National University of Singapore

CS2109S—Introduction to AI and Machine Learning

Final Assessment

Semester 1, 2025/2026

Time allowed: 2 hours

Instructions:

1. Please place your student card or identification document (NRIC, driving license, etc.) on the top right-hand corner of your desk.
2. Please switch off your personal devices with communication features and leave them on the floor next to your desk at all times.
3. If you wish to communicate with an invigilator, go to the washroom, or leave before the end of the assessment, please raise your hand to inform the invigilator.
4. Please follow the other instructions in Examplify.
5. This paper contains the context for the questions in Examplify.
6. This paper contains **Thirty-Two (32) pages** including this cover page.
7. This paper should not be submitted.
8. All questions must be answered in Examplify.
9. You may refer to the appendix provided in Examplify.

	Number of questions	Total marks
Intelligent Agents	3	4
ML Case Study	2	4
Kernel Method	2	4
Support Vector Machine	3	6
Unsupervised Learning	4	7
Neural Network I	3	6
Neural Network II	4	7
Convolutional Neural Network	5	8
Recurrent Neural Network	5	7
Attention Neural Network	4	7
Total	**35**	**60**

This page is intentionally left blank.

It may be used as scratch paper.

Part 1: Intelligent Agents (3 questions, 4 marks)

Context

A large corporation uses an AI agent, the Automated Document Sorter (ADS), to manage its internal digital filing system. The ADS agent's sole task is to process documents from a single “INBOX” folder and move them into the correct predefined subfolders: "Finance," “HR”, “Legal”, or “Marketing”.

The agent operates in a series of self-contained tasks.

For each task, it processes one document.

ADS reads the full content of the document to identify exact matching keywords that determine its category. For example, keywords such as "invoice" mean Finance; keywords such as "contract" mean Legal. If there are no exact matching keywords, then the agent will simply leave the document in the “INBOX” folder.

The agent's only actions are to move the current document from the “INBOX” to one of the four category folders or to do nothing. These are five different actions. Each action is guaranteed to produce the desired outcomes. After working with one document, the agent will automatically move on to the next document. The agent cannot move to a document on its own volition.

At timestep 0, all documents are available in the “INBOX” folder. The environment is stable. The documents in the “INBOX” do not change or get modified by anything else apart from the agent.

The agent's performance is judged on one simple metric. The metric is whether the document is placed in the correct folder. There are no other considerations like speed, resource use, or other complex trade-offs. The process for sorting one document has no bearing on the process for sorting the next.

Based on this description, answer questions 1A-1C.

Questions

1A. [2 marks] Consider the list of options below. Select all the options that correctly describe a component of the PEAS framework for the ADS agent.

a. Performance: The number of documents processed per hour.
b. Performance: Whether the document is placed in the correct folder.
c. Environment: A digital file system with an INBOX and four category subfolders.
d. Environment: A dynamic network where documents can be modified by other users.
e. Actuators: A printer to create hard copies of sorted documents.

f. Actuators: The ability to move a document.
g. Sensors: The ability to read the text content of a document.

1B. [1 mark] Which **agent type** is the **simplest** and most **appropriate** for this robot? Note: agent types sorted from simplest to complex: simple-reflex-, model-, goal-, and utility-based agent.

a. Simple reflex agent
b. Model-based reflex agent
c. Goal-based agent
d. Utility-based agent

1C. [1 mark] Classify the **environment properties** from the description. Select all that apply.

a. Fully observable, not partially observable
b. Stochastic, not deterministic
c. Episodic, not sequential
d. Static, not dynamic
e. Continuous, not discrete

Answers

1A. b, c, f, g

Correct options:

- b. Performance: Whether the document is placed in the correct folder. This matches the stated single performance metric.
- c. Environment: A digital file system with an INBOX and four category subfolders. This is exactly the operating environment described.
- f. Actuators: The ability to move a document. Moving documents between folders is the agent's only action.
- g. Sensors: The ability to read the text content of a document. The agent reads documents to detect exact keywords.

Incorrect options:

- a. Number processed per hour is explicitly not part of the performance measure.
- d. The environment is stable; documents are not modified by others.

- e. A printer is not part of the agent's actuators for this task.

1B. a

The agent applies fixed rules based on exact keyword matches.

1C. a, c, d

Correct:

- a. Fully observable: The agent reads the entire document; all relevant information for classification is available.
- c. Episodic: Each document is processed independently; outcomes don't affect the next task.
- d. Static: The environment doesn't change except by the agent's actions.

Not correct:

- b. Stochastic: It is deterministic (guaranteed outcomes).
- e. Continuous: The state and actions are discrete (finite folders, five actions, timestep processing).

Part 2: ML Case Study (2 questions, 4 marks)

Context

A data science team at a hospital is building a machine learning model to predict whether a patient will be re-admitted after 30 days. These patients are called "readmitted patients".

- Objective: To identify high-risk patients.
- Dataset: The team uses a dataset of 1,000 patient records. The target variable is "readmitted", where the target is 1 if readmitted, 0 if not. In the dataset, 5% of the patients are "readmitted". The dataset is sorted by discharge date. The first record of the dataset is the earliest patient that is discharged from the hospital.
- Features: The features include age (20-95), length_of_stay (1-14 days), lab_test_results (a normalized score from 0-1), and number_of_medical_procedures (0-6).
- Model: Logistic regression model
- Features: Polynomial degree 10
- Regularization: L1 regularization with regularization parameter $\lambda = 0$.
- 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.

Based on this description, answer questions 2A-2B.

Questions

2A. [2 marks] Based on the experimental setup described, identify potential issues. Select all that apply.

a. Unaddressed class imbalance in the target variable (readmitted).
b. Data leakage from the test set into the training set.
c. Risk of the model overfitting the training data.
d. Risk of the model underfitting and failing to capture complex patterns.
e. Training-testing split is problematic.
f. None of the above

2B. [2 marks] Suppose that the team trains the model and obtains the following results.

The positive class is readmitted = 1.

Results:

- Training Accuracy: 99.8%
- Testing Accuracy: 75.5%
- Test Set Confusion Matrix (Positive = Readmitted):
    - True Positives: 1
    - False Negatives: 9
    - True Negatives: 190
    - False Positives: 0

Given these results, what should the team do to **potentially** improve the performance of the model? Select all that apply.

a. Increase the regularization coefficient of the model.
b. Decrease the regularization coefficient of the model.
c. Duplicate the data for the readmitted class 10 times.
d. Replace the model and features with Logistic Regression without feature transformation.
e. Use a random 75/25 split for training and testing data.
f. None of the above

Answers

2A. a,b,c,e

Correct:

- a. Class imbalance is unaddressed: Only 5% of patients are readmitted. If you don't handle this, the model can just predict "not readmitted" most of the time and still look accurate, but it won't catch the cases you care about.
- b. Data leakage: They normalized all the data before splitting into train and test. That lets information from the test set "bleed" into training via the normalization stats, making results look better than they should.
- c. Overfitting risk: Using degree-10 polynomial features with no regularization ($\lambda=0$) makes the model very complex. It can memorize the training data instead of learning the patterns.
- e. Problematic train/test split: Taking the first 800 by discharge date and the last 200 for testing can introduce distribution mismatch (e.g., changes in patient compositions). For example, suppose that the earliest 800 discharges contain patients with age >65 and the latter 200 discharges contain patients with age < 35. The model trained on mostly older patients may not capture risk patterns for younger patients.

Not correct:

- d. Underfitting isn't the likely problem here because the model is extremely complex.

2B. a,c,d,e

Correct:

- a. Increase regularization: Based on the train and test results, the model clearly overfits the data. Regularization helps reduce overfitting.
- c. Duplicate (oversample) the readmitted class: Because readmitted patients are rare, oversampling them helps the model pay attention to these cases.
- d. Use Logistic Regression without high-degree feature transformations: A simpler feature set (no degree-10 polynomials) reduces the chance of overfitting and may generalize better.
- e. Fix the problematic train/test split

Not correct:

b. Decreasing regularization would worsen overfitting (it's also already zero!).

Part 3: Kernel Method (2 questions, 4 marks)

Context

Consider a regression problem where we want to train a model to predict $y \in \mathbb{R}$. The input data have two features: x_1 and x_2. There is no dummy feature x_0 (no bias term). We adopt a kernel method with a dual model:

$$h_\alpha(x) = \sum_{i=1}^{N} \alpha_i \, K\left(x^{(i)}, x\right)$$

where x_i are training points, and α_i are linear-combination coefficients. The model uses a linear kernel $K(u, v) = u \cdot v$, which corresponds to the feature map $\phi(x) = x$.

Training Data:

- $x^{(1)} = [1,1]^T,\ \alpha_1 = 0.3$
- $x^{(2)} = [2,2]^T,\ \alpha_2 = 0.7$

Based on this description, answer questions 3A-3B.

Questions

3A. [2 marks] Suppose that we derive a new kernel $K_{new}(u, v) = 4 \times K(u, v)$ which corresponds to a new feature map $\phi_{new}(x)$. Compute the output of $\phi_{new}(x)$ for a new data $x = [1,2]^T$.

Fill the output into the following blanks:

Output = [____, ____]^T

3B. [2 marks] Suppose that we replace the kernel with Gaussian kernel $K_{RBF}(u, v) = e^{-\frac{||u-v||^2}{2\sigma^2}}$ with the variance parameter $\sigma^2 = \frac{1}{2}$. Compute the output $h_\alpha(x)$ for a new data $x = [1,2]^T$. Note: You may write the answer in terms of e.
Fill the output into the following blank:
$h_\alpha(x)$ = ____

Answers

3A. $[2,4]^T$.

Since $K_{new}(u, v) = 4 \times K(u, v)$ for the linear kernel, a corresponding feature map is $\phi_{new}(x) = 2x$. For $x = [1,2]^T$, the output is $[2,4]^T$.

3B. e^{-1}

$x^{(1)} - x = [1,1]^T - [1,2]^T = [0,1]^T$

$\left\|x^{(1)} - x\right\|^2 = 0^2 + 1^2 = 1$

$K_{RBF}\left(x^{(1)}, x\right) = e^{-1}$

$x^{(2)} - x = [2,2]^T - [1,2]^T = [1,0]^T$

$\left\|x^{(2)} - x\right\|^2 = 1^2 + 0^2 = 1$

$K_{RBF}\left(x^{(2)}, x\right) = e^{-1}$

$h_\alpha(x) = 0.3 \times e^{-1} + 0.7 \times e^{-1} = e^{-1}$

Part 4: Support Vector Machine (3 questions, 6 marks)

There is no context, please answer the questions directly.

Questions

4A. [2 marks] Consider a trained SVM. Let the support vectors of the positive class +1 be given by $[-3,-1]^T$ and $[0,2]^T$. Let the support vectors of the negative class -1 be given by $[2,0]^T$. Find the normal vector w of the SVM.

Hint: Consider the offset of the SVM first. Scale the normal vector such that its length equals $\sqrt{2}$.

Fill the normal vector w into the following blanks:

w = [____, ____]^T.

4B. [2 marks] You are given the following data set.

Data point	Feature 1	Feature 2	Label
$x^{(1)}$	1	-1	+1
$x^{(2)}$	1	0	+1
$x^{(3)}$	2	0.5	+1
$x^{(4)}$	2	-1	-1
$x^{(5)}$	3	-1	-1
$x^{(6)}$	0.5	-2	-1

You train a Support Vector Machine until convergence. During the training, the algorithm is allowed to shift the offset of the hyperplane. Select the resulting support vector(s). Select all that apply.

a. $x^{(1)}$
b. $x^{(2)}$
c. $x^{(3)}$
d. $x^{(4)}$
e. $x^{(5)}$
f. $x^{(6)}$

4C. [2 marks] Consider the polynomial kernel function $k(u,v) = (u^T v)^2$, for u, v two vectors. You have found an SVM using the kernel and using a training set of vectors given in the table below. You have obtained dual coefficients, see table.

Data point	Feature 1	Feature 2	Dual coefficient
$x^{(1)}$	1	1	0.75
$x^{(2)}$	2	2	0
$x^{(3)}$	1	-1	-0.75
$x^{(4)}$	2	-1	0
$x^{(5)}$	1	-2	0

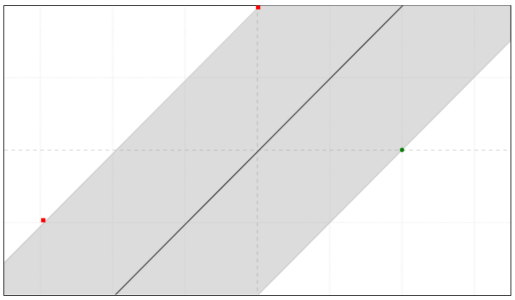
Select the prediction for a new point $x = [-2,1]^T$ using the dual SVM and the kernel function.

Hint: The dual SVM classifier is given by $h(x) = sign(\sum_j \alpha_j \ k(x^{(j)}, x))$.

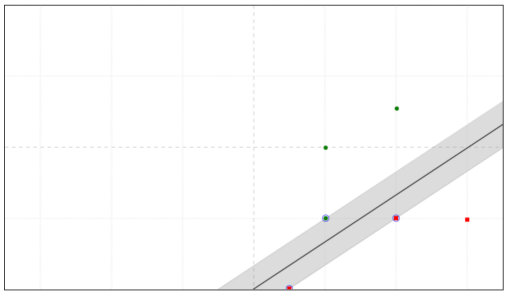
a. +1
b. -1
c. 0
d. Cannot be determined.

Answers

4A. -1,1



4B. a, d, f

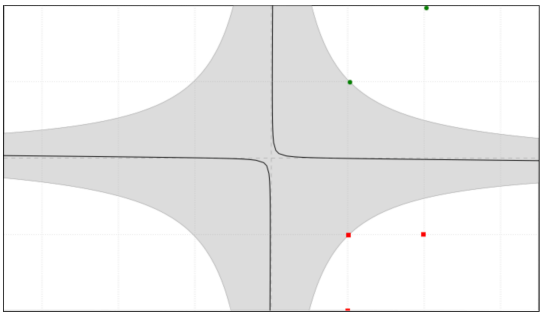


4C. b

$$k\left(x^{(1)}, x\right) = \left(x^{(1)^T} x\right)^2 = 1,$$

$k(x^{(3)}, x) = \left(x^{(3)^T} x\right)^2 = 9$

$\sum_j \alpha_j \ k(x^{(j)}, x)\ ) = 0.75(1 - 9) = -6$. Hence, the point is classified as -1.



Part 5: Unsupervised Learning (4 questions, 7 marks)

There is no context, please answer the questions directly.

Questions

5A. [2 marks] Consider the following dataset consisting of three points in 2D space:

$x^{(1)} = [2, 0]^T$

$x^{(2)} = [1, 1]^T$

$x^{(3)} = [3, 2]^T$.

Compute the centroid c1 of these three points and fill in the following blank:

c1 = ____.

Let another point be $x^{(4)} = [0, 0.5]^T$ and another centroid be $c2 = [-1, -1]^T$. Determine the assignment to cluster 1 or cluster 2 of the point $x^{(4)}$ based on the Euclidean distance.

Fill in the following blank with the string "c1" or "c2":

Assignment = ____.

5B. [2 marks] You would like to construct a classifier based on ideas from distance-based models and/or unsupervised learning. The classifier should be applied to a family of learning problems. The family of problems always has data that originates from 5 classes. Choose options that provide a **sufficient** combination of Data (D), Learning Algorithm (LA), and Classification (C). Here, "sufficient" means that the combination of D, LA, and C outputs the correct classification for new data.

Assume that the training and new data is linearly separable. Assume that many training examples $n \to \infty$ are given. And assume that no outliers/noise are present. Classification is performed for a new data point x without noise.

Select all settings given in the following options that apply.

a. D: n data points $x^{(j)}$ for all $j \in 1 \ldots n$.
   - LA: Minimize MSE using linear model and targets until convergence.
   - C: Choose minimum distance to the targets. In case of a tie, select the smallest target index among the ties.

b. D: n data points $x^{(j)}$ and true $y^{(j)} \in 1 \ldots 5$, for all $j \in 1 \ldots n$.

- LA: Compute centroids $\mu^{(c)}$ for all $c \in 1 \ldots 5$, using the corresponding data $x^{(j)}$ where $y^{(j)} = c$.
- C: Output the c that minimizes $x^T \mu^{(c)}$, for all $c \in 1 \ldots 5$. In case of a tie, select the smallest c among the ties.

c. D: n data points $x^{(j)}$ and true $y^{(j)} \in 1 \ldots 5$, for all $j \in 1 \ldots n$.
   - LA: None.
   - C: Compute $d^{(j)}(x) = \left|\left|x^{(j)} - x\right|\right|_2^2$ for all $j \in 1 \ldots n$. Determine the j values of the 10 smallest values $d^{(j)}(x)$. Determine the values $y^{(j)}$ of those j's. Output the corresponding $y^{(j)}$ of the majority value $y^{(j)}$ of those j's. In case of a tie, output the $y^{(j)}$ of the smallest j among the ties.

d. D: n data points $x^{(j)}$ for all $j \in 1 \ldots n$.
   - LA: A single k-means clustering with k = 2.
   - C: Output the cluster with minimum distance of its centroid to x. In case of a tie, select the smallest cluster index among the ties.

e. None of the above.

5C. [1 mark] Assume you have 41 samples used for the X^T transpose of the mean-centred **data** matrix. An SVD of X^T obtains a first singular value $\sigma_1 = 2$. What is the unbiased sample variance associated with this singular value?
Fill in the blank with the fixed-point number with two digits of precision.
Variance = ____

5D. [2 marks] In a company, SVD is used on the transpose of the data matrix of mean-centred customer data before employing a logistic regression model. The following left singular vectors are obtained.

$$u^{(1)} = \frac{1}{\sqrt{2}}[1, 0, 1]^T$$

$$u^{(2)} = [0, 1, 0]^T$$

$$u^{(3)} = \frac{1}{\sqrt{2}}[1, 0, -1]^T$$

SVD obtains the singular values $\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}$.

According to company-specific guidelines, data variance is reduced by removing variance corresponding to $\sigma^{(2)}$. You are presented with a new mean-centered customer data point: $x = \left[2\sqrt{2}, 6, 7\sqrt{2}\right]^T$.

You are asked to compress the data before using it as the input for the logistic regression model. What is the compressed version of the customer data? Fill your answer in the following blank. Please simplify your expression and evaluate your expression numerically, if it is possible based on the available information.

Customer data = ____

Answers

5A: [2,1]^T, and “c2”

Compute the centroid as: $c1 = \frac{1}{3}\left(x^{(1)} + x^{(2)} + x^{(3)}\right) = \frac{1}{3}([2,0]^T + [1,1]^T + [3,2]^T) = [2,1]^T$
Compute the Euclidean distance of $x^{(4)} = [0,0.5]^T$ to c1 and $c2 = [-1,-1]^T$. First, $d_{c1}^2 = 2^2 + 0.5^2 = 4.25$. Second, $d_{c2}^2 = 1 + 1.5^2 = 3.25$. Hence, fill “c2” into the blank.

5B: c

Option a is wrong, as there are no targets given in the data (we only know that the number of classes = 5). We cannot construct the MSE, and also minimum distance to the targets is non-sensical.

Option b is wrong, as the classification uses the dot product. The dot product is a measure of similarity, in contrast to a distance measure like the Euclidean distance. Minimum dot product will select dot products proportional to -1, or 0 if we minimize the absolute value, which are opposing or orthogonal vectors.

Option c is correct. In fact, this option is related to the k nearest neighbor classifier.

Option d is incorrect, as a single run of k-means (to convergence) with k = 2 will not produce a classifier for 5 classes.

5C: 0.1

The variance is computed by $var_1 = \frac{\sigma_1^2}{n-1} = 0.1$. Using the biased variance is also $\frac{\sigma_1^2}{n} = 0.1$, when we use two digits of precision. We accept also the input 0.098.

5D: [9,-5]^T

We are asked to remove the variance corresponding to $\sigma^{(2)}$ and to compress the data point. Compression means that we do not consider the basis vector $u^{(2)}$ and we compress the data using the remaining basis vectors. We compress by computing the dot product of the data with the remaining basis vectors. It gives: $(u^{(1)})^T x = 2 + 7 = 9$ and $(u^{(3)})^T x = 2 - 7 = -5$. Hence, fill into the blank “[9,-5]^T”, as the compressed data should be a column vector. We accept slightly other statements of the solution as well.

Part 6: Neural Network I (3 questions, 6 marks)

Context

You are given the following dataset for a binary classification task:

x_1	x_2	y
-1	-1	1
1	1	1
2	1	1
1	0	0
2	0	0
0	1	0

Using this dataset, answer questions 6A-6C.

Questions

6A. [2 marks] Without adding any handcrafted features, i.e., directly using only the original features x_1 and x_2 (no bias), which of the following models can generate a decision boundary that correctly classifies all 6 data points? You can assume an appropriate decision threshold is used. Select all that apply.

a. A neural network with a single neuron. Identity activation function is applied.
b. A neural network with a single neuron. Sigmoid activation function is applied.
c. A neural network with two fully-connected layers. Identity activation function is applied in first fully-connected layer, and sigmoid activation function is applied in second fully-connected layer.
d. None of the above.

6B. [2 marks] With the addition of a handcrafted feature $x_3 = x_1 x_2$, and using x_1, x_2 and x_3 (no bias), which of the following models can generate a decision boundary that correctly classifies all 6 data points? You can assume that an appropriate decision threshold is used. Select all that apply.

a. A neural network with a single neuron. Identity activation function is applied.
b. A neural network with a single neuron. Sigmoid activation function is applied.
c. A neural network with two fully-connected layers. Identity activation function is applied in first fully-connected layer, and sigmoid activation function is applied in second fully-connected layer.
d. None of the above.

6C. [2 marks] With the addition of a handcrafted feature $x_3 = 5x_1 + 3x_2$, and using x_1, x_2 and x_3 (no bias), which of the following models can generate a decision boundary that

correctly classifies all 6 data points? You can assume that an appropriate decision threshold is used. Select all that apply.

a. A neural network with a single neuron. Identity activation function is applied.
b. A neural network with a single neuron. Sigmoid activation function is applied.
c. A neural network with two fully-connected layers. Identity activation function is applied in first fully-connected layer, and sigmoid activation function is applied in second fully-connected layer.
d. None of the above.

Answers

6A: d

Option a: The output is $\hat{y} = w_1 x_1 + w_2 x_2$. The decision boundary is $w_1 x_1 + w_2 x_2 = \text{threshold}$. This is the equation of a straight line.

Option b: The output is $\hat{y} = \sigma(w_1 x_1 + w_2 x_2)$. The decision boundary is $\hat{y} = threshold = \sigma(c)$, so the boundary is $w_1 x_1 + w_2 x_2 = c$. This is also the equation of a straight line.

Option c: The output of the first layer is a set of linear combinations of the input (e.g., $a_1 = W_{11}^{[1]} x_1 + W_{21}^{[1]} x_2$, $a_2 = W_{12}^{[1]} x_1 + W_{22}^{[1]} x_2$, etc.).To generate the output, a linear combination of the first layer's outputs needs to be computed (e.g., $f = W_{11}^{[2]} a_1 + W_{21}^{[2]} a_2$). If you substitute the a_1, a_2 equations into the f equation, you get:

$$f = W_{11}^{[2]} \left(W_{11}^{[1]} x_1 + W_{21}^{[1]} x_2 \right) + W_{21}^{[2]} \left(W_{12}^{[1]} x_1 + W_{22}^{[1]} x_2 \right)$$

$$= \left(W_{11}^{[2]} W_{11}^{[1]} + W_{21}^{[2]} W_{12}^{[1]} \right) x_1 + (W_{11}^{[2]} W_{21}^{[1]} + W_{21}^{[2]} W_{22}^{[1]}) x_2.$$

This entire network collapses into a single neuron with a sigmoid activation. Its decision boundary is just a straight line.

Since the dataset is not linearly separable and all three models are linear classifiers, none of them can correctly classify all 6 points.

6B: a, b, c

With x_3, we have

x_1	x_2	x_3	y
-1	-1	1	1
1	1	1	1
2	1	2	1
1	0	0	0
2	0	0	0
0	1	0	0

The dataset now is linear separable, all three models can correctly classify all 6 data points.

6C: d

The new handcrafted feature is a linear combination of x_1, x_2.

Option a: The output is $\hat{y} = w_1x_1 + w_2x_2 + w_3(5x_1 + 3x_2)$. The decision boundary is $(w_1 + 5w_3)x_1 + (w_2 + 3w_3)x_2 = threshold$.

Option b: The output is $\hat{y} = \sigma(w_1x_1 + w_2x_2 + w_3(5x_1 + 3x_2))$. The decision boundary is $\hat{y} = threshold = \sigma(c)$, so the boundary is $(w_1 + 5w_3)x_1 + (w_2 + 3w_3)x_2 = c$.

Option c: The output of the first layer is a set of linear combinations of the input (e.g., $a_1 = W_{11}^{[1]}x_1 + W_{21}^{[1]}x_2 + W_{31}^{[1]}(5x_1 + 3x_2)$, $a_2 = W_{12}^{[1]}x_1 + W_{22}^{[1]}x_2 + W_{32}^{[1]}(5x_1 + 3x_2)$, etc.).To generate the output, a linear combination of the first layer's outputs needs to be computed (e.g., $f = W_{11}^{[2]}a_1 + W_{21}^{[2]}a_2$). If you substitute the a_1, a_2 equations into the f equation, you get:

$$f = W_{11}^{[2]}\left(W_{11}^{[1]}x_1 + W_{21}^{[1]}x_2 + W_{31}^{[1]}(5x_1 + 3x_2)\right) + W_{21}^{[2]}\left(W_{12}^{[1]}x_1 + W_{22}^{[1]}x_2 + W_{32}^{[1]}(5x_1 + 3x_2)\right)$$

$$= \left(W_{11}^{[2]}W_{11}^{[1]} + W_{21}^{[2]}W_{12}^{[1]} + 5W_{11}^{[2]}W_{31}^{[1]} + 5W_{21}^{[2]}W_{32}^{[1]}\right)x_1 + (W_{11}^{[2]}W_{21}^{[1]} + W_{21}^{[2]}W_{22}^{[1]} + 3W_{11}^{[2]}W_{31}^{[1]} + 3W_{21}^{[2]}W_{32}^{[1]})x_2.$$

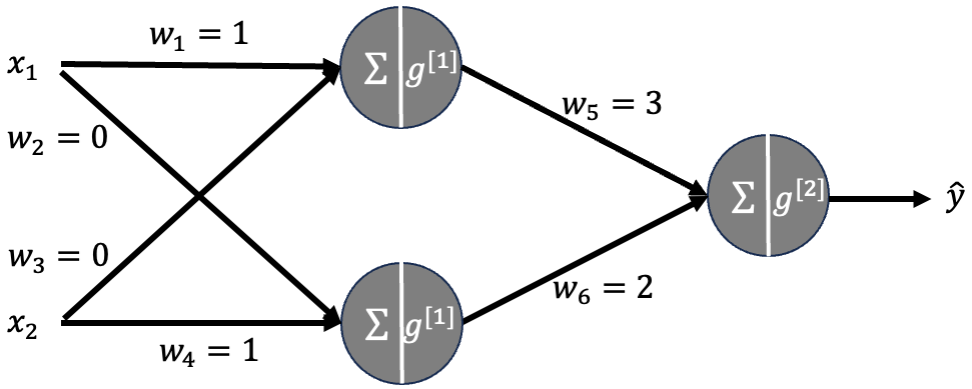
This entire network collapses into a single neuron with a sigmoid activation.

All three models are linear classifiers in the original (x_1, x_2) space, and the original dataset is not linearly separable, none of them can correctly classify all 6 points.

Part 7: Neural Network II (4 questions, 7 marks)

Context

Consider the neural network shown below, which has no bias term in its neurons.



The activation functions for the first and second fully-connected layers are $g^{[1]}$ and $g^{[2]}$, respectively. Both $g^{[1]}$ and $g^{[2]}$ are identity functions.

You are tasked with training the neural network with the loss function set as:

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2$$

where n is the number of data samples.

Based on this description, answer questions 7A-7D.

Questions

7A. [1 mark] Given a data sample with two features: x_1 =-1 and x_2 =-1, what is the value for predicted output? Fill your answer in the following blank.
Predicted output = ____

7B. [2 marks] Suppose you need to train the neural network with a single data sample (x_1 =-1, x_2 =-1, y=3). The current weights can be found in the context. What is the partial derivative of the loss function with respect to w_4? Fill your answer numerically in the following blank.
$\frac{\partial J(w)}{w_4}$= ____

7C. [2 marks] Suppose $g^{[1]}$ is now updated to ReLU activation function and $g^{[2]}$ is still identity function. Which of the following is correct? Select all that apply.

a. The partial derivative of the loss function with respect to w_3 is always 0, regardless of the model weights and inputs.
b. Given features x_1 and x_2 with arbitrary negative values, the final output is always 0, regardless of the model weights.
c. The network computes a linear transformation of the input.
d. None of the above.

7D. [2 marks] Suppose $g^{[1]}$ is now updated to sigmoid activation function and $g^{[2]}$ is still identity function. Which of the following is correct? Select all that apply.

a. If the weights are set properly, the neural network can act exactly like a logistic regression model, as long as they both use the same two features and the logistic regression model has no bias.
b. There exists a setting of weights with all non-zero values, such that the neural network can represent a logistic regression model that takes x_1 and x_2 as input and no bias.
c. The output value is between 0 and 2, regardless of the model weights and inputs.
d. None of the above

Answers

7A. -5

Output:

$$z_1 = w_1 x_1 + w_3 x_2 = 1 \times (-1) + 0 \times (-1) = -1$$

$$a_1 = g^{[1]}(z_1) = z_1 = -1$$

$$z_2 = w_2 x_1 + w_4 x_2 = 0 \times (-1) + 1 \times (-1) = -1$$

$$a_2 = g^{[2]}(z_2) = z_2 = -1$$

$$\hat{y} = w_5 a_1 + w_6 a_2 = 3 \times (-1) + 2 \times (-1) = -5$$

7B. 16

Partial derivative:

$$\frac{\partial J(w)}{\partial w_4} = \frac{\partial J(w)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial w_4} = (\hat{y} - y) w_6 \cdot 1 \cdot x_2 = (-5 - 3) \times 2 \times (-1) = 16$$

7C. d

Option a: Incorrect. $\frac{\partial J(w)}{\partial w_3} = \frac{\partial J(w)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial w_3} = (\hat{y} - y) w_5 \cdot 1_{z_1 > 0} \cdot x_2$. The partial derivative is not always 0

Option b: Incorrect. The weights can be negative. For example, if $x_1 = -1, x_2 = -1$ and we set $w_1 = -1$, $w_4 = -1$ and keep the values of other weights. Then, we have:

$$z_1 = w_1 x_1 + w_3 x_2 = -1 \times (-1) + 0 \times (-1) = 1$$

$$a_1 = g^{[1]}(z_1) = \mathrm{ReLU}(z_1) = 1$$

$$z_2 = w_2 x_1 + w_4 x_2 = 0 \times (-1) + (-1) \times (-1) = 1$$

$$a_2 = g^{[2]}(z_2) = \mathrm{ReLU}(z_2) = 1$$

$$\hat{y} = w_5 a_1 + w_6 a_2 = 3 \times 1 + 2 \times 1 = 5$$

Option c: Incorrect. The ReLU function is a non-linear activation function.

7D. a

$g^{[1]}$ is sigmoid function and $g^{[2]}$ is identity function. We have:

$$z_1 = w_1 x_1 + w_3 x_2$$

$$a_1 = g^{[1]}(z_1) = \sigma(z_1)$$

$$z_2 = w_2 x_1 + w_4 x_2$$

$$a_2 = g^{[2]}(z_2) = \sigma(z_2)$$

$$\hat{y} = w_5 a_1 + w_6 a_2 = w_5 \sigma(w_1 x_1 + w_3 x_2) + w_6 \sigma(w_2 x_1 + w_4 x_2)$$

Option a: Correct. We can set $w_5 = 1$ and $w_6 = 0$. The network's output becomes:

$$\hat{y} = \sigma(w_1 x_1 + w_3 x_2)$$

Option b: Incorrect. With all non-zero weights, the network's output is a linear combination of $\sigma(z_1)$ and $\sigma(z_2)$ (two sigmoid functions' output) and cannot represent a logistic regression model (one single sigmoid function's output).

Option c: Incorrect. The values a_1 and a_2 are between 0 and 1, but the weights w_5 and w_6 can be any value.

Part 8: Convolutional Neural Network (5 questions, 8 marks)

Context

You are given a 3-channel image, and the details are provided below:

Channel 1:

-1	-2	1	0
1	-2	-3	0
-1	-1	0	2
1	0	1	-1

Channel 2:

1	-1	1	-2
-2	-2	-1	-3
-1	2	2	0
0	1	-1	0

Channel 3:

3	0	-2	2
-1	-1	-1	-1
2	2	1	1
2	-1	1	1

Using this image, answer questions 8A-8E.

Questions

8A. [2 marks] Channel 1 of the image is passed to a convolutional layer containing 5 kernels, where the height and width of each kernel are both set to 2. The layer uses a stride of 2 and no padding.
What is the output **height** and what is the **number of output channels**?
Fill your answer in the following blanks.
Output height = ____
Output channel number = ____

8B. [1 mark] You now need to generate a feature map by taking the given 3-channel image as input using a convolutional layer with a stride of 1, no padding. There is only one kernel of shape 2 x 2 x 3 (Height x Width x Channels) with all weights set to 1 in this convolutional layer.
What is the value at position: (Row 2, Column 2) of the output feature map?
Note: Row and column indices start at 1.

Fill your answer in the following blank.
Value = ___

8C. [1 mark] You now directly apply a max pooling layer with window size 2x2 and a stride of 2 to the Channel 1 of the image. What is the sum of values in the resulting output?
Fill your answer in the following blank.
Sum of values = ___

8D. [2 marks] You now need to design a convolutional layer that generates a 3-channel output, where each output channel contains the sum of its corresponding input channel, i.e., channel-wise sum.

For example, the output's first channel contains a single value which is -5 (i.e., -1-2+1+0+1-2-3+0-1-1+0+2+1+0+1-1), the sum of all 16 values in Channel 1.

Which of the following convolutional layer settings would achieve this? Select all that apply.

a. One kernel of shape 4 x 4 x 3 (Height x Width x Channels), with all weight values set to 1, a stride of 4, and no padding.
b. Three kernels, each of shape 4 x 4 x 3 (Height x Width x Channels), with all weight values set to 1, a stride of 4, and no padding.
c. Three kernels, each of shape 1 x 1 x 3 (Height x Width x Channels), with all weight values set to 1/3, a stride of 4, and no padding.
d. None of the above.

8E. [2 marks] You are designing a neural network for a 10-class classification task. The last convolutional layer in your neural network produces an output with a shape of 2 x 2 x 8 (Height x Width x Channels).

This convolutional layer output is flattened and then fed into a fully-connected layer, which generates the final output. Cross-entropy loss is applied for model training.

Which of the following statements is correct? Select all that apply.

a. The fully-connected layer should contain 10 neurons.
b. The fully-connected layer should contain 32 neurons.
c. Without bias, the number of trainable parameters (weights) in the fully-connected layer is 10.
d. None of the above

Answers

8A: 2; 5

Output height: $H_{out} = \left\lfloor \frac{H_{in} - K + 2P}{S} \right\rfloor + 1 = \left\lfloor \frac{4-2+0}{2} \right\rfloor + 1 = 2$

Output channels: The number of output channels is always equal to the number of kernels in the layer. Since there are 5 kernels, the output has 5 channels.

8B: -4

The output at (Row 2, Column 2) is calculated by placing the kernel starting at the input's (Row 2, Column 2).

Patch from Channel 1:

-2	-3
-1	0

Patch from Channel 2:

-2	-1
2	2

Patch from Channel 3:

-1	-1
2	1

The convolution is the sum of the element-wise multiplication of the kernel and the input patch. Since all kernel weights are 1, this is just the sum of all 12 values.

$$-2 - 3 - 1 + 0 - 2 - 1 + 2 + 2 - 1 - 1 + 2 + 1 = -4$$

8C: 5

The input is split into four 2×2 quadrants (due to stride 2).

Top-Left: $\max(-1, -2, 1, -2) = 1$

Top-Right: $\max(1, 0, -3, 0) = 1$

Bottom-Left: $\max(-1, -1, 1, 0) = 1$

Bottom-Right: $\max(0, 2, 1, -1) = 2$

Sum: $1 + 1 + 1 + 2 = 5$

8D: d

Option a: Incorrect. To generate a 3-channel output, the number of kernels required is 3.

Option b: Incorrect. With all weight values in the kernels set to 1, each kernel will aggregate all 3 input channels and compute the sum of all pixels in the entire 3-channel image.

Option c: Incorrect. A kernel with its height and width set to 1 does not sum the spatial dimensions (4 x 4) of the input.

8E: a

Option a. Correct: The layer needs 10 neurons to represent the 10 classes.

Option b. Incorrect: 32 is the input size to this layer, not the output size.

Option c. Incorrect: The number of parameters (weights) is calculated as 32 x 10 = 320.

Part 9: Recurrent Neural Network (5 questions, 7 marks)

Context

A sports analytics company needs to classify video clips of basketball plays into one of four categories: “Dunk”, “Three-Point Shot”, “Pass”, or “Dribbling”.

Data:

- 8,000 video clips of basketball plays.
- Each video clip is a sequence of 20 frames.
- Each frame is an image of size 224 x 224 x 3 (Height x Width x Channels).
- A target class label is provided for each video clip.

The model designed by the company:

Part 1: A CNN processes each frame individually and outputs a 128-dimensional feature vector for each frame.

Part 2: An RNN model receives the sequence of 20 feature vectors (one 128-dimensional feature vector per frame) generated by the CNN for each video clip.

Based on this description, answer questions 9A-9E.

Questions

9A. [2 marks] Consider a single frame showing a player in mid-air near the basket. What is the CNN component (Part 1) most likely to encode in its 128-dimensional feature vector for this frame?

a. The speed at which the player is moving.
b. Features representing the player's pose.
c. The index of the current frame, e.g., this is the 16^{th} frame.
d. None of the above

9B. [1 mark] What type of RNN model (Part 2) should be used?

a. One-to-Many
b. Many-to-One
c. Many-to-Many
d. None of the above

9C. [1 mark] In the RNN model (Part 2), what does the hidden state at time step t=10 represent?

a. The features of only the 10th frame.

b. The summarized information for the entire 20-frame video.
c. The information extracted from the first 10 frames of the video.
d. None of the above

9D. [2 marks] Now, instead of using the RNN model to process the sequence of 20 feature vectors, your friend recommends computing an averaged vector from the 20 feature vectors generated by the CNN.

This averaged vector is computed by first summing the 20 feature vectors and then dividing the result by 20. This averaged vector is then used as the input for a fully-connected layer to predict the output.

Which of the following statements is correct? Select all that apply.

a. During testing, the new model will produce the same output for a video clip even if its frames are processed in reverse order.
b. During testing, the model's processing time to predict the label for one test video will decrease when running on the same machine.
c. This new model will be more likely to overfit the data than the original CNN+RNN model.
d. None of the above.

9E. [1 mark] Suppose there are 3 RNN layers in the RNN model (Part 2), to predict the final output (probability for each of the four action categories) using a fully-connected layer, what should be used as the input to the last output layer?

a. The hidden state generated in first time step of the first RNN layer.
b. The hidden state generated in last time step of the last RNN layer.
c. The hidden state generated in first time step of the last RNN layer.
d. None of the above

Answers

9A: b

Option a: Incorrect. Speed is defined by the change in position over time. A single static frame cannot explicitly show speed.

Option b: Correct. A player in mid-air describes a spatial configuration or pose visible within that single frame, which the CNN can encode.

Option c: Incorrect. The CNN only sees the pixel data of the current image. It has no external context to know that a specific image is the 16th in a sequence.

9B: b

Input: A sequence of 20 frames (Many).

Output: A single classification label for the whole video clip, such as "Dunk" (One).

9C: c

In an RNN, the hidden state at time step t acts as the network's “memory”. It is calculated based on the current input and the previous hidden state. Therefore, the hidden state at time step t=10 represents a summary of the first 10 frames of the video. It does not yet contain information about frames 11–20.

9D: a, b

Option a: Correct. The average of frames [1, 2, ..., 20] is exactly the same as the average of frames [20, 19, ..., 1]. Since the input to the final fully-connected layer is identical, the output will be identical.

Option b: Correct. RNNs must process data sequentially (step 1 must finish before step 2 starts), which can be slow. Calculating a simple average is a basic arithmetic operation that is computationally cheap.

Option c: Incorrect. The averaging model removes the RNN parameters (weights), reducing the model's complexity and capacity. Simpler models with fewer parameters are generally less likely to overfit compared to complex models

9E: b

To classify the entire video, the model needs to have seen the entire sequence. The hidden state at the last time step is the only state that contains accumulated information from the start of the video to the end.

Part 10: Attention Neural Network (4 questions, 7 marks)

Context

You are building a model for Anomaly Detection on sensor readings.

Data:

- 10,000 sequences of sensor readings.
- Each sequence has a fixed length of 5, representing consecutive sensor readings generated by a sensor (e.g., [20.2, 20.3, 20.2, 55.7, 20.4]).
- Each individual sensor value is labeled as either "Normal" or "Anomaly" (e.g., the labels for the sequence above would be [“Normal”, “Normal”, “Normal”, “Anomaly”, “Normal”]).

Based on this description, answer questions 10A-10D.

Questions

10A. [2 marks] Suppose you need to use a self-attention layer to process input sequence given in the context. The element at **each position** just contains **one value**.

When the self-attention layer calculates the output for the 4th element (55.7), which element(s) from the input sequence [20.2, 20.3, 20.2, 55.7, 20.4] is it allowed to attend to?

a. Only the third element (20.2).
b. Only the first element (20.2).
c. Only the 4th element itself (55.7).
d. All 5 elements in the sequence.

10B. [2 marks] Suppose you are asked to build a many-to-many attention neural network for this task. Which of the following setups for the final fully-connected output layer and loss function is valid? Select all that apply.

a. Output layer: 1 neuron with Sigmoid activation; Loss function: Binary Cross-Entropy.
b. Output Layer: 2 neurons with Softmax activation. Loss function: Cross-Entropy.
c. Output Layer: 5 neurons with Identity activation. Loss function: Cross-Entropy.
d. None of the above.

10C. [1 mark] Suppose you need to design an attention neural network for a real-time system. Specifically, the model needs to detect an anomaly immediately as a reading arrives, without any knowledge of future readings.

To achieve this, which attention layer should you employ?

a. Self-attention layer
b. Cross-attention layer
c. Masked self-attention layer
d. None of the above

10D. [2 marks] Your friend suggests you use a neural network with just 3 fully-connected layers for this Anomaly Detection task.

The neural network takes the sensor readings sequence as 5-dimensional feature vector input. The number of neurons in the last fully-connected is set to 5 so that the probability of being "Normal" for all 5 elements in the input sequence can be predicted.

Which of the following statements is correct? Select all that apply.

a. As the input sequence length is 5, the number of neurons in the first fully-connected layer must be set to 5.
b. For the given example sequence: [20.2, 20.3, 20.2, 55.7, 20.4], the predicted probability of being "Normal" for the two 20.2 values is the same, regardless of model weights.
c. 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.
d. None of the above

Answers

10A: d

In a self-attention layer, every element in the sequence attends to every other elements in the same sequence to compute its representation.

10B: a, b

Option a: Correct. This is the standard setup for binary classification.

Option b: Correct. This is the standard setup for multi-class classification, treating "Normal" and "Anomaly" as two distinct classes.

Option c: Incorrect. Identity activation implies the output is a real number which should not be directly used as input to compute the cross-entropy loss.

10C: c

Masked self-attention ensures that the position t can only attend to previous positions and position t.

10D: d

Option a: Incorrect. The input layer has a dimension of 5. However, the first fully-connected layer can have any number of neurons (e.g., 64, 128). The number of neurons in the first fully-connected layer is not dictated by the input size.

Option b: Incorrect. In a standard MLP, the input is treated as a flat vector where position matters explicitly. Input = $[x_1, x_2, x_3, x_4, x_5]$.

The first value (x_1 = 20.2) is multiplied by a set of weights. The third value (x_3= 20.2) is multiplied by a different set of weights, leading to potentially different output probabilities for the same numerical value at different positions.

Option c: Incorrect. The number of neurons in the last fully-connected is set to 5, so the model cannot directly be used to predict 7 labels for the sensor readings of length 7.

—END OF PAPER —