PROBLEM 1 (2 points)

A fully connected neural network is used to classify a dataset that contains 15,000 color images of hand-written digits and lowercase letters (0-9, a-z), where each image has 20x20x3 pixels. What are the sizes of the input layer and the output layer?

Input layer = 20 x 20 x 3 = 1200

Output layer = 10 + 26 = 36

PROBLEM 2 (4 points)

A labeled dataset has 9000 observations, 90 features and 20 classes. The feature matrix X is first processed by PCA. Ten 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? 80

What is the dimension of feature matrix X? 9000x90

What is the dimension of reconstructed matrix $\hat{X}$ 9000x90

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:

- 9000x20 (20 one hot encoded labels)
- 9000x110 (90 features + 20 one-hot encoded labels)

PROBLEM 3 (8 points) – Choose 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 statement is true about PCA?

(a) Most of the information is associated with its large eigenvalues
(b) Most of the information is associated with its large eigenvectors
(c) Both (a) and (b)
(d) It depends on the dataset
(e) None of the above

C). Deep neural networks can have a problem where the gradients increase 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

D). 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

E). How many eigenvalues and eigenvectors does matrix $X = \begin{bmatrix} 5 & 2 & 6 \\ 8 & 1 & 6 \\ 3 & 8 & 9 \\ 0 & 4 & 7 \end{bmatrix}$ 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). 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

G). Which of the following is NOT true about clustering?

(a) K-Means always converges
(b) K-Means may or may not converge
(c) Cluster assignment may change from run to run
(d) The elbow rule can be used to determine the number of clusters
(e) None of the above

H). 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

<u>PROBLEM 4</u> (2 points)

List all the training examples generated from the sentence ***Mary had two little lambs*** by Word2Vec if the center word is ***two*** and the window size is 3 words.

{two, Mary}

{two, had}

{two, little}

{two, lambs}

PROBLEM 5 (3 points)

SVD is a powerful decomposition technique. It decomposes matrix A 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 $A = \begin{bmatrix} 3 & 0 & 4 \\ 3 & 4 & 0 \end{bmatrix}$, determine its singular values.

$$\begin{bmatrix} 3 & 0 & 4 \\ 3 & 4 & 0 \end{bmatrix} . \begin{bmatrix} 3 & 3 \\ 0 & 4 \\ 4 & 0 \end{bmatrix} = \begin{bmatrix} 25 & 9 \\ 9 & 25 \end{bmatrix} \rightarrow \begin{vmatrix} 25-\lambda & 9 \\ 9 & 25-\lambda \end{vmatrix} = 0$$

$$\rightarrow (25-\lambda)(25-\lambda) - 81 = 0 \rightarrow (25-\lambda) = \pm 9$$

$$\rightarrow \lambda_1 = 34, \lambda_2 = 16$$

Singular values: $\sqrt{34} = 5.8310$ and $\sqrt{16} = 4$

PROBLEM 6 (3 points)

Suppose the principal components and the eigenvectors in PCA are given by matrix P and matrix V, respectively. Assume the mean of the dataset is 0.

$$P = \begin{bmatrix} 3 & 2 & 3 \\ 5 & 3 & 4 \\ 2 & 1 & 6 \end{bmatrix} \quad V = \begin{bmatrix} 6 & 0 & 3 \\ 4 & 1 & 2 \\ 1 & 4 & 0 \end{bmatrix}$$

What is the reconstructed dataset $\hat{X}$ if only the first principal component is used?

$$\hat{X} = P.V^T$$

Where P is 3x1 and V is 3x1.

$$\hat{X} = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix} \times \begin{bmatrix} 6 & 4 & 1 \end{bmatrix} = \begin{bmatrix} 18 & 12 & 3 \\ 30 & 20 & 5 \\ 12 & 8 & 2 \end{bmatrix}$$

PROBLEM 7 (3 points)

A two-dimensional dataset is used to train an SVM classifier with a polynomial kernel $(\mathbf{x}^T\mathbf{z} + 1)^2$.

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

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

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 $\mathbf{z}^T = -1{,}1$: $h = -(0.3)(0+1)^2 + (0.5)(2+1)^2 + (0.6)(0+1)^2 = 4.8 > 0$

$\rightarrow$ **positive class**