General

Hi Prof, may I clarify how partial credit is awarded for MRQ questions?For example, if the correct answer is **ABD** and I selected **ACD**, which differs by only one option, would this be considered as getting 2 out of 3 correct? Currently, it seems to be graded as 1/3.

MRQ is graded using partial marks: max(0, # selected correct - # selected incorrect) / # correct options * max marks.

1A

REVISED ANSWER: B,C,F,G or B,C

1. For Q1A do we accept B, F, G only on the grounds that C is an incomplete description (we must also include in our observation(environment) the current document)?

   Environment is what the agent interacts with. The description of the environment given by C is complete. The current file is the observation of the agent: percepts emitted by the environment, which are retrieved by the agent's sensors.

2. For Qn 1A, in the lectures it was taught that LiDAR are "sensors" while wheels of a car are "actuators". These are tangible objects/entities. Why is it that "abilities" are now also sensors and actuators instead of "entities" that actually execute actions/detect precepts such as the file system doing the email moving? Was the content taught incorrect and misleading or did I miss anything out? I believe that B,C alone should suffice as an alternative fully correct solution.
3. Please review the grading for Question 1A. What are the exact definitions for sensors/actuators? I did not select options (f) and (g) because the phrasing describes capabilities rather than the components themselves. According to the standard PEAS framework, 'Sensors' and 'Actuators' refer to the specific devices or software or hardware mechanisms that an agent uses. Unlike other correct options, Option (f) "The ability to move a document." Option (g) "The ability to read text." These options describe the Agent Function (what the agent can do) rather than the Agent Architecture (the components it possesses). When looking at the list, I rejected these options because I was looking for the mechanism that provides the ability (e.g., 'A file system' or 'A text parser module', etc.'). For comparison, Option (e) correctly identified a component ('A printer'), even though it was contextually incorrect.Because options (f) and (g) described

abstract abilities rather than concrete interface components, I found them to not matching our definitions of 'Actuators' and 'Sensors'."

For queries 2 – 3:

Sensors and actuators may not be tangible objects. Instead, they can be abstract concepts that enable agents to perceive or interact with the environment in some way. However, after reviewing the course content, we have noticed that our examples of sensors and actuators have always been tangible. Thus, we will allow omitting F and G (but not just one of them) from the answer.

1C

1. Q1C option E: the environment can be continuous, since the state space consisting of all possible documents is infinite, as there are infinitely many possible sequences of words in a document.
2. 1C. Why not continuous since state space consists of email content as well which has infinite possible sequences.

For queries 1 – 2:

An environment is discrete if it has a limited number of distinct, clearly defined percepts and actions (Definition from Lecture). It is continuous if it is the opposite: it has an unlimited/infinite number of possible percepts and actions.

In terms of action, the environment is clearly discrete because there are only a finite number of actions: 5 actions.

In terms of percepts, the environment is also discrete. A document is composed of a finite number of sentences. A sentence is composed of a finite number of words. A word is composed of a finite number of letters. The number of possible documents can be combinatorially large, but they are not infinite since they are composed of finitely many constituents.

2A

REVISED ANSWER: A,B,C,E (sequential assumption) or A,C,E (non-sequential assumption)

1. For Q2a, to repeat exactly the phrasing used in the question: "Approach: To prepare the data, the team normalizes all features. To create a training and testing set, the team takes the first 800 records for training and the last 200 for testing." No words indicated the sequence this was being done. For example, if

the question instead said: "Approach: To prepare the data, the team first normalizes all features. Then, to create a training and testing set, the team takes the first 800 records for training and the last 200 for testing." Then, and only then, in this case is it unambiguously true that the team normalized all 1000 features together, and then split into 800/200, resulting in data leakage. Since the question did not specify the sequence, we shouldn't be marked down for not selecting option b.

2. I would like to raise a small clarification regarding Question 2A, particularly option (b) on data leakage. Normalizing the entire dataset before splitting would indeed cause data leakage; however, the question does not explicitly specify the order in which normalization and dataset splitting are performed. Since the wording allows for both interpretations (normalizing before or after the split), the presence of data leakage is ambiguous. Given this, I sincerely suggest that for students who did not select option (b) should not be penalized. Thank you for considering this clarification.

For queries 1 – 2:

As the sequence of normalization and test/train splitting was not explicitly stated, we allow sequential and non-sequential interpretation of the approach.

3. For question 2A (examplify question 4), data leakage from test set into training set (option B) should not be an issue, because the question clearly states that to create a training and testing set, the team takes the first 800 records for training and the last 200 for testing. Since there are 1000 patient records, this implies that the **testing data is mutually exclusive from the training data**. Please don't penalise students who did not choose option B. In fact, I would say that selecting option B should be penalised. ...
4. For Q2a, Given that it is clearly mentioned that the team has separated the first 800 and last 200 records for testing (1000 records in total), I believe this would have completely eliminated the risk of data leakage from test set into training set. As a result, no information from the test set can be used during the model training phase, meaning no data leakage can occur in this setup. Therefore, option (b) should not be considered a valid option at all.

For queries 3 – 4:

Yes, under non-sequential interpretation. Under the sequential interpretation of the approach, the normalization happens before the split. Thus, even if the split is mutually exclusive, the data is normalized using the statistics of the testing data along with the training data. This is the source of the leakage.

5. ... Similarly, option E - train-test split is problematic is not an issue. A split of **80-20 is actually considered a good split** for most machine learning applications. No evidence from the text provided suggests that we are not using a good train-test split. Please don't penalise those who did not choose option E. In fact, I would say that selecting option E should be penalised.

It depends on how the split is performed, among other factors. In most cases, an 80-20 random split is sufficient. The key here is the "random" aspect. This ensures that the data distribution in the 80% portion and the 20% portion is similar to the original 100% data, as it is sampled randomly from it.

6. The term train–test split is **commonly interpreted** in machine learning to mean the numerical proportion used to divide the dataset (e.g., 80/20, 70/30). Many students, including myself, understood the question in this standard way. While I acknowledge that ordering by timestamp affects the validity of the split, the phrase train–test split by itself refers to the chosen ratio, not the specific procedure used to partition the data. Because of this ambiguity in terminology, my interpretation was reasonable and consistent with common usage in ML practice. Therefore, I hope that my answer can be considered for marks on the basis of this reasonable interpretation of the term. Not choosing option E should not be penalised.

There is no need to interpret the question. The procedure to perform the split is explicitly stated in the question.

7. Q2A. E is not necessarily an issue. Since the data is ordered by discharge date, using the first 800 for training and the last 200 for testing reflects a realistic temporal setting: we train on past patients and evaluate on future ones. The split ratio (80/20) is standard. The answer scheme assumes that the first 800 data could be older patients while the subsequent 200 are younger, but in no way does the information in the context suggest this. As such, I think A,B,E should be an alternative combination of correct answers.

The question ask to identify potential issues. It is true that E may not be an issue. But it can potentially be the issue, i.e., there exists a possibility that E causes the poor performance, as illustrated in the example (we did not give any assumptions).

2B

1. For question 2B (examplify question 5), D should not be accepted as a valid answer. Or at least, please don't penalise those who did not choose option D, because nothing in the data suggests that using a logistic regression model without feature transformation will improve the performance of the model. Sure, we can experiment with this technique and MAYBE obtain better results, but nothing from the evidence indicates that this would perform better. Since no evidence provided points towards this, not selecting this option should not be penalised.
2. Agree. In the mentioned results, we only know there is overfitting, and the most obvious attempt to address that is to increase lambda (initially 0). I don't think we can infer that removing the feature transformation can help, in fact removing it might be hurtful if the data is not linearly separable with just the features without the transformation. A better option would have been to reduce the degree of the polynomial transformation, but since the option mentioned to remove the transformation entirely, I don't think D is a correct option
3. I believe Option D in Q5 (Replace with bare logistic regression without feature transformation) is not strictly necessary since it does nothing specific to address the class imbalance and also doesn't explicitly address the overfitting issue (does not include regularization). I would suggest that an alternative answer without option D also get accepted. Thank you.

For queries 1 – 3:

The question presents results of a machine learning model which shows that the machine learning model doesn't work well on the test set, for which overfitting might be the cause. We know that the model uses polynomial degree 10 with no regularization, which is likely the reason for the overfitting due to its model complexity.

One way to potentially improve the performance of the model is to reduce its model complexity. One way to reduce model complexity is to use less-complex features. One way to use less-complex features is by using the features as it is, i.e., no feature transformation.

Thus, to potentially improve the performance of the model, one can use no feature transformation.

4. 2E. Why is duplicating 'readmitted' data 10 times acceptable when it would just increase chance of overfitting for the readmitted class even if it does increase test accuracy.

Statement is contradictory:
increased test accuracy → reduced overfitting ≠ increased overfitting.

3B

1. For Q3B, I was too nervous and did not simplify to e^-1 and left it as 0.3e^-1 + 0.7e^-1. would this still be accepted?

Yes, this is acceptable.

4A

1. Question 4A, can [1, -1] be accepted as the normal vector also because it is just the opposite direction of the answer
2. For question 4A, can the normal vector be [1, -1], since the offset is 0 (my calculation could be wrong), so it would represent the same decision boundary?

For queries 1 – 2:

We cannot accept this answer as the question explicitly says which of the support vectors belong to the positive class and which of the support vectors belong to the negative class.

3. Question 4A, can partial mark (1m) be given if the sign is wrong (accidentally wrote 1, 1 instead of -1, 1)

Unfortunately, we cannot accept this answer as the vector points in the wrong direction.

4. Q4A: Can the normal vector be (1, -1) instead of (-1, 1)?

We cannot accept this answer as the question explicitly says which of the support vectors belong to the positive class and which of the support vectors belong to the negative class.

5. Hi Prof, for 4A (question 8 on Examplify), I put w = [-0.5, 0.5]^T as the normal vector, since this set of weights is true to allow bias = 0. May I check if my reasoning is sound? Thank you. It's written as a hint, not a compulsory thing. Plus

for w = [-0.5, 0.5]^T, we can allow our bias to be zero. If we set w = [-1, 1]^T, the bias used for positive and negative classes become inconsistent.

Edit: Consider the following formula for calculating bias using support vectors (https://stats.stackexchange.com/a/362448, https://stats.stackexchange.com/a/590195).

If we set w = [-1, 1]^T:

b = -1 - [-1, 1][2, 0]^T = 1

b = 1 - [-1, 1][0, 2]^T = -1 (contradiction)

If we set w = [-0.5, 0.5]^T

b = -1 - [-0.5, 0.5][2, 0]^T = 0

b = 1 - [-0.5, 0.5][0, 2]^T = 0

The question tells us to find w of the SVM, so I interpreted it as such.

Since the rescaling was only mentioned in the hint, the updated grading allows other vectors pointing in the same direction with different magnitude. No partial marks are possible.

6. will there be partial marks for semi-correct answer in the Question 4A lol

We cannot provide partial marks as the vector points in the wrong direction.

7. As someone mentioned earlier for 4A, I also wrote w = (-1/2, 1/2) because I interpreted w as the actual SVM weight vector, which is defined only up to scale. Since hints are typically optional and not part of the marking criteria, I did not treat the scaling instruction as compulsory. Additionally, the $\sqrt{2}$ formatting also appeared unusual, and I interpreted it as indicating $1/\sqrt{2}$ (so that margin lengths are maximised), so I focused on identifying the actual w of the SVM model rather than applying an explicit normalization. My answer corresponds to the same separating hyperplane and classification behaviour as the expected vector, differing only in scale. I would be grateful if this could be considered for credit. Thank you.

As the rescaling was only mentioned in the hint, the updated grading allows other vectors pointing in the same direction with different magnitude. No partial marks can be given.

8. Hi Prof, I would like to clarify answer for question Q8(4A).Since the question asks only for the normal vector w that defines the SVM hyperplane, and not for the classifier or the margin constraints, the answer is purely geometric. A hyperplane $w^T x+b=0$ is unchanged if both w and b are multiplied by −1, meaning that w and −w represent the same geometric hyperplane. Therefore, the vectors $[-1, 1]^T$ and $[1, -1]^T$ are geometrically equivalent normals with the same required norm $\sqrt{2}$. Given this equivalence and the fact that the question does not restrict the orientation of the normal or specify the bias term, $[1, -1]^T$ should also be accepted as a correct answer.

We cannot accept this option as the question explicitly says which support vectors belong to the positive class and which support vectors belong to the negative class.

9. For Q4A, I wrote [-1, -1] instead of [-1, 1] (possibly mistyped during exam). Is it possible to get partial mark at least for getting the first vector value right? Thank you.

No partial marks can be given here.

5A

1. Q11 5A. Correct cluster, wrong string, awarded 0/2 instead of ½
2. Q5A: I wrote c2 for Answer 2 but the system marked me wrong. I noticed some other people had the same issue as me, so it might be worth looking into this issue.
3. Q5A: I wrote c2 as the second answer but was marked wrong for it
4. Hi prof, can there be partial marks for q11 and q22 if we got one of the answers correct?
5. Hi Prof, for Q11 (Q5A) is it possible to make it partial marks if the first answer is correct? I feel like the first blank and second blank in Q11 are testing two different concepts and hence it doesn't feel right to penalise -2 if the first answer is correct but the only the second answer is wrong. Thank you!
6. For Q11, shouldnt the marks be independent, I got the coordinates of c1 right but cluster selection wrong, so 1/2 instead of 0/2 would seem more fair.
7. For question Q11 5A, answered correctly as c2 but it was detected as wrong.

For queries 1 – 7:

We have modified the grading to allow partial marks for this question. We allow a variety of inputs to designate centroid 2. But we do not allow the string “l2”.

5B

1. For question 5B (question 12 in examplify), option B should also be a correct answer. Why is it wrong?

Option b outputs c with the minimum dot product. This output is not a correct classification.

2. For Qn 5B, using MSE works iff we train the model to output values 1 to n where each output corresponds to one of the n classes and choose the class which is closest to one of the valud n classes. Here, in the description, it state, "Minimise MSE using linear model and targets until convergence". This implies that it has access to the targets! I also asked Prof Patrick this question during the exam because it wasn't explicitly written above if this option (A) had access to the true target values and he state to follow the description. So wouldn't that imply to assume it has targets?

Option a does not have targets available. The available data is specified under (D). For each option we are given a specific combination of components (D) (LA) (C). The availability of specific combinations highlights that sometimes a wrong component will make the whole not work together. (“combination of D LA C” in the question).

3. Given the assumptions in the question 5B: linearly separable data, infinitely many clean training examples, and no noise. The class centroids converge to stable prototypes that lie in distinct linear regions. In this setting, classification by minimizing $x^T\mu(c)$ is effectively equivalent to distance-based classification, because the decision rule $\|x-\mu(c)\|^2=x^Tx-2x^T\mu(c)+\|\mu(c)\|^2$ reduces to comparing dot-product terms when norms are stable. Thus, using centroid computation (LA) and dot-product-based classification (C) is sufficient to recover the correct class under the stated assumptions. Therefore, option b provides a valid and sufficient combination and should be accepted as a valid option.

Option b outputs c with the minimum dot product. This output is not a correct classification. It is not “effectively equivalent” for the Classifier to give the wrong output. In other words, while negative dot product is related to distance as is correctly pointed out, they are not the same (similarity measure vs. distance measure).

5D

1. Question 5D, can [9 0 -5]^T be accepted as a correct answer? As I assumed "data variance is reduced by removing variance corresponding to σ(2)" meant setting the value to 0 but still keeping that column

We cannot accept this answer as no compression happens then.

2. Can we appeal for leniency and for 1 mark to be awarded for each correct value (either 9 or -5), including responses given as 1D matrix (real number) of either 9 or -5, instead of it being an all or nothing system.

We allow for a variety of options for the output. However, partial marks are not possible. A single number would not be the correct compression the company wants.

7B

1. For Q19, can the answer be -16
2. For 7B, I understand that the mathematically correct derivative is +16. However, a student who obtained −16 has demonstrated the full backpropagation structure correctly. Identifying the computational graph, computing all intermediate derivatives correctly, and correctly applying the chain rule through both layers, but making only a single sign mistake at the final stage. Since the conceptual steps are all present and correct, this seems aligned with typical marking practice where minor arithmetic sign errors incur partial deduction rather than full penalization. Therefore, I would ask for at least partial marks given for students whose answer is -16.
3. For Q7B, my answer was -16, while the correct answer was 16. I correctly computed the magnitude of the gradient, indicating I performed the derivative steps correctly, but made a sign error. Could I receive partial credit for the correct derivation?

For queries 1 – 3:

An incorrect sign means the resulting weight update will move the model in the opposite direction of the true minimum, causing the loss function to increase rather than decrease. In this context, it is treated as a major error, partial credit cannot be awarded.

7D

REVISED ANSWER: a, b or a

1. option B: since question asks that "there exists a setting of weights", and we are not required to model all logistic model. Consider logistic model \sigma(ax_1+bx_2), where a,b != 0. We can set w_1=w_2=a, w_3=w_4=b. Next let w_5=2, w_6=-1. This will model the logistic regression perfectly

2. Hi Prof! For question 21, I believe option B should be possible as well. Consider a logistic regression model with non-zero weights a and b. Then a possible configuration where the given architecture follows a logistic regression model is w1 = w3 = a, w2 = w4 = b, w5 = w6 = 0.5. This gives the result of both the 2 neurons in the hidden layer to be the same as the result of the matching logistic regression model. Then taking the average of both would give the same final result.

3. Q7D option b: I think it is possible to represent 1 sigmoid output with 2 as long as they are the same and their coefficients add up to the same too

4. I think many others have already given explanations as to why option B is a correct option so I shall not write a grandmother's story here. As such, I think A,B should be the only combination of answers that receives full credit.

5. for 7d, option b, if w1 = w2, w3 = w4, we get the same sigmoid function from both, then we js set both w5 and w6 to be 0.5, wont that give a singular sigmoid function?

For queries 1 – 5:

To represent logistic regression model: $\sigma(ax_1 + bx_2)$

If $a \neq 0$ and $b \neq 0$, the following are valid settings:

- $w_1 = w_2 = a, w_3 = w_4 = b\ , w_5 = 2, w_6 = -1$
- $w_1 = w_2 = a, w_3 = w_4 = b\ , w_5 = 0.5, w_6 = 0.5$

However, if $a = 0$ and $b \neq 0$ (or $a \neq 0$ and $b = 0$), the above settings will not work, under the all non-zero weights assumption.

Because the question does not clearly specify the full details of the logistic regression model, we accept either selecting both Option (a) and Option (b), or selecting just Option (a).

6. For question 7D, Option c (The output value is between 0 and 2, regardless of the model weights and inputs): This should be correct as even the answer states that "The values $a1$and $a2$ are between 0 and 1", which is in fact between 0 and 2.

a_1 and a_2 are between 0 and 1, and the weights w_5 and w_6 can be any value. The output is $\hat{y} = w_5 a_1 + w_6 a_2$. For example, if $a_1 = a_2 = 0.5$, and $w_5 = 50, w_6 = 100$, then $\hat{y} = 75$, which is not between 0 and 2.

7. For Q7D, Shouldn't Option (a) be incorrect because logistic regression has an output of the form $\sigma(w_1x_1 + w_2x_2)$ (a single sigmoid applied to a linear combination to give output). In contrast, the given neural network has sigmoid function inside the hidden layer but an identity activation function at the output layer, so its output becomes a linear combination of sigmoid activations w1(2) σ(z1) + w2(2)σ(z2) and this is different from the sigmoid of a linear function. Even if the NN weights are set to 0 and 1, the 2 models are still not identical or the same. Hence, I think Option (d) can also be allowed as a valid option.

In option (a), the phrase "act exactly like a logistic regression model" is key. Acting identically does not equal being identical. In the context of this course, we discuss agents performing actions, and acting identically means generating the same output given the same input. The given neural network is powerful enough to act exactly like a logistic regression model, even though it is not architecturally the same model. Therefore, option (a) is correct.

8A

REVISED: 1 mark for each blank.

1. Hello Prof, for Q22 (8A) I got the first part of the question correct and the second part wrong but I received 0/2 for the question instead of 1/2. Is this an error with the marking?
2. For question 22, my answer 1 (output height) is wrong, but answer 2 is correct (the number of channels). But I was given 0/2. The two options are independent, so I would like to appeal to be given 1 mark for my correct answer.
3. Hi prof, can there be partial marks for q11 and q22 if we got one of the answers correct?
4. Can there be partial marks for Q8A? For getting 1 of the numbers right

For queried 1 - 4:

The grade has now been updated to reflect the separate marking of the two blanks.

8E

1. For question 8E, I think the option B "The fully-connected layer should contain 32 neurons." should also be considered correct since the question never specifies whether the number of neurons in the fully connected layer refers to the input or the output.

The neurons in a fully-connected layer are the computational units whose generated values form the layer's output. The input size to the layer is determined solely by the output dimension of the preceding layer, not by the specified number of neurons in the current fully-connected layer.

9C

REVISED ANSWER: c or d (Selecting c gets full mark; Selecting d also gets full mark)

1. For question 29, the answer given is "> C: The information extracted from the first 10 frames of the video.". But at time step 10, the hidden state only contains information from the first 9 frames, as the 10th frame is still being processed. Exactly this was mentioned in the lecture too. In fact, I can pinpoint the exact timestamp of the lecture (Lecture 27/10/2025, timestamp 01:08:22). So "None of the above" should be accepted. In fact, the option C is wrong and should be rejected.
2. The question felt vague because it did not say whether “hidden state at time step t = 10” refers to the hidden state used at time 10 (before processing frame 10) or the hidden state computed at time 10 (after processing frame 10). Since the lecture notes describe the hidden state going into time t as summarising only the earlier frames, I think it is reasonable to argue that “None of the above” could also be considered a correct answer. As such, I think E should be an alternative response that receives full credit.

For queries 1 and 2:

The question does not clearly specify whether the "hidden state" refers to the input hidden state or the generated hidden state. Therefore, we will accept both options (c) and (d) as correct.

9E

1. Isn't RNN input meant to take in 2 vectors? The question asked for "the input"

The question asked for the input to the last fully-connected layer, so the correct answer is option (b).

10B

1. What is the reasoning behind this answer for 10B? Given the context, I was assuming that it is an input of 5 and output of 5, as a series of labels for each input is produced. As the question asks for a many-to-many neural network, how is outputting a single classification after consolidating all the values a viable output? Each individual input in the sequence should have its own neuron outputting its own classification, so I don't see how having only one neuron for all 5 inputs is possible?

2. For question 10B, I feel the question is ambiguous because it does not specify whether the fully connected output layer is applied per timestep or only once. Under a valid interpretation where one fully-connected output layer corresponds to the entire sequence, options (a) and (b) become invalid, leading to (d). Therefore, please consider partial credit.

For queries 1 and 2:

Please refer to the many-to-many attention neural network discussed in Lecture 11, Slide 41. The fully-connected layer is applied individually to the output of the attention layer at each time step. It is not applied once to the whole sequence.

10C

1. Question 10C, can Cross Attention Layer be accepted such that previous positions are kept by the cross attention layer. Because I thought for mask self attention, we would still need to know the values for future positions except we set the attention value manually, whereas in a real time system we dont have the values for the future positions, so it doesn't make sense to use a mask self attention layer since we dont have the future values for us to mask.

Given the data provided (described in the context) for creating the real-time anomaly detection system, we need ensure that during model training, the input does not attend to future elements. Therefore, we need the masked self-attention layer.

10D

REVISED ANSWER: c or d (Selecting c gets full mark; Selecting d also gets full mark)

1. For question 10D i would like to argue option C is valid (If the company decides to collect sensor readings of length 7, the trained neural network described in the question can be directly employed to generate the labels.), my logic is without changing the model or weights the model can still be used on a subset of the 7 weights to get the labels in two predictions

The question does not clearly specify the meaning of “directly employed”. Therefore, we will accept both options (c) and (d) as correct.

2. I would like to request partial credit for option 10D(b) because the wording “the predicted probability … is the same regardless of model weights” is ambiguous. The intended meaning is “for all possible weight configurations,” but it can also reasonably be interpreted as “independent of the particular weights learned during training,” in which case the statement might appear true for identical values in the same input sequence. Since the phrasing does not explicitly state the strict mathematical interpretation and can mislead students, I believe partial credit is justified.

A statement about a property being "regardless of weights" implies that, whatever the values of the weights are, they cannot cause the output for two identical inputs to be different. Regarding your interpretation, the word "particular" should be removed as it is not present in the original question's phrasing. With this removal, the resulting phrase, "independent of the weights learned during training" still leads to the conclusion that option (b) is incorrect.