National University of Singapore

CS2109S—Introduction to AI and Machine Learning

Semester 2, 2021/2022

Time allowed: 2 hours

INSTRUCTIONS TO STUDENTS

1. Write down your **Student Number** on the answer sheet and shade completely the corresponding bubbles in the grid for each digit or letter. **DO NOT WRITE YOUR NAME!**
2. The assessment paper contains **SIX (6) questions** and comprises **TEN (10) pages** including this cover page.
3. Weightage of questions is given in square brackets. The maximum attainable score is 100.
4. This is a **OPEN-SHEET** assessment. You are allow one A4-sized double-sided cheat-sheet.
5. You are allowed to bring a calculator, but it cannot have any form of external communication capability, i.e. not Wifi- or 4G-enabled. Mobile phones and tablets are not allowed.
6. All questions must be answered in the space provided on the answer sheet; no extra sheets will be accepted as answers.
7. You are allowed to write with pencils, as long as it is legible.
8. **Marks may be deducted** for unrecognisable handwriting and/or for not shading the student number properly.
9. You must submit only the **ANSWER SHEET** and no other documents. The question set may be used as scratch paper.
10. An excerpt of the question may be provided in the answer sheet to aid you in answering in the correct box, where applicable. It is not the exact question. You should still refer to the original question in this question booklet.

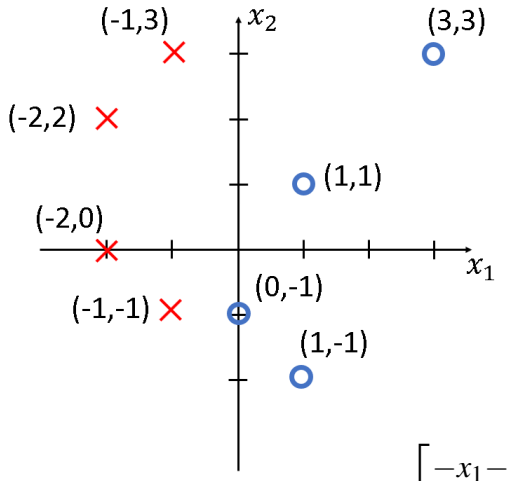
This page is intentionally left blank.

It may be used as scratch paper.

Question 1: Short Questions [14 marks]

A. Explain what you understand is the "kernel trick" and describe situations where we might want to use it. [3 marks]

B. Consider the 8 training data points in the figure below. There are 2 classes: circles and crosses. Suppose we apply an SVM to classify the data. Circle the support vector(s). [3 marks]



C. **[Alternative PCA Formulation]** Suppose that $\boldsymbol{X} = \begin{bmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_m- \end{bmatrix}$ is an array of the training samples in row format, then the following is an implementation of the SVD-based approach to PCA discussed in lecture to transform the samples x_i from n degrees to k degrees ($k < n$, by performing SVD on the "centred" covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$:

```
sigma= X.T @ X
U, s, VT = svd(sigma)
Ur = U[:,:k]
Z = X @ Ur
Xapprox = Z @ Ur.T
```

It turns out that there is an alternative formulation where we can perform SVD *directly* on $\boldsymbol{X}$ to yield the *same(!)* results:

```
U, s, VT = svd(X)
vectors = VT[:k]
Z = X @ vectors.T
Xapprox2 = Z @ vectors
```

What do you think is the practical difference between these 2 implementations and why do you think we would prefer one formulation over the other. You are expected to reason about this from first principles. [4 marks]

D. You apply Gradient Descent to a ML-problem and you found that the training takes a long time. Suggest 2 ways that might be able to help you reduce the training time without sacrificing too much accuracy. [4 marks]

Question 2: COVID-Rho [31 marks]

After the latest wave of the Omicron variant subsided, scientists were horrified to discovered a new variant, for which existing ART tests no longer work! They call this the *Rho* variant. The symptoms of the new variant, while similar to previous variants are not entirely the same. Since existing ART tests no longer work, the government needs a new way to determine if the sick people who come to A&E are likely to have contracted the new variant.

Since you have taken CS2109S, you were approached to construct a decision tree classifier. Unfortunately, there is not much data available since and all you have are the data for 12 patients shown in the table below. There are 3 possible outcomes for the current detailed lab analysis: Positive (for Rho), Negative (for Rho), and Other (some other non-Rho COVID variant). Tie breaking order on constructing tree is No Taste $>$ Tiredness $>$ Cough $>$ Fever.

Patient	Fever?	Cough?	Tiredness?	No Taste?	Rho?
A	Yes	No	Yes	No	Negative
B	No	Yes	Yes	Yes	Positive
C	Yes	Yes	No	Yes	Other
D	Yes	No	No	Yes	Other
E	Yes	No	Yes	Yes	Positive
F	Yes	Yes	No	No	Negative
G	No	Yes	Yes	No	Positive
H	No	Yes	Yes	No	Negative
I	Yes	No	Yes	Yes	Positive
K	No	Yes	No	Yes	Negative
L	No	No	No	No	Negative
M	No	No	Yes	Yes	Positive

A summary of count for these results is given in the following table for your convenience:

Summary	Fever?	Cough?	Tiredness?	No Taste?
Yes/Positive	2	2	5	4
Yes/Negative	2	3	2	1
Yes/Other	2	1	0	2
No/Positive	3	3	0	1
No/Negative	3	2	3	4
No/Other	0	1	2	0
Total	12	12	12	12

The information content (entropy) for a given probability distribution p_i, for $i = 1, \cdots, n$ is given by:

$$-\sum_{i=1}^{n} p_i \log_2(p_i)$$

A. **[Warm Up]** What is the entropy of the COVID outcomes?
Hint: remember the log *is in base 2!* [3 marks]

B. Construct a decision tree using information gain to predict whether a given patient is likely to have COVID-Rho. [15 marks]

We break the tie in favour of Cough over Fever.

C. There is likely some error in the initial data leading to overfitting. Derive a pruned decision tree that has at least 2 training data points for each leaf. [4 marks]

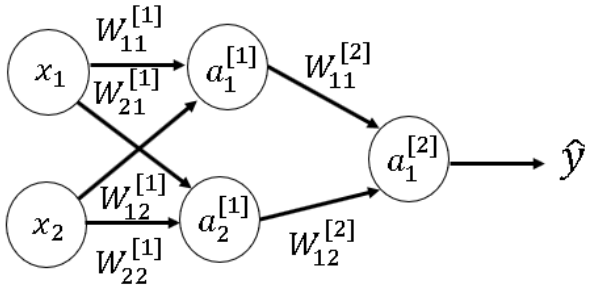
D. Given the following test data, what is the precision, recall and F1 value for the pruned decision tree in Part(C). You can consider "Other" to be "Negative." [6 marks]

Patient	Fever?	Cough?	Tiredness?	No Taste?	Ground Truth
1	Yes	Yes	No	No	Negative
2	No	No	No	No	Negative
3	Yes	Yes	No	Yes	Other
4	Yes	No	No	Yes	Positive
5	No	Yes	No	Yes	Other
6	Yes	No	Yes	No	Negative
7	No	Yes	Yes	No	Negative
8	No	Yes	Yes	No	Negative
9	No	Yes	Yes	Yes	Positive
10	Yes	No	Yes	Yes	Positive

E. Suppose you are told that COVID-Rho is very very deadly compared to the other variants, which would you try to maximize: precision or recall? Explain. [3 marks]

Question 3: BackProp Calculations [20 marks]

Consider the following simple 2-layer MLP network with 1 hidden layer and no bias terms:



Following the notation used in lecture:

$$\begin{aligned} \boldsymbol{f}^{[1]} &= (\boldsymbol{W}^{[1]})^T \boldsymbol{x} \\ \boldsymbol{a}^{[1]} &= g^{[1]}(\boldsymbol{f}^{[1]}) \\ f^{[2]} &= (\boldsymbol{W}^{[2]})^T \boldsymbol{a}^{[1]} \\ \hat{y} &= g^{[2]}(f^{[2]}) \end{aligned}$$

All the activation functions are sigmoids given by:

$$g^{[1]}(x) = g^{[2]}(x) = \sigma(x) = \frac{1}{1+e^{-x}}$$

For reference,

$$\begin{aligned} \sigma'(x) &= \sigma(x)(1-\sigma(x)) \\ \boldsymbol{W}^{[1]} &= \begin{bmatrix} W_{11}^{[1]} & W_{21}^{[1]} \\ W_{12}^{[1]} & W_{22}^{[1]} \end{bmatrix} \\ \boldsymbol{W}^{[2]} &= \begin{bmatrix} W_{11}^{[2]} \\ W_{12}^{[2]} \end{bmatrix} \end{aligned}$$

Recall that our goal is to compute $\frac{\partial \hat{y}}{\partial \boldsymbol{W}}$ so that we can use it for Gradient Descent. Also, where L is the maximum layer of an MLP (here, $L = 2$), we define $\boldsymbol{\delta}^{[l]}$ as follows:

$$\boldsymbol{\delta}^{[l]} = \frac{\partial \hat{y}}{\partial \boldsymbol{f}^{[l]}} \tag{1}$$

A. **[Warm Up: Forward Propagation]** Express $\hat{y}$ in terms of $\boldsymbol{x}$, $\boldsymbol{W}^{[1]}$, $\boldsymbol{W}^{[2]}$ and $\boldsymbol{\sigma}$. [2 marks]

B. Express $\boldsymbol{\delta}^{[2]}$ in terms of $\boldsymbol{f}^{[i]}$ and σ, where $i = 1$ or 2. [2 marks]

C. Express $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}} = \frac{\partial f^{[2]}}{\partial \boldsymbol{W}^{[2]}} \frac{\partial \hat{y}}{\partial f^{[2]}}$ in terms of $\boldsymbol{a}^{[i]}$, and σ, where $i = 1$ or 2. [2 marks]

D. Express $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$ in terms of $\boldsymbol{\delta}^{[i]}$ and $\boldsymbol{a}^{[i]}$, where $i = 1$ or 2. [2 marks]

E. Derive an expression for $\boldsymbol{\delta}^{[1]} = (\frac{\partial \boldsymbol{a}^{[1]}}{\partial \boldsymbol{f}^{[1]}})^T \frac{\partial f^{[2]}}{\partial \boldsymbol{a}^{[1]}} \frac{\partial \hat{y}}{\partial f^{[2]}}$ in terms of $\boldsymbol{\delta}^{[2]}$ from Equation (1). [3 marks]

F. Express $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}} = \frac{\partial \boldsymbol{f}^{[1]}}{\partial \boldsymbol{W}^{[1]}} (\frac{\partial \hat{y}}{\partial \boldsymbol{f}^{[1]}})^T$ in terms of $\boldsymbol{x}^{[0]}$, and $\boldsymbol{\delta}^{[i]}$, where $i = 1, 2$. [3 marks]

G. Given $\boldsymbol{x}^{[0]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\boldsymbol{W}^{[1]} = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$ and $\boldsymbol{W}^{[2]} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, compute $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}}$ and $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$ (to 3 significant digits). [6 marks]

Question 4: Unsupervised Learning [12 marks]

In this problem, you and your team are are conducting a survey on the population on behalf of the government. The survey consists of 1000 questions, ranging from simple yes/no questions, to approval ratings from 0-10. Your team has managed to gather over one million responses.

A. Your team now wants to identify if there are natural clusters found within the population, and intend to do so by performing k-means on the raw data. Suggest a method to go about identifying a good value of k to select. [4 marks]

B. In order to initialize the centroids to run k-means, Ben Bitdiddle offers you a strategy to initialize the centroids. Randomly assign every sample to one of the k clusters, and compute the k centroids of this random configuration to use as the initial centroids. Explain why this might not be a good method of random initialization of centroids. [4 marks]

C. You have run k-means on your dataset but realised that the clusters do not make much sense. What could have caused this issue, and how can you resolve it? [4 marks]

Question 5: Convolutional Neural Networks [20 marks]

You are given the following rectangular image $\mathbf{x} \in \mathbb{R}^{3\times6}$:

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R

A. If we use k 3×6 kernels, equal to the dimensions of the input image, is this simply a fully connected layer with k output neurons? Explain. [2 marks]

B. You are provided a single 3×3 kernel as shown below. Assume we apply no padding and a stride of 2×2. What is the value of the bottom-right output pixel in terms of the variables $\in$ {A, B, ..., Q, R} after applying this kernel. You should not flip the kernel. [2 marks]

1	0	-1
1	0	-1
1	0	-1

C. Using the same 3×3 kernel as above, assume we apply a zero-padding of 1 and a stride of 1×1. What is the value of the bottom-right output pixel now? You should not flip the kernel. [2 marks]

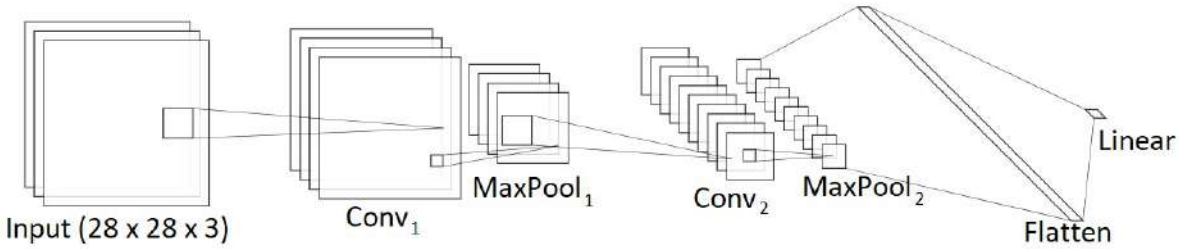
D. Your class has been tasked to solve a image binary classification problem. Your friend Ben Bitdiddle claims that his deep neural network consisting of multiple fully connected layers will be able to solve this image classification problem.

You disagree and feel that we should use convolutional layers early in the network in order to solve the problem. Who is correct? Explain. [3 marks]

E. You design a CNN binary classifier that takes in a 3-channel colour 28x28 image as the input. The model is described as a sequence of layers:

$\text{Conv}_1(4,(5,5)) \rightarrow \text{MaxPool}_1 \rightarrow \text{Conv}_2(10,(5,5)) \rightarrow \text{MaxPool}_2 \rightarrow \text{Flatten} \rightarrow \text{Linear}(x,1)$

Conv_1 applies a 5×5 kernel with 0 padding and stride 1, and has 4 output channels.
Conv_2 applies a 5×5 kernel with 0 padding and stride 1, and has 10 output channels.
Both MaxPool layers apply a 2×2 kernel with a stride equal to the kernel dimensions (2×2).
The Linear layer is fully connected from x nodes (to be determined) in the flattened layer to 1 output node for binary classification.



State the dimensions of each layer, and the number of trainable parameters. [6 marks]

F. For the image binary classification problem, you have been provided 1,000 labelled images, 900 of which are labelled class A, and the remaining are labelled class B. Describe in a step-by-step format, how you will set up your experiment to train and evaluate the performance of your classifier. [5 marks]

Question 6: Did you see the animal? [3 marks]

In the "Invisible Gorilla" experiment of 1999, volunteers were asked to watch a video where two groups of people—some dressed in white, some in black—are passing basketballs around. The volunteers were asked to count the passes among players dressed in white while ignoring the passes of those in black. About half of watchers missed a person in a gorilla suit walking in and out of the scene thumping its chest.

In CS2109S, we attempted to verify the results of this study. During one of the lectures, Prof Ben transformed into an animal. What animal was that?

If you cannot remember the animal, you can also try to tell us a joke. If it's funny enough, maybe you might get some points too. [3 marks]

Appendix

The following is one of the algorithms that was introduced in class that is reproduced here for your reference.

function DECISION-TREE-LEARNING(*examples, attributes, default*) **returns** a decision tree

inputs: *examples*, set of examples

attributes, set of attributes

default, default value for the goal predicate

if *examples* is empty **then return** *default*

else if all *examples* have the same classification **then return** the classification

else if *attributes* is empty **then return** MAJORITY-VALUE(*examples*)

else

$best \leftarrow$ CHOOSE-ATTRIBUTE(*attributes*, *examples*)

$tree \leftarrow$ a new decision tree with root test *best*

for each value v_i of *best* **do**

$examples_i \leftarrow$ {elements of *examples* with $best = v_i$}

subtree $\leftarrow$ DECISION-TREE-LEARNING($examples_i$, $attributes - best$, MAJORITY-VALUE(*examples*))

add a branch to *tree* with label v_i and subtree *subtree*

end

return *tree*

— HAPPY HOLIDAYS —

CS2109S—Introduction to AI and Machine Learning

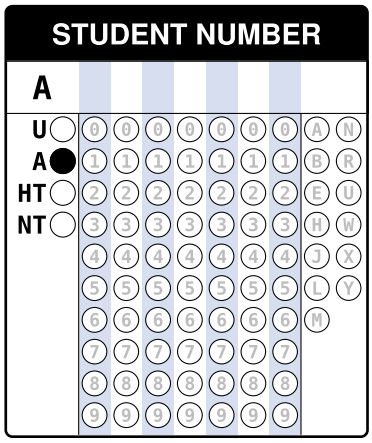
Final Assessment Answer Sheet

Semester 2, 2021/2022

Time allowed: 2 hours

Instructions (please read carefully):

1. Write down your **Student Number** on the right and using ink or pencil, **shade completely** the corresponding bubbles in the grid for each digit or letter. **DO NOT WRITE YOUR NAME!**
2. This answer booklet comprises **THIRTEEN (13) pages**, including this cover page.
3. This is a **OPEN-SHEET** assessment. You are allow one A4-sized double-sided cheatsheet.
4. Weightage of questions is given in square brackets. The maximum attainable score is 100.
5. You are allowed to bring a calculator, but it cannot have any form of external communication capability, i.e. not Wifi- or 4G-enabled. Mobile phones and tablets are not allowed.
6. All questions must be answered in the space provided on the answer sheet; no extra sheets will be accepted as answers.
7. You are allowed to write with pencils, as long as it is legible.
8. **Marks may be deducted** for unrecognisable handwriting and/or for not shading the student number properly.
9. You must submit only the **ANSWER SHEET** and no other documents. The question set may be used as scratch paper.
10. An excerpt of the question may be provided to aid you in answering in the correct box, where applicable. It is not the exact question. You should still refer to the original question in the question booklet.

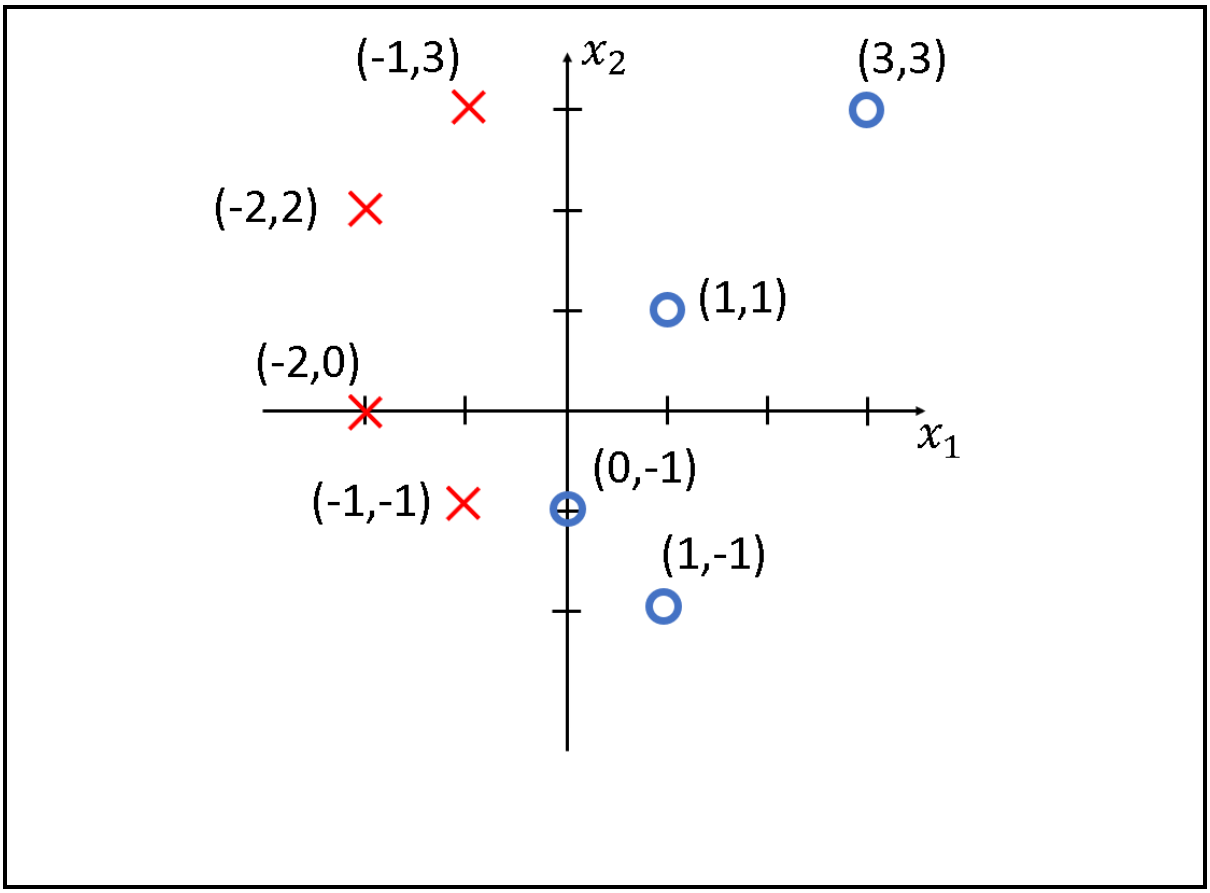


For Examiner's Use Only

Question	Marks
Q1	/ 14
Q2	/ 31
Q3	/ 20
Q4	/ 12
Q5	/ 20
Q6	/ 3
Total	/100

Question 1A What is the "kernel trick"? [3 marks]

Question 1B Circle the support vector(s). [3 marks]



Question 1C Alternative PCA Formulation. [4 marks]

Question 1D Speeding up Gradient Descent. [4 marks]

Question 2A Entropy of the COVID outcomes [3 marks]

Question 2B Decision tree to decide if patient has COVID-Rho. [15 marks]

Question 2B Decision tree to decide if patient has COVID-Rho (continued). [15 marks]

Question 2B Decision tree to decide if patient has COVID-Rho (continued). [15 marks]

Question 2C Pruned decision tree (min-leaf ≥ 2). [4 marks]

Question 2D Precision, recall and F1 value. [6 marks]

Question 2E Choice between precision and recall. [3 marks]

Question 3A Warm Up: Forward Propagation [2 marks]

Question 3B Find $\boldsymbol{\delta}^{[2]}$. [2 marks]

Question 3C Find $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$. [2 marks]

Question 3D Find $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$(again). [2 marks]

Question 3E Find $\boldsymbol{\delta}^{[1]}$. [3 marks]

Question 3F Find $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}}$. [3 marks]

Question 3G Compute $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}}$ and $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$ (to 3 significant digits). [6 marks]

Question 4A Suggest a method to select a good value of k. [4 marks]

Question 4B Explain why this random initialization method might not be good. [4 marks]

Question 4C What is causing poor k-means performance and how to resolve? [4 marks]

Question 5A Does a full 3x6 kernel behave as a fully connected layer? Explain. [2 marks]

Question 5B Value of the bottom-right output pixel? Padding 0, stride 2. [2 marks]

Question 5C Value of the bottom-right output pixel? Padding 1, stride 1. [2 marks]

Question 5D Should we use CNN or not? Explain. [3 marks]

Question 5E CNN classifier dimensions and number of parameters. [6 marks]

Layer	Layer dimension	Number of trainable parameters
Input	$28 \times 28 \times 3$	0
Conv_1		
MaxPool_1		
Conv_2		
MaxPool_2		
Linear	$\times 1$	

Question 5F Describe your CNN experiment setup. [5 marks]

Question 6 Did you see the animal? [3 marks]

This page is intentionally left blank.

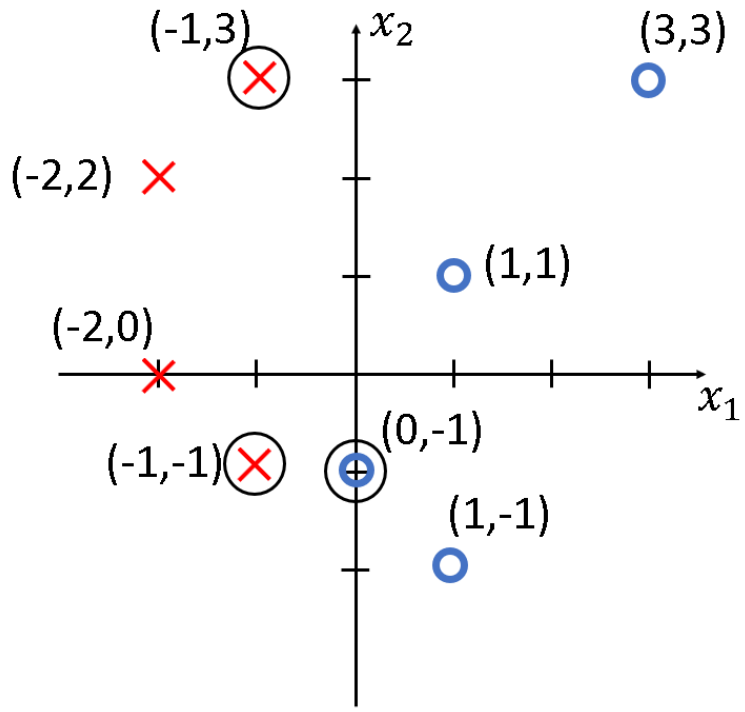
It may be used as scratch paper.

— END OF ANSWER SHEET —

Question 1A What is the "kernel trick"? [3 marks]

A kernel is a function that maps the training set into a different space where the data points might then become linearly separable. In lectures, we mentioned the use of this technique in both polynomial regression and for SVM.

Question 1B Circle the support vector(s). [3 marks]



Question 1C Alternative PCA Formulation. [4 marks]

The covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$ is of size $n \times n$, while the data matrix $\boldsymbol{X}$ is of size $m \times n$. Which formulation is preferred depends on the relative sizes of m and n because it impacts the time taken for the SVD computation. Note that there's a cost to computing $\boldsymbol{X}^T\boldsymbol{X}$ as well.

Question 1D Speeding up Gradient Descent. [4 marks]

Many possibilities including:

- Perform feature scaling and mean normalization on the training data.
- Apply PCA to reduce the dimensionality of the training data, but taking care to keep enough variance, i.e. at least 99%.
- Use a larger learning rate or an adaptive learning rate to speed up convergence.
- If the dataset is too large, possibly run stochastic gradient descent or mini-batch gradient descent instead for faster iterations.

Question 2A Entropy of the COVID outcomes [3 marks]

The outcomes counts for Positive, Negative and Other are 5, 5, and 2.

$$\begin{aligned} \text{Entropy} &= -\frac{5}{12}\log(\frac{5}{12}) - \frac{5}{12}\log(\frac{5}{12}) - \frac{2}{12}\log(\frac{2}{12}) \\ &= 0.526 + 0.526 + 0.431 \\ &= 1.483 \end{aligned}$$

Question 2B Decision tree to decide if patient has COVID-Rho. [15 marks]

We just need to consider the remaining entropy.

$$
\begin{aligned}
\text{Remainder} &= \sum_{i=1}^{v} \frac{p_i+n_i+o_i}{p+n+o} I\left(\frac{p_i}{p_i+n_i+o_i}, \frac{n_i}{p_i+n_i+o_i}, \frac{o_i}{p_i+n_i+o_i}\right) \\
\text{Remainder(Fever)} &= \frac{6}{12} I\left(\frac{2}{6}, \frac{2}{6}, \frac{2}{6}\right) + \frac{6}{12} I\left(\frac{3}{6}, \frac{3}{6}, 0\right) \\
&= -\frac{1}{2}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) \times 3\right) - \frac{1}{2}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) \times 2\right) \\
&= 1.292 \\
\text{Remainder(Cough)} &= \frac{6}{12} I\left(\frac{2}{6}, \frac{3}{6}, \frac{1}{6}\right) + \frac{6}{12} I\left(\frac{3}{6}, \frac{2}{6}, \frac{1}{6}\right) \\
&= -\frac{1}{2}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) \times 2 \\
&= 1.459 \\
\text{Remainder(Tiredness)} &= \frac{7}{12} I\left(\frac{5}{7}, \frac{2}{7}, 0\right) + \frac{5}{12} I\left(\frac{3}{5}, \frac{2}{5}, 0\right) \\
&= -\frac{7}{12}\left(\frac{5}{7}\log_2\left(\frac{5}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right) - \frac{5}{12}\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) \\
&= 0.908 \\
\text{Remainder(No Taste)} &= \frac{7}{12} I\left(\frac{4}{7}, \frac{1}{7}, \frac{2}{7}\right) + \frac{5}{12} I\left(\frac{1}{5}, \frac{4}{5}, 0\right) \\
&= -\frac{7}{12}\left(\frac{4}{7}\log_2\left(\frac{5}{7}\right) + \frac{1}{7}\log_2\left(\frac{1}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right) \\
&\quad -\frac{5}{12}\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right) \\
&= 1.105
\end{aligned}
$$

Hence, the root node should be “Tiredness?”.

Question 2B Decision tree to decide if patient has COVID-Rho (continued). [15 marks]

Thereafter, we have the following sample points remaining for "Tiredness? Yes ":

Patient	Fever?	Cough?	No Taste?	Rho?
A	Yes	No	No	Negative
B	No	Yes	Yes	Positive
E	Yes	No	Yes	Positive
G	No	Yes	No	Positive
H	No	Yes	No	Negative
I	Yes	No	Yes	Positive
M	No	No	Yes	Positive

Again, we just need to consider the remaining entropy.

$$\begin{aligned}
\text{Remainder(Fever)} &= \frac{3}{7}I(\frac{1}{3},\frac{2}{3})+\frac{4}{7}I(\frac{1}{4},\frac{3}{4}) \\
&= -\frac{3}{7}(\frac{1}{3}\log_2(\frac{1}{3})+\frac{2}{3}\log_2(\frac{2}{3}))-\frac{4}{7}(\frac{1}{4}\log_2(\frac{1}{4})+\frac{3}{4}\log_2(\frac{3}{4})) \\
&= 0.857 \\
\text{Remainder(Cough)} &= \frac{3}{7}I(\frac{1}{3},\frac{2}{3})+\frac{4}{7}I(\frac{1}{4},\frac{3}{4}) \\
&= -\frac{3}{7}(\frac{1}{3}\log_2(\frac{1}{3})+\frac{2}{3}\log_2(\frac{2}{3}))-\frac{4}{7}(\frac{1}{4}\log_2(\frac{1}{4})+\frac{3}{4}\log_2(\frac{3}{4})) \\
&= 0.857 \\
\text{Remainder(No Taste)} &= \frac{4}{7}I(\frac{4}{4},0)+\frac{3}{7}I(\frac{1}{3},\frac{2}{3}) \\
&= -\frac{4}{7}(\frac{4}{4}\log_2(\frac{4}{4}))-\frac{3}{7}(\frac{1}{3}\log_2(\frac{1}{3})+\frac{2}{3}\log_2(\frac{2}{3})) \\
&= 0.394
\end{aligned}$$

Hence, the next attribute should be "No Taste?".

Next, we have the following sample points remaining for "Tiredness? No ":

Patient	Fever?	Cough?	No Taste?	Rho?
C	Yes	Yes	Yes	Other
D	Yes	No	Yes	Other
F	Yes	Yes	No	Negative
K	No	Yes	Yes	Negative
L	No	No	No	Negative

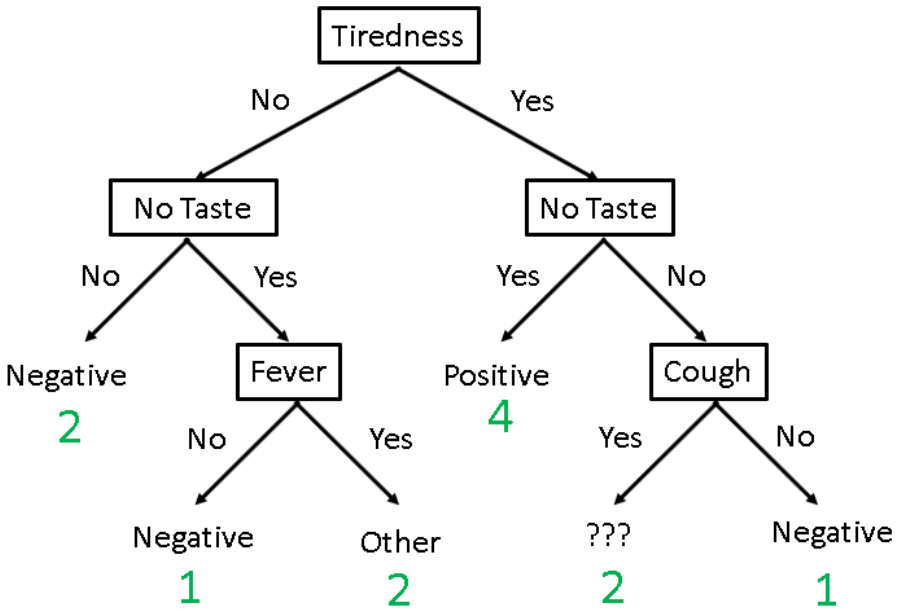
Question 2B Decision tree to decide if patient has COVID-Rho (continued). [15 marks]

Again, we just need to consider the remaining entropy.

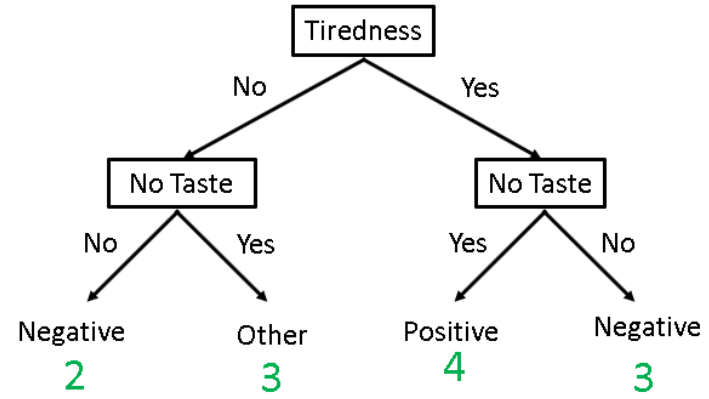
$$\begin{aligned}
\text{Remainder(Fever)} &= \frac{3}{5}I(\frac{1}{3},\frac{2}{3})+\frac{2}{5}I(0,1) \\
&= -\frac{3}{5}(\frac{1}{3}\log_2(\frac{1}{3})+\frac{2}{3}\log_2(\frac{2}{3})) \\
&= 0.551 \\
\text{Remainder(Cough)} &= \frac{3}{5}I(\frac{2}{3},\frac{1}{3})+\frac{2}{5}I(\frac{1}{2},\frac{1}{2}) \\
&= -\frac{3}{5}(\frac{1}{3}\log_2(\frac{1}{3})+\frac{2}{3}\log_2(\frac{2}{3}))-\frac{2}{5}(\frac{1}{2}\log_2(\frac{1}{2})\times 2) \\
&= 0.951 \\
\text{Remainder(No Taste)} &= \frac{3}{5}I(\frac{1}{3},\frac{2}{3})+\frac{2}{5}I(1,0) \\
&= -\frac{3}{5}(\frac{1}{3}\log_2(\frac{1}{3})+\frac{2}{3}\log_2(\frac{2}{3}) \\
&= 0.551
\end{aligned}$$

Hence, the next attribute can either be "No Taste?" or "Fever," but you are expected to break ties in favour of "No Taste?".

We can deduce the last attribute by inspection and derive the following decision tree:



Question 2C Pruned decision tree (min-leaf ≥ 2). [4 marks]



Question 2D Precision, recall and F1 value. [6 marks]

The prediction for patient 4 is wrong. It is a false negative.

	Pred -ve	Pred +ve
Actual -ve	$TN = 7$	$FP = 0$
Actual +ve	$FN = 1$	$TP = 2$

$$\begin{aligned}
\text{Precision} &= \frac{TP}{TP+FP} \\
&= \frac{2}{2+0} \\
&= 1 \\
\text{Recall} &= \frac{TP}{TP+FN} \\
&= \frac{2}{2+1} \\
&= \frac{2}{3} \\
\text{F1} &= \frac{2}{\frac{1}{P}+\frac{1}{R}} \\
&= \frac{4}{5}
\end{aligned}$$

Question 2E Choice between precision and recall. [3 marks]

If COVID-Rho is very dangerous, it can be potentially fatal if we false to detect it accurate. We want to avoid false negatives, so we want to maximize *recall*.

Question 3A Warm Up: Forward Propagation [2 marks]

$$\begin{aligned} \hat{y} &= g^{[2]}({\boldsymbol{W}^{[2]}}^T(g^{[1]}({\boldsymbol{W}^{[1]}}^T\boldsymbol{x}))) \\ &= \sigma({\boldsymbol{W}^{[2]}}^T(\sigma({\boldsymbol{W}^{[1]}}^T\boldsymbol{x}))) \end{aligned}$$

Question 3B Find $\boldsymbol{\delta}^{[2]}$. [2 marks]

$$\begin{aligned} \boldsymbol{\delta}^{[2]} &= \frac{\partial \hat{y}}{\partial \boldsymbol{f}^{[2]}} \\ &= \left[\sigma(\boldsymbol{f}^{[2]})(1-\sigma(\boldsymbol{f}^{[2]}))\right], \text{ since } \hat{y} = \sigma(\boldsymbol{f}^{[2]}) \end{aligned}$$

Question 3C Find $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$. [2 marks]

$$\begin{aligned} \frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}} &= \frac{\partial f^{[2]}}{\partial \boldsymbol{W}^{[2]}}\frac{\partial \hat{y}}{\partial f^{[2]}} \\ &= \boldsymbol{a}^{[1]}\sigma(\boldsymbol{f}^{[2]})(1-\sigma(\boldsymbol{f}^{[2]})), \text{ since } \boldsymbol{f}^{[2]} = (\boldsymbol{W}^{[2]})^T\boldsymbol{a}^{[1]} \end{aligned}$$

Question 3D Find $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$(again). [2 marks]

$$\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}} = \boldsymbol{a}^{[1]}\boldsymbol{\delta}^{[2]}$$

Question 3E Find $\boldsymbol{\delta}^{[1]}$. [3 marks]

$$\begin{aligned}
\boldsymbol{\delta}^{[1]} &= \frac{\partial \hat{y}}{\partial f^{[1]}} \\
&= (\frac{\partial \boldsymbol{a}^{[1]}}{\partial \boldsymbol{f}^{[1]}})^T \frac{\partial \boldsymbol{f}^{[2]}}{\partial \boldsymbol{a}^{[1]}} \frac{\partial \hat{y}}{\partial f^{[2]}} \\
&= (\frac{\partial \boldsymbol{a}^{[1]}}{\partial \boldsymbol{f}^{[1]}})^T \boldsymbol{W}^{[2]}\boldsymbol{\delta}^{[2]}, \text{ since } f^{[2]} = (\boldsymbol{W}^{[2]})^T \boldsymbol{a}^{[1]} \\
&= \begin{bmatrix} \frac{\partial a_1^{[1]}}{\partial f_1^{[1]}} & \frac{\partial a_1^{[1]}}{\partial f_2^{[1]}} \\ \frac{\partial a_2^{[1]}}{\partial f_1^{[1]}} & \frac{\partial a_2^{[1]}}{\partial f_2^{[1]}} \end{bmatrix}^T \boldsymbol{W}^{[2]}\boldsymbol{\delta}^{[2]}, \text{ since } \boldsymbol{a}^{[1]} = \sigma(\boldsymbol{f}^{[1]}) \\
&= \begin{bmatrix} \sigma(f_1^{[1]})(1-\sigma(f_1^{[1]})) & 0 \\ 0 & \sigma(f_2^{[1]})(1-\sigma(f_2^{[1]})) \end{bmatrix} \boldsymbol{W}^{[2]}\boldsymbol{\delta}^{[2]}, \text{ since } \boldsymbol{a}^{[1]} = \sigma(\boldsymbol{f}^{[1]})
\end{aligned}$$

Since the sigmoid function is element-wise, $\sigma(f^{[1]})(1-\sigma(f^{[1]}))\boldsymbol{W}^{[2]}\boldsymbol{\delta}^{[2]}$ is also good.

Question 3F Find $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}}$. [3 marks]

$$\begin{aligned}
\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}} &= \frac{\partial \boldsymbol{f}^{[1]}}{\partial \boldsymbol{W}^{[1]}} ((\frac{\partial \boldsymbol{a}^{[1]}}{\partial \boldsymbol{f}^{[1]}})^T \frac{\partial f^{[2]}}{\partial \boldsymbol{a}^{[1]}} \frac{\partial \hat{y}}{\partial f^{[2]}})^T \\
&= \frac{\partial \boldsymbol{f}^{[1]}}{\partial \boldsymbol{W}^{[1]}} (\boldsymbol{\delta}^{[1]})^T \\
&= \boldsymbol{x}^{[0]}(\boldsymbol{\delta}^{[1]})^T, \text{ since } \boldsymbol{f}^{[1]} = (\boldsymbol{W}^{[1]})^T \boldsymbol{x}^{[0]}
\end{aligned}$$

Question 3G Compute $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}}$ and $\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$ (to 3 significant digits). [6 marks]

$$
\begin{aligned}
\boldsymbol{f}^{[1]} &= (\boldsymbol{W}^{[1]})^T \boldsymbol{x}^{[0]} \\
&= \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 3 \\ 1 \end{bmatrix} \\
\boldsymbol{a}^{[1]} &= \sigma(\boldsymbol{f}^{[1]}) \\
&= \begin{bmatrix} \sigma(3) \\ \sigma(1) \end{bmatrix} = \begin{bmatrix} 0.953 \\ 0.731 \end{bmatrix} \\
\boldsymbol{f}^{[2]} &= (\boldsymbol{W}^{[2]})^T \boldsymbol{a}^{[1]} \\
&= \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} \sigma(3) \\ \sigma(1) \end{bmatrix} \\
&= \begin{bmatrix} 3\sigma(3) + 4\sigma(1) \end{bmatrix} = [5.781] \\
\boldsymbol{a}^{[2]} &= \sigma(\boldsymbol{f}^{[2]}) \\
&= \begin{bmatrix} \sigma(5.781) \end{bmatrix} \\
\boldsymbol{\delta}^{[2]} &= \begin{bmatrix} \sigma(5.781)(1-\sigma(5.781)) \end{bmatrix} = [0.00306] \\
\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}} &= \sigma(\boldsymbol{f}^{[2]})(1-\sigma(\boldsymbol{f}^{[2]}))\boldsymbol{a}^{[1]} \\
&= 0.00306 \begin{bmatrix} \sigma(3) \\ \sigma(1) \end{bmatrix} = \begin{bmatrix} 0.00291 \\ 0.00224 \end{bmatrix} \\
\boldsymbol{\delta}^{[1]} &= 0.00306 \begin{bmatrix} \sigma(3)(1-\sigma(3)) & 0 \\ 0 & \sigma(1)(1-\sigma(1)) \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \\
&= \begin{bmatrix} 0.000415 \\ 0.00241 \end{bmatrix} \\
\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[1]}} &= \boldsymbol{x}^{[0]} (\boldsymbol{\delta}^{[1]})^T \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0.000415 & 0.00241 \end{bmatrix} \\
&= \begin{bmatrix} 0.000415 & 0.00241 \\ 0.000415 & 0.00241 \end{bmatrix}
\end{aligned}
$$

Notice the vanishing gradients!

Question 4A Suggest a method to select a good value of k. [4 marks]

+1 Run k-means on multiple k, and compute the loss for each configuration.
+3 Plot the loss against the k-values, and apply the elbow method to select a good k.

Question 4B Explain why this random initialization method might not be good. [4 marks]

+3 By assigning samples to random clusters, each cluster's centroid would be found somewhere close to the mean of the dataset.
+1 Good random initializations of centroids usually leverage on the centroids being far away from each other/exploiting locality.
+1 Poor initialization of the clusters will increase likelihood of converging to suboptimal local minimas.
Max 4m, no marks awarded for stating that number of iterations will increase, or that random assignment of points to clusters will result in random final centroids.

Question 4C What is causing poor k-means performance and how to resolve? [4 marks]

If student mentions any of the following reasons, max 2m altogether:
2m Poor initialization of the clusters, run more iterations with different initial centroids.
2m Bad choice of k, use elbow method to select a good k.
2m Noise in the dataset, re-take the data or manual cleaning of the data.

The root cause of poor performance in a large dataset with many features is the Curse of Dimensionality, stating any the following reasons, max 4m altogether:
4m Features are different scales (i.e. some values are boolean ranging from 0-1, while others are in the 0-10 range) and will affect the clustering, normalize the data to resolve.
4m Large number of features, which means that the model suffers from the curse of dimensionality, run PCA to reduce the number of dimensions.
Marks will be deducted for not stating the problem without a reasonable resolution.

Question 5A Does a full 3x6 kernel behave as a fully connected layer? Explain. [2 marks]

Yes it is similar to a fully connected layer, but without bias terms. Since a fully connected layer has equation $W^T x + b$, then the n-by-n kernel multiplying by the image would behave as $W^T x$, without the bias term, of with $b = 0$.

`+1` Explain that each kernel will produce a single output from the entire image, k kernels will produce a k outputs similar to k output neurons.
`+1` Note that the bias term is 0 in this CNN case.

Question 5B Value of the bottom-right output pixel? Padding 0, stride 2. [2 marks]

C+I+O-E-K-Q

Note that since we have no padding and a stride of 2, the output is only a 1×2 array, and the column with F, L, R is not reached.

Question 5C Value of the bottom-right output pixel? Padding 1, stride 1. [2 marks]

K+Q

First, pad the original image with a layer of 0 values, then slide the kernel to the bottom-right.

Question 5D Should we use CNN or not? Explain. [3 marks]

Both networks would be able to solve the problem, but CNNs are better suited. CNN kernels **leverage on spatial context** (by only assigning weights to nearby pixels) to extract local features (as opposed capturing global features in a fully connected layer), enjoy **translation invariance**, so features appearing in one part of the image can also be recognized elsewhere, and **have fewer parameters**, as the kernels share weights, which means training is faster.

Mentioning any of the three points is sufficient to get `3m`, but justification provided should be stated clearly, otherwise partial marks will be awarded instead.

Question 5E CNN classifier dimensions and number of parameters. [6 marks]

Layer	Layer dimension	Number of trainable parameters
Input	$28 \times 28 \times 3$	0
Conv_1	$24 \times 24 \times 4$	$(5 \times 5 \times 3 + 1) \times 4 = 304$
MaxPool_1	$12 \times 12 \times 4$	0
Conv_2	$8 \times 8 \times 10$	$(5 \times 5 \times 4 + 1) \times 10 = 1010$
MaxPool_2	$4 \times 4 \times 10$	0
Linear	160×1	$160 + 1 = 161$

Question 5F Describe your CNN experiment setup. [5 marks]

1. Split dataset into train and test sets in a 9:1 ratio.
2. Perform image augmentation separately on since 1000 images is too little.
3. Upsample the minority class.
4. Train the classifier on the training set.
5. Score the classifier using F1-score.

Other marks may be awarded or deducted based on reasoning (or lack thereof).

Question 6 Did you see the animal? [3 marks]