National University of Singapore

CS2109S—Introduction to AI and Machine Learning

Final Assessment - Context for Examplify Questions

Semester 2, 2024/2025

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 **Six (6) 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
General Machine Learning	2	4
Kernel Method	2	3
Regularization	3	5
Support Vector Machine	4	6
Unsupervised Learning	3	6
Perceptron	2	4
Neural Network I	3	6
Neural Network II	4	7
Convolutional Neural Network	6	10
Recurrent Neural Network	3	5
Total	**35**	**60**

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

Context

A hospital employs autonomous robots to deliver medications to patient rooms. The robots must navigate corridors, avoid people or equipment in their path, and prioritize urgent deliveries for critical patients while maintaining safe operation. They are equipped with cameras and lidar, which enable them to perceive their surroundings, and utilize wheels and a gripper to interact with the environment. Additionally, the robots are powered by a very long-lasting battery. These are the only components that make up these straightforward robots. These robots are expected to fully manage the deliveries, meaning they need to keep track of which medications have been delivered and determine whether patients require follow-up medication.

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

Questions

1A. [2 marks] Which components are **correct** for the PEAS framework of the hospital delivery robot? Select all that apply.

- ✓A. **Performance**: Minimize damage to medications during transit
- ✓B. **Performance**: Prioritize deliveries for critical patients first
- C. **Performance**: Maximize battery life during operation
- D. **Environment**: Parking lots and charging stations
- E. **Environment**: Consists only of predefined grid of corridors with only fixed obstacles
- F. **Actuators**: Alarm system for collisions
- G. **Actuators**: Path-planning algorithm for route optimization
- H. **Sensors**: Temperature sensors to monitor medication integrity

1B. [1 mark] Which **agent type** is **most appropriate** for this robot?

- 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
- ✓B. Stochastic
- C. Episodic
- D. Static
- ✓E. Continuous

Solution

1A. A and B

Based on the description:

- The robot must prioritize urgent deliveries (B) and ensure safe operation (A).
- Battery life (C) is not a performance goal because the robot is equipped with very long-lasting battery.
- The environment includes dynamic obstacles (e.g., people, equipment), so it is not a static grid (C). There is no mention of parking lots (D).
- Actuators are wheels and gripper, there is no mention of alarms (F). Algorithms (G) are part of the agent's internal decision-making, not actuators.
- Sensors are cameras and lidar, there is no mention of temperature sensors (H).

1B. D

The robot must optimize decisions in a dynamic environment:

- Prioritize critical patients (assigning higher utility to urgent deliveries).
- Balance competing goals (e.g., speed vs. safety).

Why not other agents?

- A/B. Reflex agents (simple/model-based) react to current percepts but cannot handle complex trade-offs.
- C. Goal-based agents focus on achieving a single goal, not ranking multiple goals (e.g., urgency vs. efficiency).

1C. B and E

- **Partially-observable**: Sensors (cameras/lidar) have limitations (e.g., blind spots, occlusions).
- **Stochastic**: The environment have random elements (e.g., people/equipment may block paths unexpectedly/randomly).
- **Sequential:** Tasks are sequential and interdependent (e.g., tracking deliveries, follow-up medications).
- **Dynamic**: The environment changes dynamically (e.g., moving people/equipment).
- **Continuous**: The robot navigates through corridors in real-world space.

Part 2: General Machine Learning (2 questions, 4 marks)

There is no context, please answer the question(s) directly.

Questions

2A. [2 marks] You are given the following machine learning situation. At great monetary expense, you obtain a limited number of training data $(x^{(j)}, y^{(j)})$ such that $y^{(j)} = f(x^{(j)})$, and $f(x^{(j)})$ is either 0 or 1. Here, $f(x)$ is the ground truth that is unknown to you. Deployment of your machine learning method will have many imperfections. You are faced with a situation where the new data $x^{(k)}$ arrives as $x^{(k)} + q^{(k)}$, where $q^{(k)}$ is some vector of random noise, i.e., each component of the vector is noisy. The noise is relatively small, but noticeable.
Based on the description of the problem, what is the **best** learning model to choose? Here, **best** is defined as being the most appropriate for the specific properties of the new data. Your choice should not be dependent on factors that are not described in the problem.

A. Linear regression.
B. Logistic regression.
C. Two-layer neural network with logistic output layer.
✓D. Support vector machine.
E. K-Means.

2B. [2 marks] The self-attention layer can be employed to capture contextual information over the input sequence. Which of the following statements is/are true? Select all that apply.

A. Self-attention layer can only process sequential data one time step at a time.
B. To capture contextual information, we need to maintain a hidden state that stores past information.
C. The output of the self-attention layer is obtained by taking a weighted sum of the query vectors.
✓D. None of the above.

Solution

2A. D
From the problem statement, it is a classification problem in {0,1}, which rules out A and E. Also from the problem statement, we learn that noise in the test data/deployment is the key aspect of the setting. Logistic regression will find a decision boundary; however, it could be any decision boundary that separates the training data. Notice that the training data came without noise. The Support Vector Machine is the most appropriate for the new data, because the SVM finds a maximum margin decision boundary based on the noiseless training data. It separates the two classes as much as possible; relatively small noise will not lead to much misclassification of the new data.

2B. D
Option A: Given the whole input sequence, self-attention can process the input sequence simultaneously via matrix multiplications.
Option B: Unlike RNNs, self-attention doesn't rely on a running hidden state to carry past context.
Option C: The output of the self-attention layer is a weighted sum of the value vectors, not the query vectors.

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

Context

A company is developing a model to predict customer's spending using two features: time spent on a website (x_1) and pages visited (x_2). There is no dummy feature x_0 (no bias term): the model is $h_w(x) = w_1 x_1 + w_2 x_2$. They adopt a kernel method with a dual model formulation:

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

where x_i are training points and α_i are linear-combination coefficients.
Training Data:

- $x^{(1)} = [1,1]^T,\ \alpha_1 = 1.0$
- $x^{(2)} = [2,2]^T,\ \alpha_2 = 0.25$

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

Questions

3A. [1 mark] Suppose that the company wants to use the feature map $(x) = [\log(x_1), \log(x_2)]$. Write down the formula for the kernel $k(u, v)$ that corresponds to this feature map. Here, $u, v \in \mathbb{R}^2$. __1__

Note: You may use u_j to denote feature u_j, a*b to denote multiplication between a and b, and M^T to denote the transpose of matrix M.

1. log(u_1)log(v_1)+log(u_2)log(v_2)

3B. [2 marks] Suppose that the company has decided to use kernel $k(u, v) = (u^\top v)^2$. Compute the output $h_\alpha(x)$ for a new customer $x = [1, 2]^\top$. __1__

1. 18

Solutions

3A. log(u_1)log(v_1)+log(u_2)log(v_2)

$\log(u) \cdot \log(v)$

$= \log(u)^T \log(v)$

$= [\log(u_1)\ \log(u_2)][\log(v_1)\ \log(v_2)]^T$

$= \log(u_1)\log(v_1) + \log(u_2)\log(v_2)$

The equation can be written in different ways. We will accept them as long as they are valid.

3B. 18

$k\left(x^{(1)}, x\right) = (1 \cdot 1 + 1 \cdot 2)^2 = 9$

$k\left(x^{(2)}, x\right) = (2 \cdot 1 + 2 \cdot 2)^2 = 36$

$h_\alpha(x) = 1.0 \cdot 9 + 0.25 \cdot 36 = 18$

Part 4: Regularization (3 questions, 5 marks)

Context

A hospital is developing an AI model to predict heart disease risk using logistic regression. The dataset includes 1,200 patients with 40 features (e.g., cholesterol levels, exercise habits, genetic markers). The initial model achieves 94% training accuracy but 58% validation accuracy. The team runs two experiments:

1. **Experiment A**: L1 regularization → Validation accuracy: 76%, 12 features have non-zero weights.
2. **Experiment B**: L2 regularization → Validation accuracy: 71%, all 40 features have non-zero weights.

Clinicians emphasize that interpretability is critical for trust in diagnostics.

Based on this description, answer questions 4A-4C.

Questions

4A. [2 marks] Which factor(s) explain why Experiment A achieved higher validation accuracy than Experiment B? Select all that apply.

A. L1 regularization increased model bias.
✓B. L1 reduced overfitting by eliminating irrelevant features.
C. L2 regularization is unsuited for medical datasets.
✓D. L2 retained noise from features that have low correlation with the output.
E. L1 inherently improves computational efficiency.

4B. [2 marks] If the team prioritizes identifying lifestyle factors (e.g., exercise, diet) for targeted patient interventions, which experiment is preferable?

✓A. Experiment A: Explicitly reveals key predictors due to sparse weights.
B. Experiment B: Maintains higher model complexity for nuanced insights.
C. Neither; SVMs should replace logistic regression for clinical use cases.
D. Both are equally valid since accuracy differences are minimal.

4C. [1 mark] Increasing the L2 regularization strength in Experiment B caused validation accuracy to drop to 65%. Which of the following explain this outcome? Select all that apply.

✓A. The model's weights became too small to capture patterns.
✓B. The model prioritized minimizing weights over fitting trends.
C. Increased regularization strength led to higher model variance.
✓D. High penalty led to underfitting.

Solution

4A. B and D

- **B. L1 reduced overfitting by eliminating irrelevant features.**
  - **Mechanism:** L1 regularization penalizes the absolute values of weights, forcing some weights to **zero** (as seen in tutorial). This effectively removes irrelevant features (e.g., noisy genetic markers or redundant variables) from the model.
  - **Impact:** By retaining only 12/40 features, Experiment A simplified the model, reducing its tendency to memorize noise in the training data (overfitting). This improved generalization to the validation set.
- **D. L2 retained noise from features with low correlation.**
  - **Mechanism:** L2 regularization penalizes squared weights, shrinking coefficients but **keeping all features** (as seen in tutorial).
  - **Impact:** Even weakly correlated features (e.g., minor genetic markers) retained small weights, introducing noise. This preserved complexity, leading to worse validation accuracy compared to L1.
- **A. "Increased model bias":**
  - The improved validation accuracy (76% vs. 71%) shows better generalization, not increased bias which leads to lower performance.
- **C. "L2 is unsuited for medical data":**
  - It has nothing to do with the observations.
- **E. "Computational efficiency":**
  - While L1 can be faster models due to sparsity, the question focuses on accuracy, not computational speed.

4B. A

- **A. Experiment A: Explicitly reveals key predictors due to sparse weights.**
  - **Why:** L1's sparsity (12 non-zero weights) highlights the **most influential features** (e.g., exercise, diet). Clinicians can directly identify actionable lifestyle factors for interventions.
  - **Example:** If "exercise frequency" has a large non-zero weight, clinicians can prioritize exercise plans for high-risk patients.
- **B. "Higher complexity for nuanced insights":**
  - L2's 40 non-zero weights obscure key drivers. Clinicians cannot easily distinguish critical lifestyle factors from less relevant ones (e.g., genetic markers).
- **C. "SVMs should replace logistic regression":**
  - SVMs don't help with identifying important features.
- **D. "Both are equally valid":**
  - Accuracy differences (76% vs. 71%) and interpretability needs make Experiment A strictly preferable.

4C. A, B, and D

- **A, B, and D** are correct:

  - A: Increasing L2 strength forces weights toward zero, shrinking their magnitudes. This weakens the model's ability to learn meaningful relationships (e.g., between cholesterol levels and heart disease).
  - B: The loss function overly prioritizes weight minimization (to avoid penalty) at the expense of fitting the training data. This leads to oversimplification.
  - C: The model becomes too simple to capture underlying patterns (high bias) – It underfits the data. Validation accuracy drops (65%) because it fails to generalize.
- ***C* is incorrect**:
  - Increased regularization reduces variance (less sensitivity to noise). The drop in accuracy is due to high bias (underfitting), not variance.

Part 5: Support Vector Machine (4 questions, 6 marks)

There is no context, please answer the questions directly.

Questions

5A. [2 marks] From the following statements, select correct statement(s). Select all that apply.

A. SVMs cannot be applied to multi-class classification.
B. The primal objective of the SVM can be simplified to a non-convex optimization problem.
✓C. The SVM is formalized with an optimization problem with inequality constraints.
✓D. Both primal and dual formulations are valid formulations for SVMs.
✓E. The dual formulation allows the use of kernel functions for classification of non-linear data.
F. None of the above.

5B. [2 marks] You are given the following data set

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

You train a Support Vector Machine. 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. $x^{(5)}$

5C. [1 mark] You have trained a SVM on customer data for a classification problem of VIP (+1) or Not-VIP (-1) customers. (VIP means very important person). The training of the SVM outputs a hyperplane normal vector $w = [-1\ 1.1\ -2\ -0.2\ -0.1\ 2]^T$ and zero offset. You receive the new customer $x = [0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5]^T$. What do you report to your manager about this customer?

A. VIP.
✓B. Not-VIP.
C. The customer is inside the margin hence cannot be classified.
D. The customer is on the wrong side of the hyperplane and hence cannot be classified.

5D. [1 mark] You have data about 10000 customers, each being described by 30 real-valued attributes and a VIP/Not-VIP label. Training a SVM has obtained 17 support vectors, and outputs an array of size 10000 of dual coefficients α. The dual coefficients contain the label information.

You are given a storage medium of limited size. You want to store the SVM classifier on the medium. With the stored information, you want to be able to classify new customers. Assume that you do not need to store additional information about the offset of the hyperplane. Storing each real number has a unit cost, i.e., a cost of 1. What is the smallest size of your storage medium to store the classifier? Fill the size into the blank using unit cost as the metric. __1__

1. 527

Solution

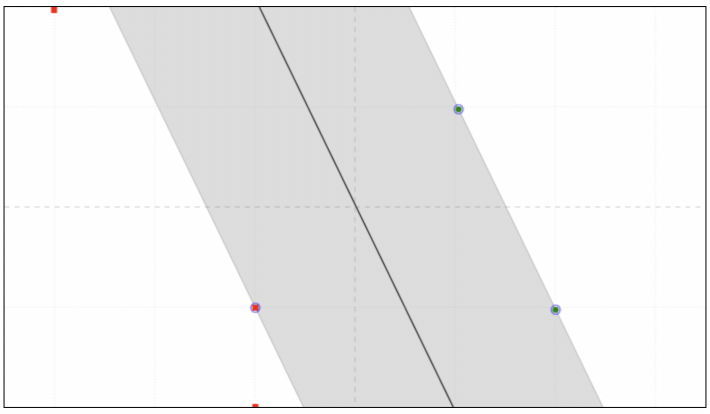
5A. CDE

For A, the SVMs can be applied to multi-class classification by training one-vs-one or one-vs-all classifiers. For B, the primal objective of the SVM can be simplified to a convex optimization problem, ||w||^2 is a convex function of w.

For C, the SVM is formalized with an optimization problem with inequality constraints. We saw the inequality constraints for the primal problem discussed in tutorial. (The dual formulation also has inequality constraints, not directly discussed in class.) For D, both primal and dual formulations are indeed valid formulations for SVMs. For E, indeed the dual formulation allows the use of kernel functions for classification of non-linear data.

5B. ABD

Visualize the problem by drawing the points in the 2D plane. Notice that points 1 and 4 are the closest together from the two different classes, hence they will be on the margin. So we found two support vectors. Drawing the maximum margin hyperplane using only these two points, we see that points 2 and 3 fall inside the margin. If we use point 3 as an additional support vector, we find that point 4 is still inside the margin. On the other hand, if we use point 2 as an additional support vector, we find that point 3 is outside the margin. Hence points 1,2, and 4 are the support vectors. To maximize your score in the presence of uncertainty, it is important to pick only the points that you are sure about.



5C. B

Computing the dot product between $w = [-1\ 1.1\ -2\ -0.2\ -0.1\ 2]^T$ and $x = [0.5\ 0.5\ 0.5\ 0.5\ 0.5]^T$ gives -0.1 for which sign(-0.1) = -1 hence it is a non-VIP customer.

5D. 527

The storage is initially empty and needs to store only data required for classifying new customers. While we obtain an array of 10000 of dual coefficients α, it is also reported that we have found 17 support vectors. Hence, this array contains many 0s which we do not have to store (sparsity property of SVMs).

For each support vector (customer), we have to store all the features. That is we have to store 17*30 = 510 units. For each support vector (customer), we also have to store the corresponding dual coefficient α_j, hence we have to store additional 17 units.

In total, the size of the storage is required to be 17*30 + 17 = 527.

Part 6: Unsupervised Learning (3 questions, 6 marks)

There is no context, please answer the questions directly.

Questions

6A. [2 marks] Consider the following dataset consisting of three points in 2D space:
$x^{(1)}$ =[−0.50.5]^T
$x^{(2)}$ =[0.50.5]^T
$x^{(3)}$ =[3.53.5]^T .
Initialize two cluster centers by choosing $x^{(1)}$ and $x^{(3)}$. Then, run the K-means algorithm, using Euclidean distance, until convergence. Input the resulting cluster centers as follows. Choose the cluster center with the **smallest first coordinate,** e.g., given cluster center 1: [2 5]^T and cluster center 2: [3 4]^T, the cluster center with the smallest first coordinate is cluster center 1, since 2 is the smallest value among all first coordinates: 2, 3. For this cluster, fill in the blanks [1 , 2] with the first and the second coordinate. Then, input the first and the second coordinate of the other cluster center into the blanks [3 , 4]. Use decimal format for fractions (e.g., 0.25).

1. 0
2. 0.5
3. 3.5
4. 3.5

6B. [2 marks] You are given data from a customer survey. After performing Singular Value Decomposition on the transpose of the mean-centered data matrix, you obtain the following matrix:

$$\Sigma = \begin{bmatrix} 6 & 0 & 0 & 0 & 0 \\ 0 & 4.5 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{bmatrix},$$

among other outputs of the SVD. You are given the task to retain 90% of the variance of the data. Select the diagonal entries of Σ that you should keep. Select all that apply.

A. 0.5
B. 1
C. 2
✓D. 4.5
✓E. 6

6C. [2 marks] You have 500 greyscale images of size 64 × 64. First, you choose to represent each pixel by a single feature. Next, you compute the means for all features using the available data. Finally, you define a mean-centred data matrix and perform a Singular Value Decomposition (SVD) on the transpose of the mean-centered data matrix. You retain 50 significant singular values and use the corresponding output of the SVD to compress the data.

You have a storage medium of limited size. Given only the data on the storage medium, you would like to be able to reconstruct the images and compress other images. Each pixel and each real number has a unit cost for storage, i.e., a cost of 1 to store on the medium.

What is the **smallest** size of your storage medium? Fill the size of the storage medium into the blank using unit cost as the metric. __1__

Hint: The computed means have to be stored as well.

1. 233896

Solution

6A 0**,** 0.5**,** 3.5**,** 3.5

Given the initial assignments, compute the Euclidean distance of point 2 to 1 and 3. The distance of point 1 and point 2 is 1. The distance of point 3 and point 2 is sqrt(18) which is greater than 1. Hence point 2 is assigned to cluster centre at point 1. For updating the cluster centres, compute the mean of point 1 and point 2. It is $([-0.5 \quad 0.5]^T + [0.5 \quad 0.5]^T)/2 = [0 \quad 0.5]^T$. The other cluster centre remains at point 3. The next assignment step does not change the assignments.

6B DE

In class we have learned how the singular values of the transpose of the mean-centred data matrix relate to the variances. The key aspect is to square the singular values, and compute the fraction of the r largest squared singular values over all squared singular values. To retain 90% of the variance we have to first compute the denominator, i.e. the total variance (6^2 + 4.5^2 + 2^2 + 1 + 0.5^2) /(N-1)= 61.5/(N-1). Note that we do not need to know N-1 for the task as it will cancel with the numerator in the formula given in class. Then, compute:

6^2/61.5 = 0.585. This does not retain 90% of the variance.

(6^2 + 4.5^2)/61.5 = 0.915. This retains more than 90% of the variance. Hence, select option D and E.

Given the interpretation of the singular values as importance values for different basis vectors in the new basis, we do not use other combinations that give a variance of more than 90%. We have to start from the top important singular values.

6C 233896

Representing each pixel by a feature means that we have 64 × 64 = 4096 features. Computing the mean for all features leads to 4096 means which we have to store, each one is a single real number (1 unit of storage).

Retaining 50 singular values leads to a compression matrix ~U that is of dimension 4096 * 50 which equals 204800 real numbers. To compress new images we have to store this matrix.

Applying ~U^T X^T leads to the compressed data set of 500 images each having 50 features. Hence, we have to store 50*500 = 25000 real numbers. To be able to reconstruct the given images we have to store this compressed data, in addition to ~U.

In total, we hence have to store: 4096 (means) + 204800 (~U) + 25000 (compressed data) = 233896.

Part 7: Perceptron (2 questions, 4 marks)

Context

Consider the following dataset (x_1 and x_2 are features, and y represents the label):

Sample	x_1	x_2	y
p_1	1	1	-1
p_2	0.5	0	1
p_3	3	-0.5	1

You are tasked with using the Perceptron model discussed in the lecture to process these three data samples. The initial weights of the Perceptron model are given as follows: w_0 = -2 (bias), w_1 = 1, w_2 = 0.5.
Note: The following Sign function is used as the activation function in the Perceptron model:

$$sgn(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

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

Questions

7A. [2 marks] Which data point(s) is/are correctly classified by the Perceptron model when initialized with the given weights? Select all that apply.

✓A. p_1
B. p_2
✓C. p_3
D. None of the above

7B. [2 marks] Perform a single step of the Perceptron update rule starting from the initial weights using a learning rate of 3. The updated value for w_0 is __1__, and the updated value for w_1 is __2__.

1. 4
2. 4

Solution

7A. A and C
Predicted label for p_1: $sgn(-2 * 1 + 1 * 1 + 0.5 * 1) = -1$
Predicted label for p_2: $sgn(-2 * 1 + 1 * 0.5 + 0.5 * 0) = -1$
Predicted label for p_3: $sgn(-2 * 1 + 1 * 3 + 0.5 * (-0.5)) = 1$
Therefore, p_1 and p_3 can be correctly classified.
7B. w_0 =4, w_1 =4
p_2 is the misclassified instance. According to the Perceptron update rule:

$$\boldsymbol{w} = \boldsymbol{w} + \gamma\left(y^{(2)} - \hat{y}^{(2)}\right)\boldsymbol{x}^{(2)}$$
$$= \begin{bmatrix} -2 \\ 1 \\ 0.5 \end{bmatrix} + 3\left(1 - (-1)\right)\begin{bmatrix} 1 \\ 0.5 \\ 0 \end{bmatrix} = \begin{bmatrix} -2+6 \\ 1+3 \\ 0.5+0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 0.5 \end{bmatrix}$$

So the updated weights are w_0 =4, w_1 =4, w_2 =0.5.

Part 8: Neural Networks I (3 questions, 6 marks)

Context

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

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

Using this dataset, answer questions 8A-8C.

8A. [2 marks] If we set the decision threshold to 0.5 and use x_1 and x_2 as inputs, a single neuron using a sigmoid activation function can be employed to classify all six training data samples correctly.
Is the above statement True or False?
Note: You are allowed to add a bias term by introducing a dummy input variable. The dummy input variable should be set to 1.

A. True
✓B. False

8B. [2 marks] Your friend suggests that you use one additional feature, $x_1^2x_2^2$. Now you can use x_1, x_2 and $x_1^2x_2^2$ as inputs to a neural network. Which of the following statements is/are true if we set the decision threshold to 0.5? Select all that apply.
Note:
(i) You are allowed to add a bias term by introducing a dummy input variable for each neuron. The dummy input variable should be set to 1 for all neurons.
(ii) A fully-connected layer is a layer of neurons in which each neuron is connected to every neuron in the previous layer.

✓A. A single neuron using a sigmoid activation function can be employed to correctly classify all six training data samples.
✓B. A neural network with two fully-connected layers, a linear activation function in the first layer, and a sigmoid activation function in the second layer can be employed to correctly classify all six training data samples.
✓C. A neural network with two fully-connected layers, a ReLU activation function in the first layer, and a sigmoid activation function in the second layer can be employed to correctly classify all six training data samples.
D. None of the above.

8C. [2 marks] It turns out that the label for the data sample (1, 1) is incorrect, the correct label should be 1. After correcting this error, which of the following statements is/are true if we set the decision threshold to 0.5 and use x_1 and x_2 as inputs to a neural network? Select all that apply.
Note:

(i) You are allowed to add a bias term by introducing a dummy input variable for each neuron. The dummy input variable should be set to 1 for all neurons.
(ii) A fully-connected layer is a layer of neurons in which each neuron is connected to every neuron in the previous layer.

✓A. A single neuron using a sigmoid activation function can be employed to correctly classify all six training data samples.

✓B. A neural network with two fully-connected layers, a linear activation function in the first layer, and a sigmoid activation function in the second layer can be employed to correctly classify all six training data samples.

✓C. A neural network with two fully-connected layers, a ReLU activation function in the first layer, and a sigmoid activation function in the second layer can be employed to correctly classify all six training data samples.

D. None of the above.

Solution

8A. False

The data points on the decision boundary of the single-neuron model must satisfy:

$$\sigma\left(\sum_{j=0}^{2} w_j x_j\right) = 0.5$$

Since $\sigma(0) = 0.5$, it follows that

$$\sum_{j=0}^{2} w_j x_j = 0$$

Therefore, the decision boundary is linear (a straight line). However, the given dataset is not linearly separable, so the statement is false.

8B. A, B and C

x_1	x_2	$x_3 = {x_1}^2{x_2}^2$	$\sigma(1-2x_3)$	$\sigma(-0.5+ReLU(1-2x_3))$	y
0	0	0	$\sigma(1)$	$\sigma(0.5)$	1
0	1	0	$\sigma(1)$	$\sigma(0.5)$	1
1	0	0	$\sigma(1)$	$\sigma(0.5)$	1
1	1	1	$\sigma(-1)$	$\sigma(-0.5)$	0
-1	1	1	$\sigma(-1)$	$\sigma(-0.5)$	0
-1	-1	1	$\sigma(-1)$	$\sigma(-0.5)$	0

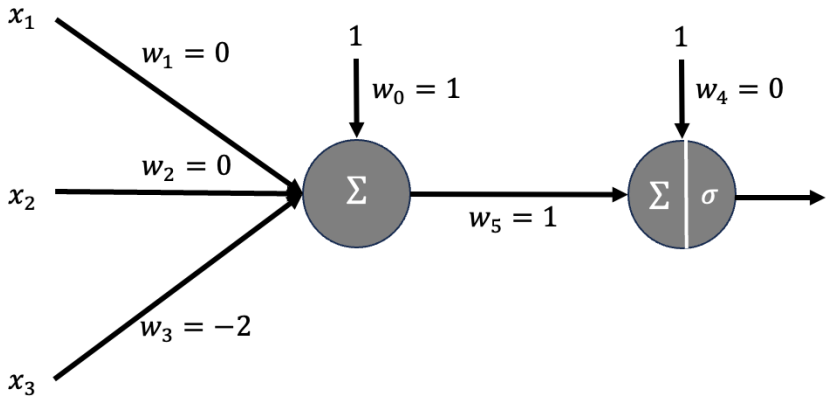
Let's use x_3 to represent ${x_1}^2{x_2}^2$.

For option A, the predicted value for one data sample is $\sigma\left(\sum_{j=0}^{3} w_j x_j\right)$. If we set

$$w_0 = 1, w_1 = 0, w_2 = 0, w_3 = -2$$

The predicted value becomes $\sigma(1-2x_3)$. When the decision threshold is set as 0.5, all the data samples can be classified correctly.

For option B, one possible way to design the neural network is:

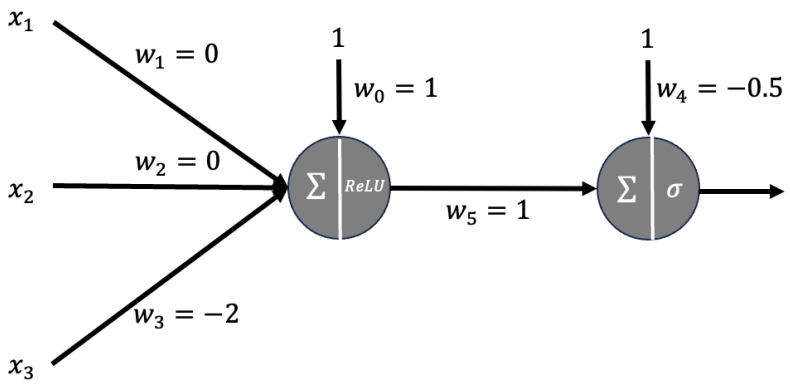


The predicted value for one data sample is:

$$\sigma(w_4 + w_5(w_0 + w_1x_1 + w_2x_2 + w_3x_3)) = \sigma(1-2x_3)$$

When the decision threshold is set at 0.5, all the data samples can be classified correctly.

For option C, one possible way to design the neural network is:



The predicted value for one data sample is:

$$\sigma(w_4 + w_5\, ReLU(w_0 + w_1x_1 + w_2x_2 + w_3x_3)) = \sigma(-0.5 + ReLU(1 - 2x_3))$$

When the decision threshold is set at 0.5, all the data samples can be classified correctly.

8C. A, B and C

The updated dataset is:

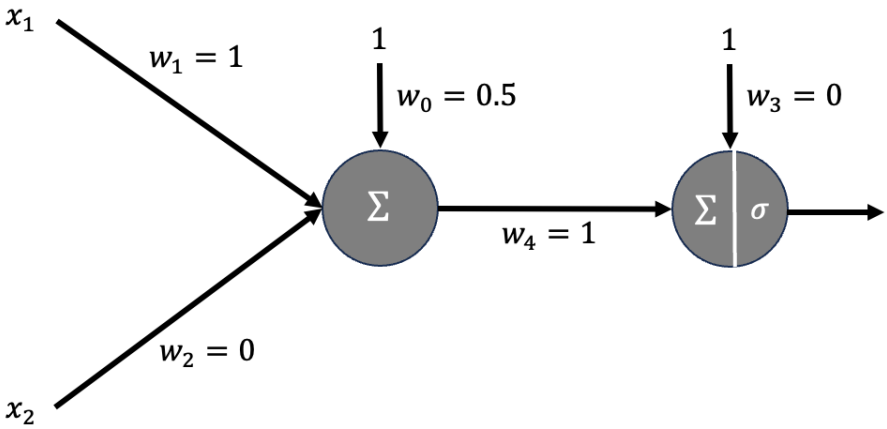
x_1	x_2	$\sigma(0.5 + x_1)$	$\sigma(-0.2 + ReLU(0.5 + x_1))$	y
0	0	$\sigma(0.5)$	$\sigma(0.3)$	1
0	1	$\sigma(0.5)$	$\sigma(0.3)$	1
1	0	$\sigma(1.5)$	$\sigma(1.3)$	1
1	1	$\sigma(1.5)$	$\sigma(1.3)$	1
-1	1	$\sigma(-0.5)$	$\sigma(-0.2)$	0
-1	-1	$\sigma(-0.5)$	$\sigma(-0.2)$	0

For option A, the predicted value for one data sample is $\sigma\left(\sum_{j=0}^{2} w_j x_j\right)$. If we set

$$w_0 = 0.5, w_1 = 1, w_2 = 0$$

The predicted value becomes $\sigma(0.5 + x_1)$. When the decision threshold is set at 0.5, all the data samples can be classified correctly.

For option B, one possible way to design the neural network is:

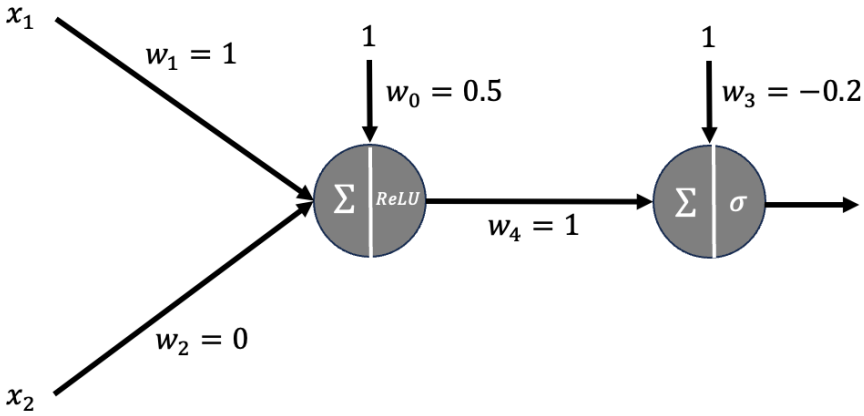


The predicted value for one data sample is:

$$\sigma(w_3 + w_4(w_0 + w_1x_1 + w_2x_2)) = \sigma(0.5 + x_1)$$

When the decision threshold is set at 0.5, all the data samples can be classified correctly.

For option C, one possible way to design the neural network is:



The predicted value for one data sample is:

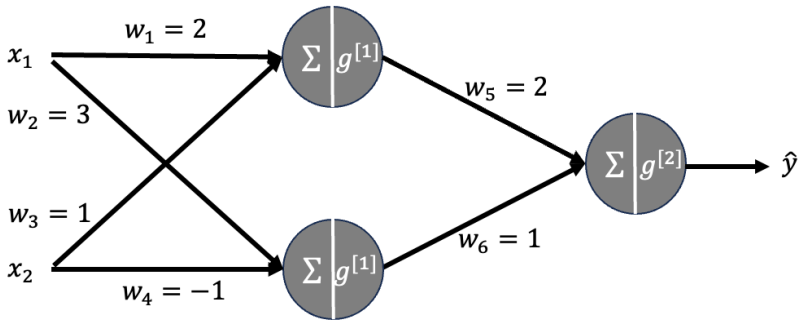
$$\sigma(w_3 + w_4\, ReLU(w_0 + w_1 x_1 + w_2 x_2)) = \sigma(-0.2 + ReLU(0.5 + x_1))$$

When the decision threshold is set as 0.5, all the data samples can be classified correctly.

Part 9: Neural Networks II (4 questions, 7 marks)

Context

Consider the following multi-layer neural network.



Suppose $g^{[1]}$ is linear activation function and $g^{[2]}$ is ReLU activation function.

$$\text{ReLU}(x) = \max(0, x)$$

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

$$L = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2$$

where N is the number of data samples.

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

9A. [1 marks] Given a data sample with two features: x_1 =1 and x_2 =1, what is the predicted value y? __1__

1. 8

9B. [2 marks] Suppose you need to train the neural network with the following 2 data samples together:

x_1	x_2	y
1	0	3
0	-1	2

What is the partial derivative of loss function with respect to w_1? __1__

1. 4

9C. [2 marks] If now you want to apply this neural network in a binary classification task and change the $g^{[2]}$ to be Sigmoid activation function, which of the following statements is true if we set the decision threshold to 0.5?

✓A. The decision boundary generated by using this neural network is always linear.

B. The decision boundary generated by using this neural network is always non-linear.

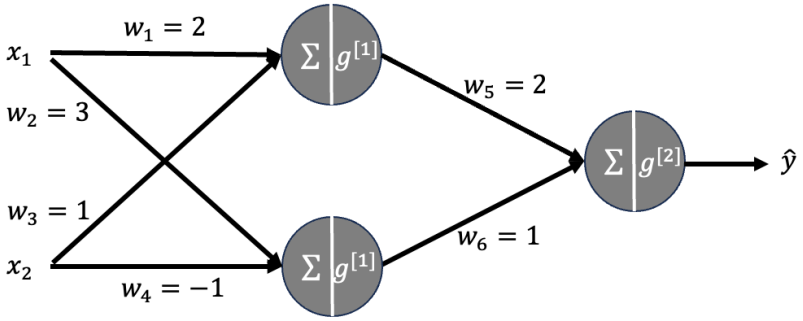
C. The decision boundary generated by using this neural network can be linear or non-linear depending on the weights of the model.

D. None of the above

9D. [2 marks] You need to train this neural network using binary cross-entropy loss, and $g^{[2]}$ has been changed to Sigmoid activation function. Which of the following statements is/are true? Select all that apply.

A. The partial derivative of loss function with respect to any weight should always be smaller than 0.5.
✓B. The partial derivative of loss function with respect to the weights can exceed 0.5.
C. The input feature values do not affect the magnitude of the partial derivative for any weight.
D. None of the above

Solution



9A. 8

x_1 =1 and x_2 =1. $g^{[1]}$ is linear activation function and $g^{[2]}$ is ReLU activation function. Therefore, the predicted value is

$$a_1 = w_1x_1 + w_3x_2 = 3$$
$$a_2 = w_2x_1 + w_4x_2 = 2$$
$$a_3 = w_5a_1 + w_6a_2 = 8$$
$$\hat{y} = ReLU(8) = 8$$

9B. 4

The loss function is defined as:

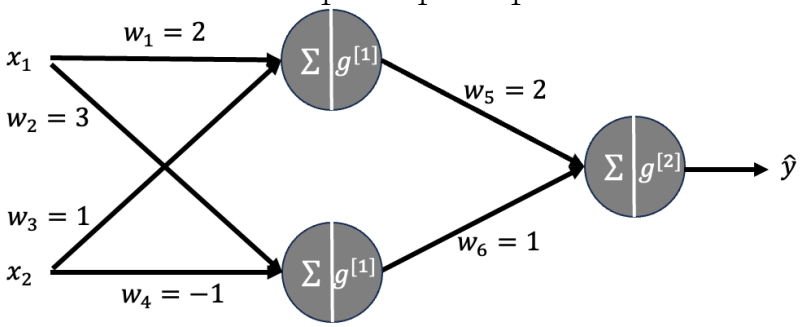
$$L = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2$$

Since we need to train the neural network with the 2 data samples:

$$L = \frac{1}{2 * 2}((\hat{y}^{(1)} - y^{(1)})^2 + (\hat{y}^{(2)} - y^{(2)})^2)$$

Let $L_1 = \frac{1}{4}(\hat{y}^{(1)} - y^{(1)})^2$ and $L_2 = \frac{1}{4}(\hat{y}^{(2)} - y^{(2)})^2$, so $L = L_1 + L_2$ and

$$\frac{\partial L}{\partial w_1} = \frac{\partial L_1}{\partial w_1} + \frac{\partial L_2}{\partial w_1}$$



With the given neural network, $\hat{y}^{(i)}$ can be predicted as follows:

$$a_1{}^{(i)} = w_1x_1{}^{(i)} + w_3x_2{}^{(i)}$$
$$a_2{}^{(i)} = w_2x_1{}^{(i)} + w_4x_2{}^{(i)}$$
$$a_3{}^{(i)} = w_5a_1{}^{(i)} + w_6a_2{}^{(i)}$$
$$\hat{y}^{(i)} = ReLU\left(a_3{}^{(i)}\right)$$
$$L_i = \frac{1}{4}(\hat{y}^{(i)} - y^{(i)})^2$$

The partial derivative is:

$$\frac{\partial L_i}{\partial w_1} = \frac{\partial L_i}{\partial \hat{y}^{(i)}}\frac{\partial \hat{y}^{(i)}}{\partial a_3{}^{(i)}}\frac{\partial a_3{}^{(i)}}{\partial a_1{}^{(i)}}\frac{\partial a_1{}^{(i)}}{\partial w_1} = \frac{1}{4}2\left(\hat{y}^{(i)} - y^{(i)}\right)1_{a_3{}^{(i)}>0}w_5x_1{}^{(i)}$$

$$= \frac{1}{2}\left(\hat{y}^{(i)} - y^{(i)}\right) 1_{a_3{}^{(i)}>0} w_5 x_1{}^{(i)}$$

For the first data sample, $x_1{}^{(1)} = 1, x_2{}^{(1)} = 0, y^{(1)} = 3$

$$a_1{}^{(1)} = w_1 x_1{}^{(1)} + w_3 x_2{}^{(1)} = 2$$
$$a_2{}^{(1)} = w_2 x_1{}^{(1)} + w_4 x_2{}^{(1)} = 3$$
$$a_3{}^{(1)} = w_5 a_1{}^{(1)} + w_6 a_2{}^{(1)} = 7$$
$$\hat{y}^{(1)} = ReLU\left(a_3{}^{(1)}\right) = 7$$
$$\frac{\partial L_1}{\partial w_1} = \frac{1}{2}\left(\hat{y}^{(1)} - y^{(1)}\right) 1_{a_3{}^{(1)}>0} w_5 x_1{}^{(1)} = \frac{1}{2} * (7-3) * 1 * 2 * 1 = 4$$

For the second data sample, $x_1{}^{(2)} = 0, x_2{}^{(2)} = -1, y^{(2)} = 2$

$$a_1{}^{(2)} = w_1 x_1{}^{(2)} + w_3 x_2{}^{(2)} = -1$$
$$a_2{}^{(2)} = w_2 x_1{}^{(2)} + w_4 x_2{}^{(2)} = 1$$
$$a_3{}^{(2)} = w_5 a_1{}^{(2)} + w_6 a_2{}^{(2)} = -1$$
$$\hat{y}^{(2)} = ReLU\left(a_3{}^{(2)}\right) = 0$$
$$\frac{\partial L_2}{\partial w_1} = \frac{1}{2}\left(\hat{y}^{(2)} - y^{(2)}\right) 1_{a_3{}^{(2)}>0} w_5 x_1{}^{(2)} = \frac{1}{2} * (0-2) * 0 * 2 * 0 = 0$$

So

$$\frac{\partial L}{\partial w_1} = \frac{\partial L_1}{\partial w_1} + \frac{\partial L_2}{\partial w_1} = 4 + 0 = 4$$

9C. A

Since $g^{[1]}$ is linear activation function and $g^{[2]}$ is Sigmoid activation function.

$$a_1{}^{(i)} = w_1 x_1{}^{(i)} + w_3 x_2{}^{(i)}$$
$$a_2{}^{(i)} = w_2 x_1{}^{(i)} + w_4 x_2{}^{(i)}$$
$$a_3{}^{(i)} = w_5 a_1{}^{(i)} + w_6 a_2{}^{(i)}$$
$$\hat{y}^{(i)} = \sigma\left(a_3{}^{(i)}\right)$$

Combining the above equations:

$$\hat{y}^{(i)} = \sigma\left(w_5(w_1 x_1{}^{(i)} + w_3 x_2{}^{(i)}) + w_6(w_2 x_1{}^{(i)} + w_4 x_2{}^{(i)})\right)$$
$$= \sigma\left(w_5 w_1 x_1{}^{(i)} + w_5 w_3 x_2{}^{(i)} + w_6 w_2 x_1{}^{(i)} + w_6 w_4 x_2{}^{(i)}\right)$$
$$= \sigma\left((w_5 w_1 + w_6 w_2) x_1{}^{(i)} + (w_5 w_3 + w_6 w_4) x_2{}^{(i)}\right)$$

The decision threshold is set at 0.5. For the data samples on the decision boundary:

$$\sigma\left((w_5 w_1 + w_6 w_2) x_1{}^{(i)} + (w_5 w_3 + w_6 w_4) x_2{}^{(i)}\right) = 0.5$$
$$(w_5 w_1 + w_6 w_2) x_1{}^{(i)} + (w_5 w_3 + w_6 w_4) x_2{}^{(i)} = 0$$

Thus, the decision boundary is linear (a straight line)

9D. B

Since $g^{[1]}$ is linear activation function and $g^{[2]}$ is Sigmoid activation function.

With the given neural network, $\hat{y}^{(i)}$ can be predicted as follows:

$$a_1{}^{(i)} = w_1 x_1{}^{(i)} + w_3 x_2{}^{(i)}$$
$$a_2{}^{(i)} = w_2 x_1{}^{(i)} + w_4 x_2{}^{(i)}$$
$$a_3{}^{(i)} = w_5 a_1{}^{(i)} + w_6 a_2{}^{(i)}$$
$$\hat{y}^{(i)} = \sigma\left(a_3{}^{(i)}\right)$$

Now, we need to use the binary cross-entropy loss:

$$L = \sum_{i=1}^{N} L_i$$

$$L_i = \frac{1}{N} BCE(y^{(i)}, \hat{y}^{(i)})$$

To compute the partial derivative of the loss with respect to w_1:

$$\frac{\partial L_i}{\partial w_1} = \frac{\partial L_i}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial {a_3}^{(i)}} \frac{\partial {a_3}^{(i)}}{\partial {a_1}^{(i)}} \frac{\partial {a_1}^{(i)}}{\partial w_1} = \frac{\partial L_i}{\partial \hat{y}^{(i)}} \sigma\left({a_3}^{(i)}\right)\left(1 - \sigma\left({a_3}^{(i)}\right)\right) w_5 {x_1}^{(i)}$$

The final value depends on the input feature ${x_1}^{(i)}$, so the partial derivative of loss function with respect to the weight can exceed 0.5. The input feature value affect the magnitude of the partial derivative for the weight.

Part 10: Convolutional Neural Network (6 questions, 10 marks)

Context

You are given a 3-channel image, and the details are provided below:
Channel 1:

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

Channel 2:

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

Channel 3:

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

Using this 3-channel image, answer questions 10A-10F.

10A. [1 mark] Suppose you use only Channel 1 as the input to a convolution layer called C_layer, which employs a kernel of shape 3x3, a stride of 1, no padding. What will be the output height and width? Output height: __1__ ; Output width: __2__

1. 2
2. 3

10B. [1 mark] The kernel of the C_layer (kernel size: 3×3, stride: 1, no padding) mentioned in the previous question is given below:

-1	1	2
1	1	-1
0	1	1

What is the value of the element in the bottom-right corner (i.e., last row and last column) of the output generated by C_layer when Channel 1 is used as input? __1__

1. 10

10C. [2 marks] A friend suggests using the following two convolution layers to process the Channel 1:
Layer 1: Uses a kernel of size 3x3, stride 1, and no padding.
Layer 2: Uses a kernel of size 2x2, stride 1, and no padding.
The receptive field size of a neuron in Layer 2 relative to the input is represented by NxN (N times N). What is the value for N? __1__

1. 4

10D. [2 marks] You must now use all three channels as input for a convolution layer. Apply the 2D convolution operation introduced in the lecture to produce a 5-channel output, using kernels with a height and width of 3. What is the total number of weights/parameters across all kernels in this convolution layer? 1

1. 135

10E. [2 marks] You need to generate an output of shape 4×5, where each element in the output is the average of the corresponding elements across the three input channels. For instance, the top-left output element should be the average of the top-left elements in channels 1, 2, and 3, i.e., (1+0+2)/3, and the bottom-right output element should be the average of the bottom-right elements in channels 1, 2, and 3, i.e., (2+0+0)/3. Which of the following settings would help you to generate the required output? Select all that apply.

A. One kernel with height set to 1 and width set to 1, all weight values set to 1/3, stride 3, padding 1.
B. One kernel with height set to 1 and width set to 1, all weight values set to 1, stride 1, no padding
C. One kernel with height set to 3 and width set to 1, all weight values set to 1/3, stride 1, no padding.
D. One kernel with height set to 1 and width set to 3, all weight values set to 1, stride 1, no padding.
✓E. None of the above.

10F. [2 marks] Regarding CNNs, which of the following statements is/are true? Select all that apply.

✓A. Convolution layers exploit local spatial correlations through convolution operations.
B. Pooling layers are used to increase the resolution of feature maps.
✓C. Dropout can be used to prevent overfitting when training CNN
D. None of the above.

Solution

10A. Output height: 2 ; Output width: 3
Input height: H=4
Input width: W=5
Kernel shape: K x K, K=3
Stride: S=1
No padding
The output height: $H_1 = \left\lfloor \frac{H-K}{S} \right\rfloor + 1 = \left\lfloor \frac{4-3}{1} \right\rfloor + 1 = 2$
The output width: $W_1 = \left\lfloor \frac{W-K}{S} \right\rfloor + 1 = \left\lfloor \frac{5-3}{1} \right\rfloor + 1 = 3$

10B. 10
Channel 1:

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

Kernel:

-1	1	2
1	1	-1
0	1	1

The value of the element in the bottom-right corner:

$$1 * (-1) + 3 * 1 + 2 * 2 + 1 * 1 + 1 * 1 + 1 * (-1) + 0 * 0 + 1 * 1 + 2 * 1 = 10$$

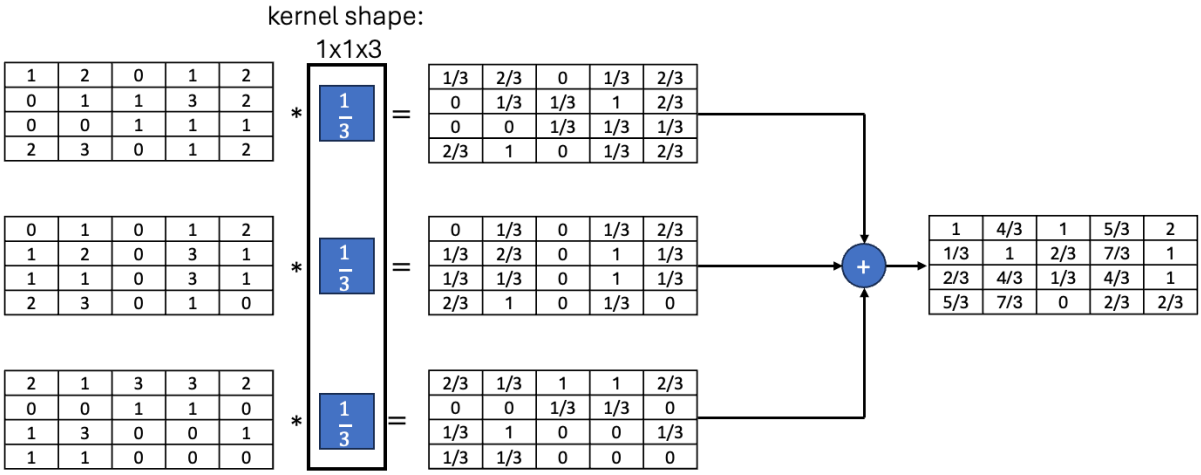
10C. 4
The receptive field r_i for a neuron in layer i is

$$r_i = r_{i-1} + (K_i - 1) \times j_{i-1}$$
$$j_i = j_{i-1} \times S_i$$

Here, K_i is the kernel size of layer i, and S_i is the stride of layer i. $r_0 = 1, j_0 = 1$.
So

$$r_1 = r_0 + (K_1 - 1) \times j_0 = 1 + (3 - 1) \times 1 = 3$$
$$j_1 = j_0 \times S_1 = 1 \times 1$$
$$r_2 = r_1 + (K_2 - 1) \times j_1 = 3 + (2 - 1) \times 1 = 4$$

The value for N is 4.
10D. 135
We now need to use all three channels as input, so the shape of one kernel is 3x3x3 (height x width x number of input channels).
In order to generate the 5-channel output, we need to use 5 kernels.
So the total number of weights across all kernels is 3x3x3x5=135.
10E. E
To generate the required output, we need one kernel with height set to 1 and width set to 1, all weight values set to $\frac{1}{3}$, stride 1, and no padding.



10F. A and C

Option A: Convolution layers apply kernels over small spatial regions, thereby exploiting local spatial correlations.

Option B: Pooling layers (e.g., max or average pooling) reduce the spatial resolution of feature maps.

Option C: Dropout randomly "drops" neurons' output during training, which helps prevent overfitting in CNNs

Part 11: Recurrent Neural Network (3 questions, 5 marks)

Context

Video activity recognition is a task in computer vision where a model identifies and classifies actions or activities occurring in a video. Suppose you are working with a dataset of human actions and need to design a model to predict whether the person in the video is "jumping" or not.

Based on this description, answer questions 11A-11C.

11A. [2 marks] Suppose you plan to first extract features for each video frame. What models can be used? Select all that apply.

✓A. CNN
✓B. RNN
✓C. Transformer Encoder
D. None of the above

11B. [2 marks] Features of multiple video frames can be treated as sequential input to the RNN model to predict the probability that the person in the video is jumping. Which of the following statements is/are true? Select all that apply.

✓A. For one RNN layer, the same set of weights is shared across all time steps to generate the outputs.
B. An RNN model can only take sequences of fixed length as input.
✓C. The feature dimension for each video frame should be the same.
D. None of the above

11C. [1 mark] You now need to design the architecture of your RNN model for this Video Activity Recognition task, which takes in the sequence of features from video frames as input. What type of RNN model should be employed?

A. One-to-Many
✓B. Many-to-One
C. Many-to-Many
D. None of the above

Solution

11A. A, B and C
Each video frame is an image, so CNN, RNN and Transformer Encoder can all be used to extract features.
11B. A and C
Option A: In one RNN layer, the same weight matrices are reused at every time step.
Option B: RNNs can process variable-length sequences.
Option C: Although the sequence length can vary, the feature dimensionality (i.e., the size of the input vector at each time step) must remain the same.
11C. B
Input: The sequence of features from video frames
Output: Whether the person in the video is "jumping" or not.
Therefore, we need a Many-to-One RNN model.

—END OF PAPER —