NAME ________________________________ COEN 240 FINAL EXAM – FALL 2021

- This exam is open book, open notes, open slides. 6 problems. Total 20 points (normalized to 100%).

PROBLEM 1 (2 points)

An input image is convolved with a filter with stride = 1 in a convolutional layer.

Input Image			Filter	
1	2	3	-1	2
4	5	6	0	1
7	8	9		

Fill in the table below representing the output of this layer.

1(-1) + 2(2) +4(0) + 5(1) = **8**	2(-1) + 3(2) + 5(0) + 6(1) = **10**
4(-1) + 5(2) + 7(0) + 8(1) = **14**	5(-1) + 6(2) + 8(0) + 9(1) = **16**

PROBLEM 2 (10 points) - Choose one answer per question.

Consider a dataset with 5 dimensions, 3 classes and 25 observations for questions A and B.

A). How many principal components are required to fully represent the dataset?

[✓ (a) 5]
(b) 3
(c) 25
(d) None of the above

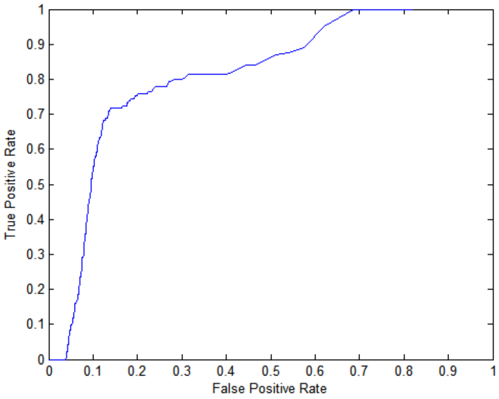
B). The dataset is used to train a Bayesian Linear Discriminant Analysis classifier with equal prior probabilities. What is the value of $P(\omega_i)$?

(a) 1/5
[✓ (b) 1/3]
(c) 1/25
(d) None of the above

C). Full SVD is performed on matrix A = $\begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}$. What are the dimensions of the matrices that contain the left singular vectors, the singular values and the right singular vectors, respectively?

(a) 4x4, 4x3, 3x3
(b) 3x3, 4x4, 4x4
[✓ (c) 3x3, 3x4, 4x4]
(d) None of the above

D). What is the most likely AUC for the following ROC curve of a binary classifier?



(a) 0.95
(b) 0.65
(c) 0.35
(d) 0.15

E). Which one of the following statement accurately describes *Early Stopping*?

(a) Stop training when the training error reaches minimum
(b) Stop training when the validation error reaches minimum
(c) Stop training when the training error drops below the validation error
(d) Stop training when the validation error drops below the training error
(e) None of the above

F). Your client is using a large dataset to train a binary classifier but they have no idea what the underlying probability distribution is. What model would you suggest they use?

(a) Linear Regression
(b) Logistic Regression
(c) Bayes Decision Theoretic
(d) Any of the above since the training set is large
(e) None of the above

G). Having more features always results in higher classification accuracy.

(a) True
(b) False
(c) It depends on the size of the dataset
(d) None of the above

H). What causes underfitting?

(a) Low bias, low variance
(b) Low bias, high variance
(c) High bias, low variance
(d) High bias, high variance
(e) None of the above

I). Deep neural networks can have a problem where the gradients drop dramatically toward zero during back-propagation. What is this problem called?

(a) Unknown gradients
(b) Exploding gradients
(c) Vanishing gradients
(d) None of the above

J). Which of the following is true about GANs?

(a) During training the weights of the discriminator and the generator are optimized at the same time.
(b) During training the weights of the discriminator and the generator are "frozen" at the same time.
(c) The generator takes poor fake images as input and produces images that look real enough to fool the discriminator.
(d) GAN-generated images (i.e., fake images) come from the output of the discriminator.
(e) None of the above

PROBLEM 3 (2 points)

Mercer's Theorem is a powerful theorem. It states, among others, that if a mapping function exists, there is an equivalent kernel representation of the inner product operation of this function. For example, for the mapping function $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$, where $\phi(\boldsymbol{i}) = \left(i_1^2, \sqrt{2}\, i_1 i_2, i_2^2\right)$, the associated kernel is $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}.\boldsymbol{z})^2$ where $\boldsymbol{x}$ and $\boldsymbol{z}$ are two-dimensional vectors. However, since not every function satisfies the Mercer's conditions, a kernel may not exist for that function. Mathematically show that if the function $\phi(\boldsymbol{i})$ is altered slightly, say $\phi(\boldsymbol{i}) = (i_1^2,\ i_1 i_2, i_2^2)$, then the kernel $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}.\boldsymbol{z})^2$ is no longer associated with that function.

$$\langle \phi(x), \phi(z) \rangle = (x_1^2, x_1 x_2, x_2^2).(z_1^2, z_1 z_2, z_2^2) = x_1^2 z_1^2 + x_1 x_2 z_1 z_2 + x_2^2 z_2^2$$

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}.\mathbf{z})^2 = [(x_1, x_2).(z_1, z_2)]^2 = (x_1 z_1 + x_2 z_2)^2 = x_1^2 z_1^2 + 2 x_1 x_2 z_1 z_2 + x_2^2 z_2^2$$

They are not identical → the kernel $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}.\boldsymbol{z})^2$ is not the kernel for the mapping function ϕ.

PROBLEM 4 (2 points)

You want to use gradient descent to find the optimal solution to a cost function, but the cost function is concave. Is it okay? If not, what do you do?

Multiply the cost function by -1 to make it convex

If you say "gradient ascent", that's ok too.

PROBLEM 5 (2 points)

You are designing a fully connected neural network to classify a dataset that contains 10,000 color images of uppercase and lowercase letters of the alphabet (A-Z, a-z), where each image is 30x30x3 pixels (the third dimension is the color dimension). What are the sizes of the input layer and the output layer, i.e., the number of "neurons" in the input layer and the number of neurons in the output layer?

Input layer = 2700

Output layer = 52

PROBLEM 6 (2 points)

Draw an overcomplete stacked autoencoder with one hidden layer. The input and output layers have 1000 units.

The hidden layer must have at least 1000 units