UNIVERSITY OF BIRMINGHAM

School of Computer Science

Postgraduate Papers

January Examinations 2021

Contents

30236 LM Advanced Networking (Extended)

Question 1

Ethernet was one of many candidates for local area networking, but is now the dominant technology. There is always some amount of luck and contingency in market success, but Ethernet's dominance is now so total that it is likely there were also sound technical reasons.

(a) One of the technologies that Ethernet succeeded against was IBM Token Ring. How does Ethernet differ in operation from Token Ring? Why do you think that Ethernet proved more successful? Justify your answers. **[4 marks]**

(b) Ethernet has increased in bitrate from 3Mbps through 10, 100 and 1000 Mbps and is now pushing to 10Gbps, 40GBps and soon 100GBps. Explain how switching and full-duplex operation enabled this increase. Your answer should pay close attention to the practicality of collision detection in networks with bitrates in excess of 1Gbps. **[4 marks]**

(c) Ethernet was originally very exposed: a passive observer could read all traffic, and an active attacker could send traffic posing as any other station. Explain how ethernet switches can be used to provide some protection against these attacks. How effective are these mechanisms, and to what extent are they useful? Justify your answer. **[12 marks]**

Question 2

TCP is approaching its fortieth birthday. While modern implementations still interwork with older ones, many additional features have been added. These are largely in response to changes in the functionality and performance of underlying networks, and take advantage of the increased performance of modern computers.

For reference, the TCP header is shown here, taken from RFC793:

```
0                   1                   2                   3
   0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |          Source Port          |       Destination Port        |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                        Sequence Number                        |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                    Acknowledgment Number                      |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |  Data |           |U|A|P|R|S|F|                               |
  | Offset| Reserved  |R|C|S|S|Y|I|            Window             |
  |       |           |G|K|H|T|N|N|                               |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |           Checksum            |         Urgent Pointer        |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                    Options                    |    Padding    |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                             data                              |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

In this question, we are going to design a hypothetical "TCP version 2" which, which being recognisably TCP and therefore leveraging the extensive experience with this protocol, can:

- fully utilise the performance available on high-speed, long-distance networks.
- replace UDP for fast "one question and one answer" protocols such as DNS and SNMP.
- take advantage of modern computers' increased memory and processing power as compared to 16-bit computers of the early 1980s.
- deliver high-quality streaming video and audio, including low-latency video conferencing.

(a) Describe and justify the changes you would make to TCP in order to solve the problems listed above. You can change the header format, the state machine and the typical libraries used to access network facilities. Your answer should include a new packet header format, showing your improvements, as well as some form of diagram to show the flow of packets in the areas you have changed. **[12 marks]**

(b) Explain the ways in which your new version of TCP will operate more effectively on modern networks. You can assume these networks have higher bitrates, lower loss and error rates and similar latency to the networks for which TCP was originally designed. **[8 marks]**

Marks will be awarded based upon:

- How well your proposals cover the four listed issues;
- How technically plausible your solutions are;
- The clarity and precision of your answer.

Question 3

Wireless networks are increasingly common, and in many organisations have almost completely replaced wired networks for client-device access.

(a) Explain the difference between WPA2 PSK (pre-shared key, also called WPA2 Personal) and WPA2 Enterprise. What criteria should one use to inform the choice between these two systems? Why might some companies be tempted to use WPA2 PSK instead of WPA2 Enterprise? **[6 marks]**

(b) You are the network manager of a large shared office building which has many tenants from different organisations, all using a single shared wireless network secured with WPA2 PSK. You have been approached by one of the tenant organisations, who has concerns about the security of the wireless network, given the fact that everyone is using the same pre-shared key. They have asked you the following questions:

(i) Can other tenants within the building read the emails that we send and receive?

(ii) Can other tenants send emails that look like they come from our organisation?

(iii) Can other tenants see what URLs we visit when web browsing? What about the content of the web pages?

(iv) Can other tenants access the shared drive we have in the building (it is meant to be accessible only to members of our organisation).

Answer the questions as best you can. Mention any clarifications you would need to obtain, in order to give more precise answers. **[4 marks]**

(c) The tenant organisation that approached you suggests you should upgrade the whole network to WPA2 Enterprise. You have been asked to respond to the CEO of the tenant organisation, advising her of the best course of action.

Write a memo, of approximately half a page, briefing her on the issues. You should consider:

- What are the problems that using a single shared-key network for multiple enterprises will cause?
- What will need to be done to deploy WPA2 enterprise, and what will the costs be?
- What will be the benefits?
- Someone has suggested having both WPA2 PSK and WPA2 Enterprise at the same time. Could that work?

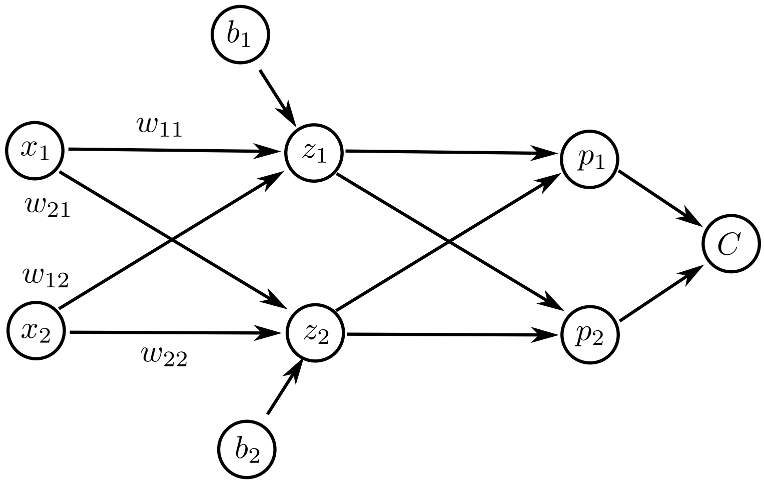
[10 marks]

Question 1 Regression and Model Selection

The following procedure is designed by someone to (given a dataset) find a regression model function for deployment and estimate its error for new datasets in future. However, there may exist some problems in the procedure. Please list all of them and briefly explain why they are a problem. **[20 marks]**

- Step 1. Split the data for training and testing. For all the data, arbitrarily split them into two portions, one major portion (e.g., 80%) for training and the rest (e.g., 20%) for testing.
- Step 2. Evaluate all the models. For each of the considered models (e.g., linear and quadratic), firstly train it on the training data with the regularisation term (e.g., try the regularisation parameter $0, 0.1, 0.2, 0.4, 0.8, ..., 10$); secondly evaluate the trained model (i.e., calculate its error) on the testing data with the corresponding regularisation parameter used in training; and lastly choose the regularisation parameter that leads to the smallest error in the evaluation for the model.
- Step 3: Select the model and return its function and error. Among all the models, the model with the smallest error (denoted by ϵ) in the evaluation is selected. Its function (i.e., the trained model in Step 2), along with the regularisation parameter, is returned for deployment, and the error ϵ is returned as the predicted error for new data in future.

Question 2 Neural Network



Consider the neural network above where the units are defined for $j \in \{1, 2\}$ as follows:

$$
\begin{aligned}
z_j &= w_{j1}x_1 + w_{j2}x_2 + b_j, \\
p_j &= \frac{e^{z_j}}{Q}, \text{ where } Q = e^{z_1} + e^{z_2}.
\end{aligned}
$$

The network is trained using stochastic gradient descent (i.e., minibatch size 1). For a training example with inputs $x_1, x_2 \in \mathbb{R}$ and correct output $y \in \{1, 2\}$, the cost C is defined as

$$C = \log(Q) - z_y.$$

(a) Compute the partial derivatives $\frac{\partial C}{\partial z_1}$ and $\frac{\partial C}{\partial z_2}$ **[7 marks]**

(b) Suppose that the gradient step increases the weight parameter w_{11}, and that $x_1 > x_2 > 0$. What can you say about the correct output y? Justify your answer. **[13 marks]**

Question 3 Principal Component Analysis

Consider a large 5-dimensional data set $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, $\mathbf{x}^i \in \mathbb{R}^5$, $i = 1, \ldots, N$, $N = 100,000,000$. By applying Principal Component Analysis (PCA) to $\mathcal{D}$, one obtains eigenvalues $\lambda_1 = 10, \lambda_2 = 4, \lambda_3 = \lambda_4 = \lambda_5 = 0.05$ of the data covariance matrix $\mathbf{C}$, together with their corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_5 \in \mathbb{R}^5$.

(a) Describe the nature of the data distribution in the 5-dimensional space. How many eigenvectors are needed so that when the data is projected onto the low-dimensional subspace spanned by them only 1% of data variability is lost? **[7 marks]**

(b) Assume now that you were told that the data $\mathcal{D}$ will be transmitted to your friend, but unfortunately the communication channel is unreliable and all elements of the data will be corrupted by an additive independent identically distributed (i.i.d.) Gaussian noise of zero mean and variance $\sigma^2 > 0$.

Your friend will apply PCA on the received data by calculating the data covariance matrix $\mathbf{C}'$ and finding its eigenvectors $\mathbf{u}'_1, \ldots, \mathbf{u}'_5$ and $\lambda'_1, ..., \lambda'_5$. What relation would you expect between

- the covariance matrices $\mathbf{C}$ and $\mathbf{C}'$,
- eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_5$ and $\mathbf{u}'_1, \ldots, \mathbf{u}'_5$,
- eigenvalues $\lambda_1, ..., \lambda_5$ and eigenvalues $\lambda'_1, ..., \lambda'_5$?

Justify your answer. **[13 marks]**

34221 LM Computer Systems

Question 1

(a) You are building a system that records student marks in an in-memory data structure. It is important that the system works efficiently and correctly for multiple concurrent users. There are 2 classes of operation: recording a mark for a student and generating a report of a student's marks. Whenever an operation takes place an audit record is kept of that operation. This is done by appending a record (user, date-time, operation performed) to a global string that records all of the operations that have been performed on the database. You can ignore practical issues e.g. of data persistence or memory use and can assume that reading/writing a mark is an atomic operation.

Explain what issues could arise from having multiple concurrent users of this system. You should propose a solution to these potential problems and provide an argument for why you have chosen this solution rather than other alternatives. Explain what the implications are of implementing your proposal. **[8 marks]**

(b) You have been given the following algorithm:

```
for i=1 to n
   for j=1 to n/2
      for k=1 to n/3
         a=a+i*j*k
```

(i) Derive and justify its time complexity using 'Big-O' notation

(ii) You have been given the following execution times for a program which implements this algorithm. The program has two parts. The first is constant and independent of n – a fixed overhead, the second implements the algorithm above.

N	250	500
Execution time	110 msec	180 msec

Estimate and justify how long it would take to execute for n=1000

[6 marks]

(c) You need to represent data for up to 40,000 people. For each person you need to record:

- Their name (100 bytes)

- Their date of birth (10 bytes)
- Their person id (8 bytes)
- An image (64k bytes)
- A string which records details of any additional information (2k bytes)

Estimate how much space you will need to allow to record this information. Express your result as the number of Mbytes. How many bits would be required to address this many bytes? Explain your working.

[6 marks]

Question 2

(a) Translate the following Java code into MIPS assembly code. Briefly, explain each line of your code.

```
for (i=0;i<10;i++) vals[i]=i*i;
```

[3 marks]

(b) (i) A program makes three system calls: one to open a file; one to read a byte from it; and then one to close the file. How many times will the mode bit change from 0 to 1 during this process? Justify your answer.

[2 marks]

(ii) How can the CPU's registers be used to pass parameters to system calls, and why might this cause problems if system calls take many parameters? Propose an alternative method of passing parameters, that avoids this problem.

[4 marks]

(iii) Give three examples of when a context-switch between processes is performed. What are the actions taken by a kernel during the context-switch? Explain the advantages and disadvantages of context-switching.

[5 marks]

(c) Using RSA, choose p =5 and q =13. For each of the following, show every step of your working:

(i) Encode the first three digits (upper-case letters) of your first name by encrypting each letter separately. If your first name contains less than three letters, combine it with the first letter(s) from your surname (last name) so you will have 3 digits in total. Hint: Replace each upper-case letter by its position in the alphabet, for example A=1, B=2, Z=26.

(ii) Apply the decryption algorithm to the encrypted version in (i) to recover the original plaintext message, i.e, the three digits that you have encrypted.

[6 marks]

Question 3

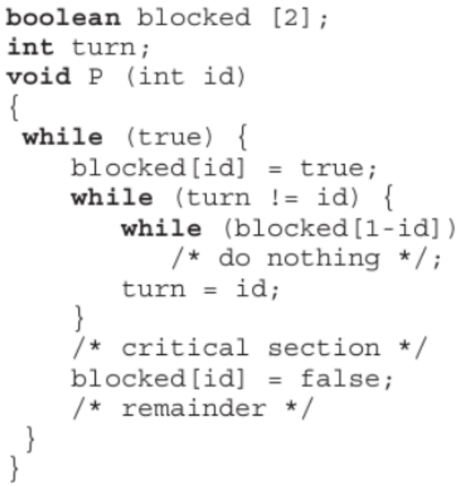
(a) Consider the following workload:

Process	Burst Time	Priority	Arrival Time
P1	60	2	0
P2	20	4	0
P3	50	1	10
P4	10	5	20
P5	20	3	50

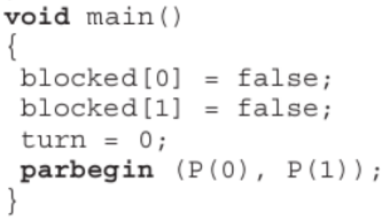
Schedule the above processes using Shortest Remaining Time First (SRTF) and Preemptive Priority algorithms. Please note that for priority scheduling, a lower number indicates higher priority.

Compute the AWT (Average Waiting Time), Average Turnaround Time (ATAT) and Average Response Times (ART) for both of the given policies. Which one of these policies is better-suited to the given workload? **[6 marks]**

(b) Consider the following program:



```
boolean blocked [2];
int turn;
void P (int id)
{
 while (true) {
    blocked[id] = true;
    while (turn != id) {
       while (blocked[1-id])
          /* do nothing */;
       turn = id;
    }
    /* critical section */
    blocked[id] = false;
    /* remainder */
 }
}
```



```
void main()
{
 blocked[0] = false;
 blocked[1] = false;
 turn = 0;
 parbegin (P(0), P(1));
}
```

The above software solution to the mutual exclusion problem for two processes has been suggested in the literature. Do you agree that the proposed solution is correct? If yes, explain with the help of supporting arguments and examples. If no, find a counter-example that demonstrates that this solution is incorrect and does not preserve mutual exclusion? **[6 marks]**

(c) Consider a computer system that runs 10,000 jobs per month with no deadlock-prevention or deadlock-avoidance scheme. Deadlocks occur about three times per month, and the system operator must terminate and rerun about 40 jobs per deadlock. Each job is worth about \$3 (in CPU time + resources), and the manually terminated jobs tend to be about half-done when they are aborted.

A systems' programmer has estimated that a deadlock-avoidance algorithm (like the banker's algorithm) could be implemented in the system with an increase in the average execution time per job of about 5%. Since the machine currently has 30% idle time, all of the 10,000 jobs per month could still be run and completed on-time. Estimate the cost of running this system per month with and without any deadlock-avoidance scheme. In your calculations, also include the cost of an operator who charges \$400 every time he manually intervenes and restarts the deadlocked processes. Finally, indicate which one of these two schemes would be more cost effective. Support your conclusions with appropriate calculations. **[8 marks]**

34244 LM Designing and Managing Secure Systems

Question 1

Two variants for storing a local password database are presented in the table below: Here, `usrX` is the user name, `H()` is a cryptographically secure hash function (e.g. SHA256), and `pwdX` and `saltX` are a user defined string and a random public value respectively, associated to `usrX`.

variant1	variant2
usr1 H(pwd1)	usr1 salt1 H(salt1,pwd1)
usr2 H(pwd2)	usr2 salt2 H(salt2,pwd2)
...	...
usrN H(pwdN)	usrN saltN H(saltN,pwdN)

(a) Assuming perfect cryptography which *variant* is more secure and why? **[6 marks]**

(b) Assume that an attacker which can construct rainbow tables efficiently gains access to a populated, in-use database. Which of the two variants is more vulnerable to this attacker and why? **[8 marks]**

(c) Finally, in the variants above you replace the `H()` function with a key derivation function `KDF()`.

(i) Give the main the inputs to a KDF() and explain their purpose from a security point of view. **[6 marks]**

(ii) Would the security provided by either of the to variants increase? Why? Discuss for both variants. **[4 marks]**

(iii) Give an example of a type of system that uses passwords and will benefit from using a `KDF()` instead of a hash function and state why using a hash function is insufficient in your chosen example. **[6 marks]**

Question 2

CloudServe is a company that offers outsourced Linux virtual servers. It has equipment in a variety of data centres around the world; it owns the equipment, but rents space in the data centres and connects to third-party networks.

CloudServe employs system administrators to maintain both its own and, for additional payment, its customers' systems.

You have recently been appointed as the Chief Security Officer for CloudServe.

You are responsible for CloudHR's IT security, in particular compliance with legal and regulatory requirements.

You recently employed a penetration testing company to evaluate the technical and operational security of your systems. They have submitted a lengthy report. It lists:

- Weaknesses in the configuration of two systems of which you were unaware and which do not feature in your asset register;
- Several instances of systems running old software which is now regarded as not best practice for new installations;
- Weak physical security practices at data centres you use which allowed penetration testers to gain access to the building;
- Poorly configured WiFi in your offices which they were able to access.

Your CEO is, of course, very concerned about this.

Write a memo to the CEO, of approximately one page, summarising your response. What do you propose to do immediately (within seven days), in the medium term (within a month) and in the long term (three months or longer) in order to:

(a) Remove or mitigate the problems that are described; **[10 marks]**

(b) Prevent these or similar problems happening again; **[10 marks]**

(c) Improve your processes and security awareness more generally. **[10 marks]**

30244 LM Intelligent Robotics (Extended)

Question 1

Your company develops software to assist a local cricket club. Based on statistical data, you develop a Bayesian network (see below) that considers the state of a cricket field and the viability of a match on a particular day based on the occurrence of rain and the operation of a sprinkler earlier that day.

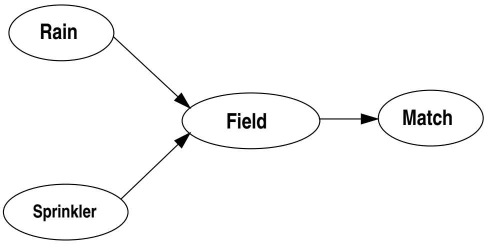


Figure 1: Bayesian network example.

R	**S**	$\mathbf{P(F\|R,S)}$
$\neg r$	$\neg s$	0.1
$\neg r$	s	0.3
r	$\neg s$	0.8
r	s	0.9

Table 1: Conditional probabilities.

The random variables and probabilities encoded in this Bayesian network are:

- Rain (R): $r = \text{rain}$; $\neg r = \text{no rain}$. $p(r) = 0.4$.
- Sprinkler (S): $s = \text{operated}$; $\neg s = \text{not operated}$. $p(s) = 0.2$.
- Field (F): $f = \text{wet outfield}$; $\neg f = \text{dry outfield}$.
- Match (M): $m = \text{played}$; $\neg m = \text{abandoned}$. $p(m|f) = 0.2$, $p(m|\neg f) = 0.9$.

(a) What is the probability that a cricket match is played when there is neither rain nor the operation of a sprinkler? **[2 marks]**

(b) What is the probability that the match is played when the sprinkler is not operated? **[4 marks]**

(c) What is the probability the sprinkler was operated that morning given that the outfield is observed to be wet? **[5 marks]**

(d) Expand and compute the values of $P(r, s, f, m)$ and $P(\neg r, s, \neg f, \neg m)$. **[4 marks]**

(e) You also process images to determine how and when a batsman is out. Assume that a batsman can be bowled or stumped, and the probability of being bowled is 0.6. Assume that a key visual feature v can be used to determine how a batsman got out, and the probability of observing v is 0.3 when the batsman is bowled and 0.6 when he is stumped. If v is observed in a previously unseen image, compute whether the batsman was bowled or stumped. Also state the decision rule if a batsman can get out in K different ways (e.g., bowled, stumped, caught, LBW etc). **[5 marks]**

Question 2

You are asked to program an unmanned aerial vehicle (UAV) to secretly locate a prisoner held in one of three enemy camps. You give the UAV a map of the region with the three camps, but the UAV's estimate of its position and that of the prisoner are uncertain. The UAV can move to any camp, check if it is in a particular camp, check for the prisoner's presence in a camp, and confirm the prisoner's presence in a camp. The UAV's field of view is limited to the camp it is in. The UAV's movement from one camp to another succeeds 85% of the time. The UAV estimates its own location correctly 90% of the time. If the UAV is exploring the camp with the prisoner, the UAV finds the prisoner 90% of the time. The UAV can observe its location, or that of a prisoner, in an incorrect camp 8% of the time.

(a) What mathematical formulation is suitable for this problem and why? **[2 marks]**

(b) Describe the chosen mathematical formulation for the task of locating the hidden prisoner, using the appropriate format. Specify any related probabilistic functions for at least one movement action, one sensing action related to the UAV, one sensing action related to the prisoner, and one terminal action. **[14 marks]**

(c) State key challenges in using the chosen mathematical formulation for complex robotics problems, and describe some potential solutions. **[4 marks]**

Question 3

You are writing software for Gaussian filter-based localization (i.e., estimation of 2D pose) of an indoor assistant robot. The robot has a learned map of the home with N strategically positioned visual landmarks. Assume that the robot has an odometry-based motion model, and that observations of landmarks are in the form of range and orientation measurements relative to the robot.

(a) Determine and describe the update equations for the domain. Also describe some limiting assumptions of this formulation. **[4 marks]**

(b) Choose a formulation for the state estimation (i.e., localization) problem that allows non-linear functions in the update equations and linearizes these functions based on a tangent at the mean. Describe this formulation by computing the corresponding equations and expressing them in matrix form. **[16 marks]**.

30255 LM Machine Learning and Intelligent Data Analysis (Extended)

Question 1 Dimensionality Reduction

(a) Explain what is meant by "dimensionality reduction" and why it is sometimes necessary. **[4 marks]**

(b) Consider the following dataset of four sample points $\{\mathbf{x}^{(i)}\}_{i=1}^{4}$ with $\mathbf{x}^{(i)} \in \mathbb{R}^2 \; \forall i$:

$$\mathbf{X} = \begin{pmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{pmatrix}$$

Explain how to calculate the principal components of this dataset, outlining each step and performing all calculations up to (but not including) the computation of eigenvectors and eigenvalues. **[6 marks]**

(c) What does principal component analysis (PCA) tell you about the nature of a multivariate dataset? Explain how it can be used for dimensionality reduction? **[4 marks]**

(d) What are the limitations of PCA and what other dimensionality reduction techniques may be used instead? **[2 marks]**

(e) You are given a dataset consisting of 100 measurements, each of which has 10 variables. The eigenvalues of the covariance matrix are shown in the following table:

Eigenvalue number	1	2	3	4	5	6	7	8	9	10
Eigenvalue	1382.0	508.4	187.0	68.8	25.3	9.3	3.4	1.3	0.46	0.17

What can you say about the underlying nature of this dataset? **[4 marks]**

Question 2 Classification

(a) Consider the Soft Margin Support Vector Machine learnt in Lecture 4e. Consider also that $C = 100$ and that we are adopting a linear kernel, i.e., $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)T}\mathbf{x}^{(j)}$. Assume an illustrative binary classification problem with the following training examples:

$\mathbf{x}^{(1)} = (0.3, 0.3)^T$, $y^{(1)} = 1$
$\mathbf{x}^{(2)} = (0.6, 0.6)^T$, $y^{(2)} = 1$
$\mathbf{x}^{(3)} = (0.6, 0.3)^T$, $y^{(3)} = -1$
$\mathbf{x}^{(4)} = (0.9, 0.6)^T$, $y^{(4)} = -1$

Which of the Lagrange multipliers below is(are) a plausible solution(s) for this problem? **Justify your answer.**

(i) $a^{(1)} = 0$, $a^{(2)} = 2$, $a^{(3)} = 2$, $a^{(4)} = 10$

(ii) $a^{(1)} = 0$, $a^{(2)} = 44$, $a^{(3)} = 22$, $a^{(4)} = 22$

(iii) $a^{(1)} = 0$, $a^{(2)} = 200$, $a^{(3)} = 100$, $a^{(4)} = 100$

[6 marks]

(b) Consider a binary classification problem where around 5% of the training examples are likely to have their labels incorrectly assigned (i.e., assigned as -1 when the true label was +1, and vice-versa). Which value of k for k-Nearest Neighbours is likely to be better suited for this problem: $k = 1$ or $k = 3$? **Justify your answer.** **[6 marks]**

(c) Consider a binary classification problem where you wish to predict whether a piece of machinery is likely to contain a defect. For this problem, 0.5% of the training examples belong to the defective class, whereas 99.5% belong to the non-defective class. When adopting Naïve Bayes for this problem, the non-defective class may almost always be the predicted class, even when the true class is the defective class. Explain why **and** propose a method to alleviate this issue. **[8 marks]**

Question 3 Document Analysis

(a) In a small universe of five web pages, one page has a PageRank of 0.4. What does this tell us about this page? **[2 marks]**

(b) Compare and contrast the TF-IDF and word2vec approaches to document vectorisation. You should explain the essential principles of each method, and highlight their respective advantages and disadvantages. **[8 marks]**

(c) One possible approach to searching a large linked set of documents is to combine a measure of document similarity such as TF-IDF similarity with a measure of a page's importance such as that provided by PageRank. Suggest three ways in which this could be done and discuss the advantages and disadvantages of each of them. **[10 marks]**

32250 LM Mathematical Foundations of Artificial Intelligence (AI) and Machine Learning

Question 1

Consider a "cigar shaped" d-dimensional data set $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N\}$, $\mathbf{x}^i \in \mathbb{R}^d$, $i = 1, 2, ..., N$, where even though $d > 1$, the data predominantly stretches along a single direction (not necessarily a co-ordinate axis). Assume the data has been centred to zero mean, i.e. $\sum_{i=1}^{N} \mathbf{x}^i = \mathbf{0}$. The data dimensionality needs to be reduced to 1. You and your friend come up with different approaches to achieve this:

- You suggest to find a line ℓ in the data space $\mathbb{R}^d$ passing through the origin $\mathbf{0}$, such that when the data points $\mathbf{x}^i \in \mathcal{D}$ are orthogonally projected onto ℓ, the sum of distances of the original points $\mathbf{x}^i$ to their projections $\mathbf{x}_p^i \in \ell \subset \mathbb{R}^d$ is minimised. You argue that in this way the loss of structure in $\mathcal{D}$ due to its projection onto the one-dimensional line ℓ is minimised.
- Your friend has another idea: Find a line $\ell \subset \mathbb{R}^d$ passing through the origin $\mathbf{0}$ such that the variance of the projections $\mathbf{x}_p^i \in \ell$ along ℓ is maximised. They argue that even though ℓ is inherently a one-dimensional object, it is positioned in $\mathbb{R}^d$ so that as much data variability in $\mathcal{D}$ as possible is preserved in the projections.

(a) Draw a schematic picture illustrating the two approaches for the case $d = 2$. Make an intuitive argument why maximising variability of data projections in ℓ may correspond to minimising the distance from projections to the original points. **[6 marks]**

(b) Formulate both approaches as formal optimisation problems. **[12 marks]**

(c) Show that the two approaches are in fact equivalent, i.e. they both define the same optimal line ℓ. **[12 marks]**

Question 2

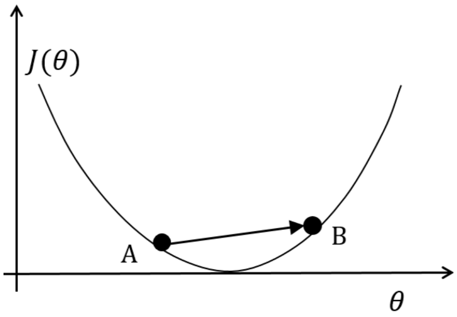
(a) You would like to find a linear model $f(x) = \theta_0 + \theta_1 x$ to fit the dataset shown in the table below. Rather than using the mean squared error as the cost function, you would like to use the mean absolute cubic error $J(\theta) = 1/m \sum_{i=1}^{m} |f(x^i) - y^i|^3$, where $m = 2$ is the number of data examples in the dataset, $f(x^i)$ denotes the predicted value for the i-th data example, and y^i is the associated target value. We use gradient descent to iteratively update the parameters θ_0 and θ_1 starting from $\theta_0 = \theta_1 = 0$.

Data	x	y
example no. 1	1.0	2.0
example no. 2	2.0	1.0

(1) Calculate the cost $J(\theta_0)$ when $\theta_0 = \theta_1 = 0$. **[4 marks]**

(2) Calculate θ_0 and θ_1 after the first iteration of the update process in the gradient descent, where the learning rate is set to 0.1. Please show the step-by-step calculations. **[12 marks]**

(b) Consider a univariate linear regression model $f(x) = \theta x$ (where the intercept is zero), and a cost function $J(\theta)$ shown in the figure below. We use the gradient descent to try to iteratively approach the lowest cost. In the update process we observed an overshoot illustrated in the figure as moving from position **A** to position **B**.



(1) Of the following cases, which will happen? **[4 marks]**

(A) θ will eventually diverge.

(B) θ will eventually converge.

(C) θ will go back and forth between the two positions **A** and **B**,

(D) θ can proceed randomly in any direction.

(2) Provide an explanation. **[10 marks]**

34231 LM Network Security + Cryptography

Question 1 (Symmetric-key cryptography)

A bank app and the bank server share a key K. When the user requests a transfer of \$1000.00 to account 12345678, the app uses AES128 in CTR mode to encrypt using the key K the message

```
TRANS "1000.00" TO "12345678"
```

(a) The message consists of 29 bytes. After encryption by AES128 in CTR mode, how many bytes are there in the ciphertext? (Don't forget to include the block for the nonce and the counter.) **[3 marks]**

Suppose the attacker can intercept and modify this message ciphertext as it travels from the app to the server.

(b) Explain how the attacker can modify the ciphertext so that when it is decrypted by the server, it will say

```
TRANS "9999.99" TO "12345678".
```

Your explanation should mention explicitly which bytes of the ciphertext the attacker will modify. **[4 marks]**

To avoid this attack, the bank programmer decides to use a MAC function to authenticate the message. Her design is as follows:

- From the key K, the app and the server derive two keys, $K_1 = \text{HMAC}_K(1)$ (to be used for encryption) and $K_2 = \text{HMAC}_K(2)$ (to be used for MAC).
- Send the encryption by AES128 in CTR mode using the key K_1 of the message `TRANS "1000.00" TO "12345678"`, together with the HMAC using key K_2 of the plaintext message `TRANS "1000.00" TO "12345678"`. When the server receives the message, it decrypts the first part with K_1, applies HMAC with K_2 to the result, and checks that this value equals the received HMAC. It rejects the message if the check fails.

(c) Is this method secure against the integrity attack of part (ii)? Explain your answer. **[5 marks]**

(d) Is this method secure against confidentiality attacks? Explain your answer. **[3 marks]**

Question 2 (Public-key cryptography)

Consider MElGamal, a variant of the ElGamal public key encryption (PKE) scheme, with encryption algorithm that only accepts plaintexts in $\{0,1\}$. MElGamal is defined as follows:

- Key generation (λ)
  - Let q, p be primes w.r.t. security parameter λ such that $q | p-1$
  - Let $g \neq 1$ be such that $g^q = 1 \mod p$
  - Let $\mathbf{G}$ be the subgroup of $\mathbf{Z}_p^\star$ generated by g
  - Let $x \overset{R}{\leftarrow} \mathbf{Z}_q$. Let $h = g^x \mod p$
  - Public-key : $PK = (p, q, g, h)$. Private-key : $SK = x$
- Encryption ($PK, m \in \{0,1\}$)
  - m must be a single bit (i.e. $m \in \{0,1\}$).
  - Let $r \overset{R}{\leftarrow} \mathbf{Z}_q$
  - Output $c = (g^r \mod p,\ h^r \cdot g^m \mod p)$
- Decryption ($SK, C = (c_1, c_2)$)
  - ...

(a) Define the decryption algorithm of MElGamal. Show that MElGamal satisfies the completeness property for PKE schemes. **[4 marks]**

(b) Is MElGamal an IND-CPA secure scheme? Justify your answer. **[3 marks]**

(c) Let us define an operation $\otimes$ over MElGamal ciphertext as follows: $(c_1, c_2) \otimes (c_1', c_2') := (c_1 \cdot c_2 \mod p,\ c_1' \cdot c_2' \mod p)$. Show that for any $m, m' \in \{0,1\}$ such that $m \neq m'$ holds that $\mathsf{Enc}(PK, m) \otimes \mathsf{Enc}(PK, m') = \mathsf{Enc}(PK, m \oplus m')$, where $\oplus$ is the XOR operation. **[5 marks]**

(d) Is MElGamal non-malleable? Justify your answer. **[3 marks]**

Question 3 (Network security)

Wireless networks are increasingly common, and in many organisations have almost completely replaced wired networks for client-device access.

(a) Explain the difference between WPA2 PSK (pre-shared key, also called WPA2 Personal) and WPA2 Enterprise. What criteria should one use to inform the choice between these two systems? Why might some companies be tempted to use WPA2 PSK instead of WPA2 Enterprise? **[9 marks]**

You are the network manager of a large shared office building which has many tenants from different organisations, all using a single shared wireless network secured with WPA2 PSK. You have been approached by one of the tenant organisations, who has concerns about the security of the wireless network, given the fact that everyone is using the same pre-shared key. They have asked you the following questions:

Q1 Can other tenants within the building read the emails that we send and receive?

Q2 Can other tenants send emails that look like they come from our organisation?

Q3 Can other tenants see what URLs we visit when web browsing? What about the content of the web pages?

Q4 Can other tenants access the shared drive we have in the building (it is meant to be accessible only to members of our organisation).

(b) Answer the questions as best you can. Mention any clarifications you would need to obtain, in order to give more precise answers. **[6 marks]**

The tenant organisation that approached you suggests you should upgrade the whole network to WPA2 Enterprise. You have been asked to respond to the CEO of the tenant organisation, advising her of the best course of action.

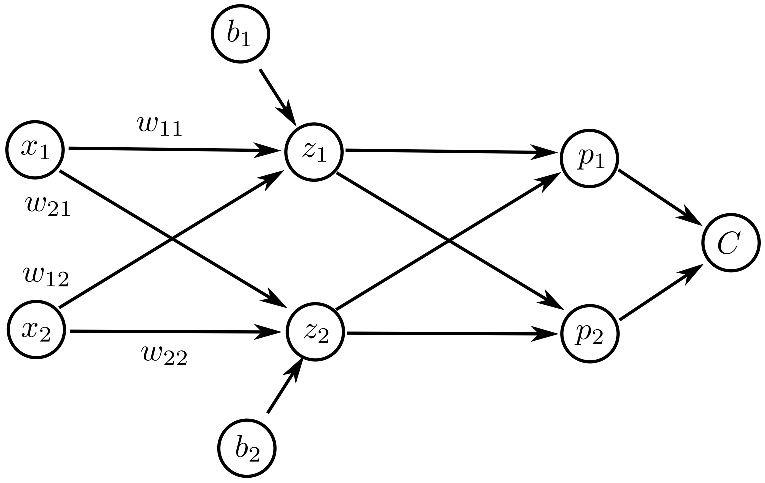
(c) Write a memo, of approximately half a page, briefing her on the issues. You should consider:

- What are the problems that using a single shared-key network for multiple enterprises will cause?
- What will need to be done to deploy WPA2 enterprise, and what will the costs be?
- What will be the benefits?
- Someone has suggested having both WPA2 PSK and WPA2 Enterprise at the same time. Could that work?

[15 marks]

32212 LM Neural Computation (Extended)

Question 1



Consider the neural network above where the units are defined for $j \in \{1, 2\}$ as follows:

$$z_j = w_{j1}x_1 + w_{j2}x_2 + b_j,$$
$$p_j = \frac{e^{z_j}}{Q}, \text{ where } Q = e^{z_1} + e^{z_2}.$$

The network is trained using stochastic gradient descent (i.e., minibatch size 1). For a training example with inputs $x_1, x_2 \in \mathbb{R}$ and correct output $y \in \{1, 2\}$, the cost C is defined as

$$C = \log(Q) - z_y.$$

(a) How does the sum $p_1 + p_2$ relate to the inputs x_1 and x_2?

[4 marks]

(b) Compute the partial derivatives $\frac{\partial C}{\partial z_1}$ and $\frac{\partial C}{\partial z_2}$

[6 marks]

(c) Suppose that the gradient step increases the weight parameter w_{11}, and that $x_1 > x_2 > 0$. What can you say about the correct output y? Justify your answer.

[10 marks]

Question 2

Consider a set of two input-output pairs $\big(\binom{1}{2}, 1\big), \big(\binom{-1}{2}, -1\big)$. Let the loss function be the least square and we wish to find a linear model $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$, where $\mathbf{x}^\top$ is the transpose of $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{w} = \binom{w_1}{w_2} \in \mathbb{R}^2$. Then the objective function becomes

$$J(\mathbf{w}) = \frac{1}{2}\left(\frac{1}{2}\Big((1,2)\mathbf{w} - 1\Big)^2 + \frac{1}{2}\Big((-1,2)\mathbf{w} + 1\Big)^2\right).$$

Let us build our prediction model by minimizing the above objective function.

(a) Simplify the objective function to the form of

$$J(\mathbf{w}) = c_1 w_1^2 + c_2 w_2^2 + c_3 w_1 + c_4 w_2 + c_5 w_1 w_2 + c_6,$$

where $c_k \in \mathbb{R}$ are coefficients, and w_i is the i-th coordinate of $\mathbf{w} \in \mathbb{R}^2$. After that, compute the gradient of $J(\mathbf{w})$ in terms of c_k.

[8 marks]

(b) Consider gradient descent with the initial point $\mathbf{w}^{(0)} = \binom{3}{1}$ and learning rate $\varepsilon = 0.5$. Write down the process of calculating $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$. After that, calculate $J(\mathbf{w}^{(1)})$ and $J(\mathbf{w}^{(2)})$.

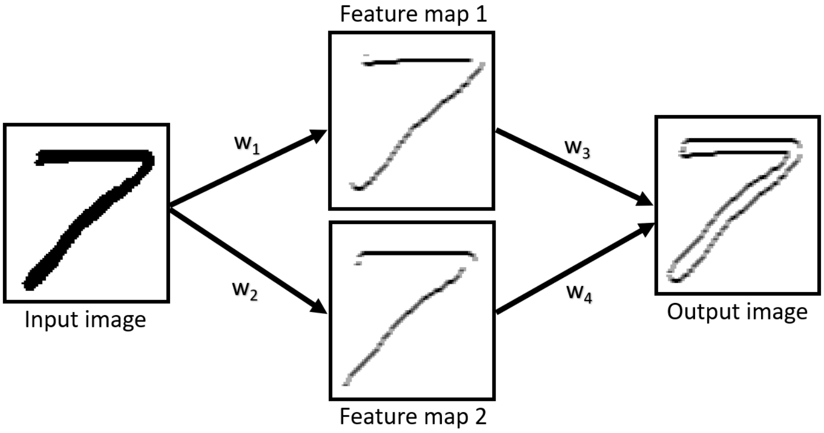
[6 marks]

(c) Consider gradient descent with momentum. Assume $\mathbf{w}^{(0)} = \binom{3}{1}$, learning rate $\varepsilon = 0.5$ and momentum rate $\alpha = 0.5$. Write down the process of calculating $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$. After that, calculate $J(\mathbf{w}^{(1)})$ and $J(\mathbf{w}^{(2)})$.

[6 marks]

Question 3

A convolutional network is shown in the figure below. We use the nonlinear ReLU activation function in the first layer and the linear activation function in the output layer. Note that for better visualisation, in the images we use white regions to denote 0 and darker regions to denote larger values.



(a) Design appropriate convolution kernels of size 3×3 for the first layer such that the feature maps 1 and 2 are these displayed in the figure. Justify your answer.

[5 marks]

(b) Design appropriate convolution kernels of size 3×3 for the output layer such that the output is that displayed in the figure. Justify your answer.

[5 marks]

(c) How many (free) parameters (weights) in the convolutional network do we need to train? Justify your answer.

[4 marks]

(d) In convolutional networks, apart from ReLU there exist other nonlinear activation functions such as Sigmoid and tanh. For the collection of neurons in a single layer, what would be the Jacobian matrices of the Sigmoid and tanh functions, respectively?

[6 marks]

30256 LM Programming Language Principles, Design, and Implementation (Extended)

Question 1

(a) Consider the following function, which we call M:

$$\lambda f : \mathbb{B} \to \mathbb{B}.\lambda g : \mathbb{B} \to \mathbb{B}.\lambda x : \mathbb{B}.\lambda y : \mathbb{B}.\texttt{if}\ x\ \texttt{then}\ f\ y\ \texttt{else}\ g\ y$$

(i) Prove (using a proof tree) that M is a well-typed expression of the Simply Typed λ-Calculus.

(ii) Provide examples of expressions F and G such that $(M\ F\ G)$ is the exclusive or function (i.e., it returns `true` iff its two arguments are different).

(iii) The expression $(M\ F\ G\ \texttt{false}\ \texttt{true})$ has type $\mathbb{B}$. Describe the safety guarantee we can deduce about this expression from the fact that it is a well-typed expression of the Simply Typed λ-Calculus.

(iv) Prove (using proof trees) that $(M\ F\ G\ \texttt{false}\ \texttt{true})$ computes to a value.

[10 marks]

(b) Similar to the encoding of numbers in System F, we can define stacks as follows (we only define stacks of numbers here). The stack type is defined as follows:

- $\texttt{Stack} = \forall\alpha.(\mathbb{N} \to \alpha \to \alpha) \to \alpha \to \alpha$

For example, the empty stack is $s_0 = \lambda\alpha.\lambda f : \mathbb{N} \to \alpha \to \alpha.\lambda x : \alpha.x$; the stack $s_1 = \lambda\alpha.\lambda f : \mathbb{N} \to \alpha \to \alpha.\lambda x : \alpha.f\ n\ x$ is the stack that contains only the value n; and the stack $s_2 = \lambda\alpha.\lambda f : \mathbb{N} \to \alpha \to \alpha.\lambda x : \alpha.f\ n\ (f\ m\ x)$ is the stack that contains the value m on top of the value n. More generally, a function of the form $\lambda\alpha.\lambda f : \mathbb{N} \to \alpha \to \alpha.\lambda x : \alpha.f\ n_1\ (\ldots(f\ n_k\ x)\ldots)$ encodes a stack where n_k is stacked on top of n_{k-1}, which is stacked on top of n_{k-2} etc., and where n_1 is at the bottom of the stack. The `push`, `peek`, and `pop` operations are defined as follows:

- $\texttt{push} = \lambda n : \mathbb{N}.\lambda s : \texttt{Stack}.\lambda\alpha.\lambda f : \mathbb{N} \to \alpha \to \alpha.\lambda x : \alpha.s\{\alpha\}\ f\ (f\ n\ x)$
- $\texttt{peek} = \lambda d : \mathbb{N}.\lambda s : \texttt{Stack}.s\{\mathbb{N} \to \mathbb{N}\}\ G\ I\ d$, where
  - $G = \lambda n : \mathbb{N}.\lambda g : \mathbb{N} \to \mathbb{N}.\lambda x : \mathbb{N}.g\ n$
  - $I = \lambda x : \mathbb{N}.x$
- $\texttt{pop} = \lambda s : \texttt{Stack}.\lambda\alpha.\lambda f : \mathbb{N} \to \alpha \to \alpha.\lambda x : \alpha.s\{(\alpha \to \alpha) \to \alpha\}\ G\ C\ I$, where
  - $G = \lambda n : \mathbb{N}.\lambda g : (\alpha \to \alpha) \to \alpha.\lambda h : \alpha \to \alpha.h\ (g\ (f\ n))$
  - $C = \lambda h : \alpha \to \alpha.x$
  - $I = \lambda x : \alpha.x$

The function push pushes a number onto a stack, peek returns the last number pushed onto a stack, and pop removes the last element pushed onto a stack, i.e., the furthest one to the right. Note that peek takes a default value d as argument, which is returned when the stack is empty, i.e., (peek d s_0) computes to d, while (peek d s_1) computes to n, and (peek d s_2) computes to m.

(i) Provide an example of a function of type Stack that encodes the stack where 1 is stacked on top of 0, which is stacked on top of another 0, which is at the bottom of the stack.

(ii) Prove (using a proof tree) that peek is a well-typed expression of System F.

(iii) Prove (using proof trees) that (peek d s_2) computes to m, where d is an expression of type $\mathbb{N}$, and s_2 is defined above.

(iv) Using the extension of System F with existential types we saw in the lectures, define a stack abstract datatype that hides this implementation and provides the following interface: the empty stack, push, peek, pop.

[10 marks]

Question 2

Consider the following recursion combinator:

$$Y = \lambda t.(\lambda f.t(\lambda z.ffz))(\lambda f.t(\lambda z.ffz))$$

(a) Express Y as an ASG. **[5 marks]**

(b) Explain, using an ASG abstract machine, how the following expressions are evaluated, given that the multiplication operator $(*)$ is a "shortcut" operator which does not evaluate its second argument if the first argument is 0.

(i) $Y(\lambda f.\lambda x.0)1$

(ii) $Y(\lambda f.\lambda x.f(x) * 0)1$

(iii) $Y(\lambda f.\lambda x.0 * f(x))1$

If an expression evaluates to a constant value show clearly the result and also the key intermediate steps in the ASG rewrite. If an expression diverges make a clear argument as to why it diverges and what the runtime behaviour will be. **[10 marks]**

(c) Furthermore:

(i) How would your answers above differ if multiplication $*$ always needed to evaluate both arguments?

(ii) Mathematically, for any x, $x * 0 = 0 * x = 0$. Which one of the second and third expressions above can be *optimised* into the first expression? Explain your answer.

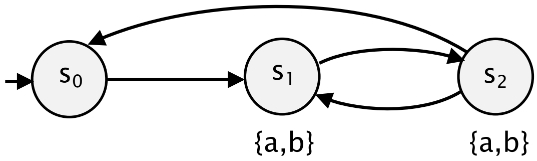
(iii) Can the first expression be optimised just to the constant value 0? Justify your answer.

[5 marks]

Hint: You may want to use a drawing tool to allow you to evaluate ASGs by copying-and-pasting then editing rather than drawing every stage of the rewrite from scratch.

Question 3

Consider the labelled transition system (LTS) below:



(a) Below are several properties that could be verified on this LTS. For each one, state which classes of property it belongs to (e.g., invariant, safety, liveness), justifying your answer, and translate it into LTL. Assume that a and b are atomic propositions.

(i) a is true whenever b is true;

(ii) a and b are simultaneously true only a finite number of times;

(iii) every b is immediately followed by a;

(iv) exactly one of a or b (but not both) is eventually true;

(v) if a ever becomes true, then it remains true forever, and this is immediately preceded by a state where b was true.

[6 marks]

(b) Illustrate the LTL model checking procedure for verifying property (iii) from above against the LTS. **[6 marks]**

(c) Let us now relax property (iii) to "every a is followed by b within at most k steps". Assume a new LTL operator $\Diamond^{\leq k}\psi$, which states that ψ becomes true within k steps (in other words, for a trace $\sigma = A_0A_1A_2\ldots$, we have $\sigma \models \Diamond^{\leq k}\psi$ if and only if $A_iA_{i+1}A_{i+2}\ldots \models \psi$ for some $i \leq k$). Write the new property in this extension of LTL and discuss how it could be model checked on the LTS above. Then discuss how general LTL model checking could be extended to include this new $\Diamond^{\leq k}$ operator. **[8 marks]**

32245 LM Storing and Managing Data

Some of the following questions relate to the Pagila database, the details of which are available at https://canvas.bham.ac.uk/courses/46252/pages/pagila-schema. Ensure that your queries are well formatted. This paper contains THREE questions across TWO pages.

Question 1

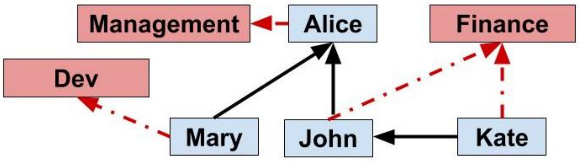
(a) Write a query to find the top three categories ranked (high to low) by the number of times English films in those categories were rented. Your must use joins. **[10 marks]**

(b) How many times has any customer whose first name starts with "A" rented movies in which "JULIA MCQUEEN" (first name Julia) acted? Your solution must use joins and your output should be a single number. **[10 marks]**

Question 2

(a) We've decided to give a 5% discount on Sci-Fi films that have so far earned us less than £100 through rentals across all stores. Sci-Fi films that have earned us £100 or more are not eligible for the discount. Write a query to output the names of *ALL* Sci-Fi movies and the associated rental rate rounded to 2 decimal places. Your solution must use joins and output two columns: The name of the film and the new rental price, which is either: the original rental price for films that are not eligible for the the discount or the original rental price with a discount of 5% for films that are eligible for the discount. **[10 marks]**

(b) Draw an Entity Relation Diagram (ERD) that captures the relations in the Instance Diagram in Figure below. The solid black lines represent the relation "Reports To" and the dashed red lines represent the relation "Belong to department". Your ERD should capture both the Employee and Department relations and also include any other relevant attributes of these Entities. Also write queries to create the corresponding tables for these entities (including any required weak entities) being sure to include any relevant constraints, especially primary keys and foreign keys.



[10 marks]

Question 3

Consider the following scenario. A banking company uses a single PostgreSQL database server from 2012 on a machine with an operating system from a similar era to store its customers' data. All user details, including transaction details and login details are stored in the same database in a plaintext format to make sure the user login process is quick. The bank has a web server on the same hardware as the database server that customers can use to do online banking. The bank has a certificate for their website `www.securebank.com` and customers can access their bank accounts via a website accessed at `http://banking.securebank.com`. Customers can search for transactions from a certain period and there is a search box for a user to input the date that they want which is then passed as a plaintext string into a query string to return the relevant data. The bank does have security measures in place and has a firewall with the credentials set as default to make sure no one forgets them.

What are the problems with both storing and transmitting the data securely? For each problem please pose a solution or criteria for ensuring the secure storage or transmission of the data in question. **[20 marks]**

35447 Programming for Data Science

Question 1

Imagine that you are a data scientist working on a recommender system for a music streaming company. You are given the following toy dataset (`https://canvas.bham.ac.uk/files/10125436/download?download_frd=1`) to work on to produce a simple recommender program in Python using objects, classes and methods. You should not use any external libraries when developing your solution to this question.

UserID	Username	Genre	Listens
1	Dave	Rock	100
1	Dave	Country	1245
1	Dave	Pop	1
2	Dolly	Country	100312
2	Dolly	Trance	100
3	Amadeus	Classical	34912390
3	Amadeus	Country	5

(a) Write a program that does the following:

- Reads in the attached text file. Each field is delimited by a comma. You must ignore the header.
- Your program must create an object for each unique user. To do this you need to define a class called `User`. Each instance of this class must take in the `UserID` and the `Username` via its constructor. There should only be one object created for each unique UserID.
- Each `User` object must contain a dictionary representing the listening history of the user. This should be stored as an instance variable. As your program reads in each line of the file, you must update this dictionary with the genre and the number of listens, whereby genre is the key. Assume that each genre will only appear once per user in the input file. You must define a method called `addToHistory` to add each key-value pairing to the dictionary. You must also define a getter method called `getHistory` that returns a dictionary of a user's listening history.
- In the `User` class you must create a method called `printSummary` that prints out a summary of the listening history for each user, with the total number of genres listened to, and the total number of listens for that user. An example for Dave would be: 'Dave has listened to 3 genres a total of 1,346 times.'

- Create a main method that takes input from the file, creates a list of User objects containing their listening history. Then for each User, call the method that prints a summary of the User's listening history.

Please save your solution to this in a file called "1a.py" and submit it with your other solutions in your zipped submission file. [10 marks]

(b) Write a class called `Recommender`. In its constructor, it should be passed a list of User objects and store these as an instance variable. In this class you should write a method called `generateRecommendations`. This should iterate through the list of User objects and for each one make a single recommendation based upon the listening history of the next user in the list. A simple recommendation algorithm is as follows:

- Load user A and the next user in the list, user B. For all the genres that user B has listened to, if there is one that does not exist in user A's listening history, then print this out as a recommendation. e.g sample output could be: "Dave is recommended Trance". Only produce a single recommendation, even if there are potentially more. For the last user in the list, they should generate their recommendations based on the first person in the list.

After writing the class, `Recommender`, create a runnable main method to demonstrate the class working using a list of `User` objects from part (a). As a comment in the main method, list the output showing what recommendations are generated for each user in the above dataset.

Please save your solution to this in a file called "1b.py" and submit it with your other solutions in your zipped submission file. [10 marks]

Question 2

In this question, you have to write a program that processes the data for a book lending library. You have the following input data:

- The available books per author (numbers are book IDs):
  "Smith": 1, 3, 3, 3, 1, 5, 5, 2, 4, 4
  "Jones": 6, 6, 7, 8, 9, 9, 7, 9, 9, 10
  "Hope": 11, 12, 13, 14, 15, 11, 12, 13, 14, 15
  "Mason": 16, 17, 18, 18, 18, 19, 20, 20, 16, 20
  "Doe": 21, 21, 21, 21, 21, 22, 22, 22, 22, 22
- Information about the genre of each author:
  "Romantic": "Smith", "Hope", "Doe"
  "Adventure": "Jones", "Mason"

Your program must do the following:

- Stores the information about the books and their genre in Pandas DataFrames (N.B can be more than one dataframe).
- Generates at random the borrowing of books for a week as follows: each week 10 unique books are borrowed at random. At the end of each week, all books are returned. Your program should use vectorization and make use of NumPy to make the process more efficient.
- Generate the borrow statistics for 10 weeks and print the following information:
  - The number of borrows per authors, organized by genre.
  - The most popular book in each genre.

Please save your solution to this in a file called "2.py" and submit it with your other solutions in your zipped submission file.

[20 marks]

Question 3

In this question you are required to perform a multi-class classification task on a car evaluation data set. The task is to categorize cars into one out of four evaluation results – "unacc", "acc", "good" and "v-good". Please follow the instruction below ("About the data set") to obtain the data needed for this question.

About the data set: Car evaluation is a public data set, which can be found here `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`. After clicking the "Data Folder" on the webpage, you will see a data file – "car.data", containing the data samples for learning; a data description file – "car.names", explaining data features and labels. Please download them. They can be opened by normal text editors. Because the data file is not a standard csv file read by Scikit-learn, you can manually change the file extension to csv, and start working on it. You can also find the same data file (already changed to csv) here `https://canvas.bham.ac.uk/files/10044903/download?download_frd=1`.

Once you have downloaded the aforementioned data file, please answer the following two parts of this question with Python programs. **For each question, write a Python program, and include an explanation of your results and findings as indicated in the question in a comment within each respective file.**

Scikit-learn and *Scipy* libraries are needed for implementation. You are not limited to these two libraries and may use others if you wish.

(a) Read data into your Python program. Perform feature engineering techniques to transform all the object type of features into numbers (N.B: do not delete any features). Then, build the following TWO classifiers *using the default settings* in Scikit-learn:

- Decision tree in sklearn.tree.DecisionTreeClassifier
- AdaBoost in sklearn.ensemble.AdaBoostClassifier

Compare the accuracy between the decision tree model and the AdaBoost model using 10-fold cross validation. The comparison should be based on an appropriate statistical test. Discuss why you have chosen this test and whether one model performs significantly better than the other having run your code.

Please save your solution to this in a file called "3a.py" and submit it with your other solutions in your zipped submission file. [10 marks]

(b) Propose and implement a strategy to improve the accuracy of your decision tree model in Python. You should write code to perform this, and state briefly in a comment within your code file what strategy you use, why you believe the strategy may work, and whether your strategy is effective based on the statistical test result. N.B you will not be penalised if the final statistical test result shows your strategy does not significantly improve accuracy.

Please save your solution to this in a file called "3b.py" and submit it with your other solutions in your zipped submission file. [10 marks]