How many parameters are required to describe this joint probability distribution? Show your working.

**Answer**: The total number of parameters is

| Conditional Probability | number of parameters |
|---|---|
| $P(T)$ | 1 |
| $P(F_g)$ | 1 |
| $P(G\|T, F_g)$ | 4 |
| $P(F_a)$ | 1 |
| $P(A\|F_a, G)$ | 4 |

So the total number of parameters is

$$1 + 1 + 4 + 1 + 4 = 11$$

## Question 2 (2022 AI2 Exam question)

As a machine learning expert for an AI cyber security company, your task is to design an automated network intrusion detection system. You have collected a large number of records of network activities. Each record includes the log information about network activity, such as protocol types, duration, number of failed logins, which are random variables, denoted as $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^\intercal$. Each record also includes a binary random variable $Y$ called label that was labelled by cyber security experts as intrusions ($Y = 1$) or normal connections ($Y = 0$).

(a) Consider feature selection based on mutual information to reduce the number of independent variables.

(i) Explain to your colleague, who knows nothing about information theory, the concept of mutual information. **[2 marks]**

(ii) Explain the loop, i.e., lines 4-7 of the pseudocode in Table 1. Note $I(Y; X_i)$ is the mutual information between $Y$ and $X_i$.

1: **Initialisation:** Set $F \leftarrow \boldsymbol{X}$ and $S \leftarrow \emptyset$
2: $f_{\max} = \text{argmax}_{X_i \in \boldsymbol{X}} I(Y; X_i)$
3: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max}$
4: **Repeat** until $|S| = K$:
5: $f_{\max} = \text{argmax}_{X_i \in F} I(Y; X_i) - \beta \sum_{X_s \in S} I(X_s; X_i)$
6: Set $F \leftarrow F \setminus \{f_{\max}\}$ and $S \leftarrow f_{\max} \cup S$
7: **End**
8: **End**

Table 1: Pseudocode of Mutual Information based Feature Selection Algorithm.

**[3 marks]**

(b) After applying your feature selection algorithm, assume you selected four random variables as features, denoted as $F_1, F_2, F_3, F_4$. Based on these features, you now work with a cyber security expert to construct a Bayesian network to harness the domain knowledge of cyber security. The expert first divides intrusions into three cyber attacks, $A_1, A_2, A_3$, which are marginally independent from each other. The expert suggests the presence of the four features are used to find the most probable type of cyber attacks. The four features are conditionally dependent on the three types cyber attacks as follows: $F_1$ depends only on $A_1$, $F_2$ depends on $A_1$ and $A_2$. $F_3$ depends on $A_1$ and $A_3$, whereas $F_4$ depends only on $A_3$. We assume all these random variables are binary, i.e., they are either 1 (true) or 0 (false).

(i) Draw the Bayesian network according to the expert?s description. **[2 marks]**

(ii) Write down the joint probability distribution represented by this Bayesian network. **[3 marks]**

(iii) How many parameters are required to describe this joint probability distribution? Show your working. **[5 marks]**

(iv) Suppose in a record we observe $F_2$ is true, what does observing $F_4$ is true tell us? If we observe $F_3$ is true instead of $F_2$, what does observing $F_4$ is true tell us? **[5 marks]**

**Model answer / LOs / Creativity:**

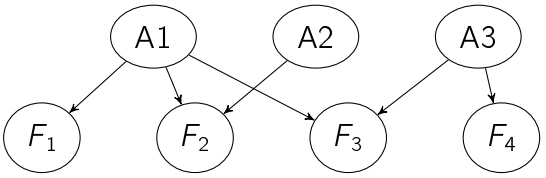
(a) Mutual Information Feature selection

(i) Basic motivation behind mutual information is to measure the information that two random variables $X$ and $Y$ share. In other words, it measures how much knowing one of these variables reduces uncertainty about the other.

(ii) The student should explain the two lines similar to the following: “The two lines in the loop are used to select $K$ features. In this loop, we also find the feature $f_{\max}$ which achieves the maximum mutual information $I$ among all the remaining independent variables in set $F$. However, because some features highly correlated with each other, selecting them will increase the number of features but does not improve the prediction. Therefore, we need to make sure there must be minimal redundancy between the candidate feature $X_i$ and the set of selected features $S$. That's exactly the second term of the equation ( $f_{\max}$) You then add this feature into $S$ and then subtract it from set $F$ and repeat until we got $K$ features".

(b) Bayesian Networks

(i) The following draw is acceptable:




(ii) The joint probability distribution represented by the Bayesian network:

$$\begin{aligned} &P(A_1, A_2, A_3, F_1, F_2, F_3, F_4) \\ &= P(A_1)P(A_2)P(A_3)P(F_1|A_1)P(F_2|A_1, A_2)P(F_3|A_1, A_3)P(F_4|A_3) \end{aligned} \tag{3}$$

(iii) Total number of parameter is 15, but the student should show the following working:

| Conditional Probability | number of parameters |
|---|---|
| $P(A_1)$ | 1 |
| $P(A_2)$ | 1 |
| $P(A_3)$ | 1 |
| $P(F_1\|A_1)$ | 2 |
| $P(F_2\|A_1, A_2)$ | 4 |
| $P(F_3\|A_1, A_3)$ | 4 |
| $P(F_4\|A_3)$ | 2 |

(iv) With $F_2 = 1$, observing $F_4 = 1$ still gives us information only about $A_3$. If we observe $F_3 = 1$ instead of $F_2$, then observing $F_4 = 1$ will give us information about $A_1$ and $A_3$ due to competing causes (aka. explaining away).