1. [10 points] [L5, CO 1] Suppose $\mathbf{X}$ represents the data matrix (samples along columns) containing information about 100 individuals

with features $\begin{cases} \text{age,} \\ \text{annual income,} \\ \text{credit limit,} \\ \text{gender (female or male),} \\ \text{education level (graduate, high school, post graduate),} \end{cases}$ and categories $\begin{cases} \text{does not default payment,} \\ \text{defaults payment.} \end{cases}$

(a) Suppose we want to apply softmax classifier to the dataset. What will be the shape of the weights matrix $\mathbf{W}$ assuming that the bias trick has been done?

(b) What will be the shape of the raw scores matrix comprising the raw scores of all 100 samples?

(c) In plain English and using the data as context, explain what each of the following represents assuming indexing starts from 1:

$$w_{:,2},\ w_{1,:},\ w_{1,8},\ w_{2,5}.$$

8 features:

(age, annual_income, credit limit,

gender_female, gender_male,

edu_grad, edu_high_school,

edu_pg)

8 + 1 bias

a) $W_{2\times 9}$

b) $Z = W_{2\times 9}\, X_{9\times 100}$

$\therefore Z_{2\times 100}$

c) $W_{:,2}$ – weights applied to all output category raw scores and annual-income feature

$W_{1,:}$ – weights applied to all features and 'does not default payment' class

$W_{1,8}$ – weights applied to all features and 'does not default payment' class and 'edu_19' feature

Illy $w_{2,5}$.

2. [10 points] [L5, CO3] Consider the following initial weights matrix (assuming that the bias trick has been done) for dense layer l in a deep neural network:

$$\mathbf{W} = \begin{bmatrix} 0.01 & -0.01 & 0.08 & 0.1 \\ 0.01 & -0.01 & 0.08 & 0.1 \end{bmatrix}.$$

(a) How many nodes are there in layer $l-1$ and layer l?

(b) Is there any issue with the initial values of the weights given here? Justify your answer in 1-2 lines.

(c) Calculate the $L2$-regularization loss for dense layer l.

a) layer $l-1$: 4

layer l : 2

b) "issue of symmetry" since both neurons have same initial values, they will learn same thing

c) $2 \times (0.01)^2 + 2 \times (-0.01)^2 + 2 \times (0.08)^2$

(regularization not applied on bias & hence last column is excluded)

3. [10 points] [L5, CO3] Suppose we want to implement a dropout layer after dense layer l of a deep neural network with a dropout probability of 0.2. Consider the following dropout-matrix:

$$D = \begin{bmatrix} 0.49 & 0.47 & 0.7 & 0.99 \\ 0.86 & 0.49 & 0.76 & 0.96 \\ 0.13 & 0.98 & 0.87 & 0.54 \\ 0.48 & 0.96 & 0.76 & 0.32 \\ 0.62 & 0.15 & 0.23 & 0.58 \end{bmatrix}$$

(a) What is the number of neurons in dense layer l?

(b) What is the batch size?

(c) Each batch sample contributes to the learning of specific neurons of dense layer l. Identify those neurons for each batch sample and fill in the table below (counting starts from 0):

Batch sample	Neurons
0	?, ?, ...
⋮	⋮

(d) How does the activations vector for dense layer l for the 0th sample denoted as $a^{[l](0)}$ gets forward propagated through this dropout layer? Your answer should be a vector whose elements involve the elements of the vector $a^{[l](0)}$.

(e) Compare and contrast dropout vs. loss-based regularization in not more than 2-3 lines.

(f) Is dropout applied to test data? In one line, justify your answer.

(no. of features + bias)

a) $Z_{m \times n} = W_{m \times p} \times X_{p \times n}$ no. of samples

no. of neurons

will have same shape as $D_{m \times n} = D_{5 \times 4}$

$\therefore$ no. of neurons in layer $l = 5$

b) Batch size is 4

c) all activated scores less than (1 − drop-out probability) will be passed to next layer c)

Batch sample	Neurons
0	0, 2, 3, 4
1	0, 1, 3, 4
2	0, 1, 3, 4
3	2, 3, 4

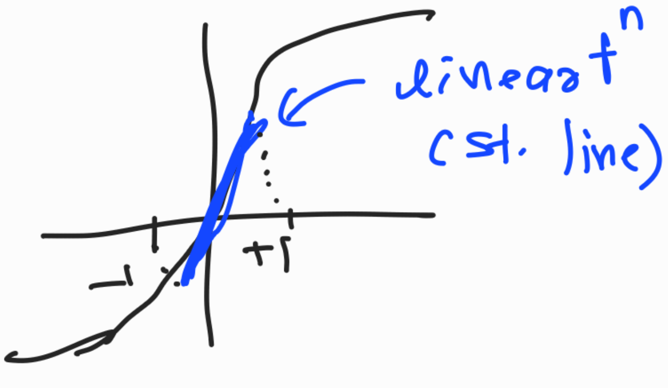
d)

$$a^{[0][0]} = \begin{bmatrix} a_0/0.8 \\ a_1/0.8 \\ a_2/0.8 \\ a_3/0.8 \\ a_4/0.8 \end{bmatrix}$$

e)

→ Dropout reg^n is computationally less intensive whereas l2/l1 reg^n is $comp^n$ intensive.
(L1) L2 + loss based)

→ L1/L2 reg^n tends to shrink weight values to zero, which make raw score values to be near zero making it look like linear f^n



dropouts avoid this

+ loss based reg^n applied to loss f^n, dropouts ← just randomly dropping connections b/w neurons

f) dropouts are not applied on test data, only on training data as regularization which is done to avoid overfitting is required only during training

4. [10 points] [L5, CO4] Consider the following sample matrix:

$$\mathbf{X} = \begin{bmatrix} -6 & -5 & 6 & 4 \\ 7 & -10 & -2 & 5 \\ 1 & 1 & 7 & -8 \\ -2 & 2 & 1 & 1 \end{bmatrix}.$$

(a) Suppose you are told that this sample represents a grayscale image. Justify why you see negative numbers in the sample matrix.

(b) Convolve this sample with the kernel $K = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ using no zero padding and unit stride.

a) mean centering

↓

(pixel value − mean pixel value)

↓

was done on all values in initial matrix

b) idond neene madko

c) Edge detection

interpret the effect of convolving the image with the given kernel.

5. [10 points] [L3, CO4] Consider the convolutional neural network defined by the layers in the left column in the table below. Fill in the shape of the output volume and the number of parameters corresponding to each layer using the notations below:

- CONVx-N denotes a convolutional layer with N filters with kernel height and width both equal to x. Padding is 2, and stride is 1.
- POOL-N denotes an $N \times N$ max-pooling layer with stride of N and no zero padding.
- FLATTEN flattens its inputs.
- FC-N denotes a fully-connected layer with N neurons.

Layer	Output Volume Shape	Number of Parameters
Input	$32 \times 32 \times 3$	0
CONV3-16	?	?
Leaky ReLU	?	?
POOL-2	?	?
FLATTEN	?	?
FC-10	?	?

$$\frac{32 - 3 + 2(2)}{1} + 1$$

Layer	o/p vol shape	no. of parameters
Input	32x32x3	0
CONV3-16	34 x 34 x 16	16(3x3x3+1)
Leaky ReLU	34x34x16	0
POOL-2	17 x 17 x 16	0
FLATTEN	4624	0
FC-10	10 x 1	4625