- Open book/notes/slides. 1 hour. 25 points (normalized to 100%). Show your work as appropriate.
- ***You must complete this exam entirely on your own with no collaboration or consultation with any other person.***

PROBLEM 1 (2 points)

A fully connected neural network is used to classify a dataset that contains 10,000 color images of uppercase and lowercase letters (A-Z, a-z), where each image is 30x30x3 pixels (the third dimension is the color dimension). What are the dimensions of the input layer and the output layer?

Input layer = 2700

Output layer = 52

PROBLEM 2 (4 points)

A labeled dataset has 5000 observations, 50 features and 20 classes. The feature matrix X is first processed by PCA. Three principal components are used to reconstruct the dataset $\hat{X}$. Combined with the class label, it is then used to train a Softmax Regression classifier.

What is the total number of unused principal components? 47

What is the dimension of feature matrix X? 5000x50

What is the dimension of reconstructed matrix $\hat{X}$ 5000x50

What is the dimension of the matrix that contains *all* of the labels for training the Softmax Regression?

Possible answers depending on how you interpret the question:

- 5000x20 (20 one hot encoded labels)
- 5000x70 (50 features + 20 one-hot encoded labels)

PROBLEM 3 (10 points) – Choose only one answer per question.

A). Which of the following is NOT true about Mercer's Theorem?

(a) It can be used to find the mapping function ϕ
(b) Computing the kernel is sufficient
(c) It is not necessary to find the dimensionality of ϕ
(d) Both (b) and (c)
(e) None of the above

B). Which of the following is NOT true for SVM?

(a) Training examples that are support vectors can be discarded
(b) Training examples that are support vectors cannot be discarded
(c) The decision boundary is placed in the middle of the "street"
(d) The decision boundary is placed at the optimal location
(e) None of the above

C). Your friend is training a classifier using a dataset whose underlying probability distribution is unknown. What should they use?

(a) Linear Regression
(b) Logistic Regression
(c) Bayes Decision Theoretic
(d) K-means
(e) None of the above

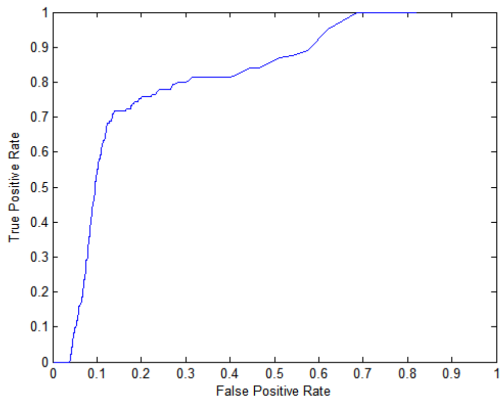
D). How many output neurons are needed to classify a dataset with 5 dimensions, 3 classes and 25 observations?

(a) 5
(b) 3
(c) 25
(d) 5x3
(e) None of the above

E). How many eigenvalues and eigenvectors does matrix $X = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}$ have?

(a) 3 eigenvalues, 4 eigenvectors
(b) 4 eigenvalues, 3 eigenvectors
(c) 3 eigenvalues, 3 eigenvectors
(d) 4 eigenvalues, 4 eigenvectors
(e) None of the above

F). What is the most likely AUC for the following ROC curve?



(a) 0.95
(b) 0.70
(c) 0.30
(d) 0.15
(e) 0.05

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

(a) True
(b) False
(c) It depends on the size of the dataset
(d) It depends on whether the model is overfitting or underfitting
(e) None of the above

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

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

I). 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 fake images as input and produces real images.
(d) GAN-generated images come from the output of the discriminator.
(e) None of the above

J). Which of the following statement is true about PCA?

(a) Most of the information is associated with large eigenvalues
(b) Most of the information is associated with large eigenvectors
(c) Most of the information is associated with small eigenvalues
(d) Both (a) and (b)
(e) None of the above

PROBLEM 4 (3 points)

SVD is a powerful decomposition technique. Given matrix A, it decomposes it into the product of three matrices: $A = U.S.V^T$ where U is the left singular vectors of A (the eigenvectors of AA^T), S is the singular values of A (the square root of the eigenvalues of AA^T) and V is the right singular vectors of A (the eigenvectors of A^TA). If $A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$, determine its singular values. Show your work.

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} . \begin{bmatrix} 1 & 3 \\ 2 & 2 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 10 \\ 10 & 14 \end{bmatrix} \rightarrow \begin{vmatrix} 14-\lambda & 10 \\ 10 & 14-\lambda \end{vmatrix} = 0$$

$$\rightarrow (14-\lambda)(14-\lambda) - 100 = 0 \rightarrow \lambda^2 - 28\lambda + 98 = 0 \rightarrow (\lambda - 24)(\lambda - 4) = 0$$

$$\rightarrow \lambda_1 = 24, \lambda_2 = 4$$

Singular values: $\sqrt{24} = 4.8990$ and $\sqrt{4} = 2$

PROBLEM 5 (3 points)

An SVM classifier with a polynomial kernel $(\mathbf{x}^T\mathbf{z} + 1)^2$ is trained using a two-dimensional training set.

i	$\mathbf{x}^T$	y	Langrage Multiplier	Support Vector
1	1, 0	-1	0.1	Yes
2	2, 4	-1	0.2	Yes
3	1, 1	-1	0.3	No
4	3, 1	+1	0.4	Yes
5	1, 3	+1	0.5	No
6	4, 4	+1	0.6	No

Determine the class label for the instance (-1,1). Assume the bias is 0. Show your work.

Use only the support vectors: M = 3

$$h = \sum_{i=1}^{M} \alpha_i y_i K(\mathbf{x}_i^T \mathbf{z}) = \sum_{i=1}^{M} \alpha_i y_i (\mathbf{x}_i^T \mathbf{z} + 1)^2$$

For $\mathbf{z}^T = -1{,}1$: $h = -(0.1)(-1+1)^2 - (0.2)(2+1)^2 + (0.4)(-2+1)^2 = -1.4 < 0$

$\rightarrow$ **negative class**

PROBLEM 6 (3 points)

Mercer's Theorem states that if a mapping function exists, there is a kernel representation of the inner product of this function. For instance, for the mapping function ϕ: $\mathbb{R}^2 \rightarrow \mathbb{R}^3$, 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. 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.