TEST DATE: 0900-1200 HRS, 04-MAY-2025

DA623: Computing with Signals

End-Semester Examination | Jan-May 2025, IIT Guwahati

Total Questions: Seven (8) — Total Points: Forty five (55)

- **Thought of the day** – "Be calm and answer the questions." — **Time: 180 mins**
- Write with clarity, stating any assumptions you make and putting down equations wherever there is a slight chance for it.

Question 1 [1 × 10 = 10 points]

State True or False, and the reason for it.

(a) The number of bits used to represent each pixel in an image directly determines its spatial resolution.
(b) Pooling layers in CNNs reduce the number of parameters in the model.
(c) In CNNs, the same filter is applied to different regions of the input image, allowing parameter sharing.
(d) Words with similar spelling will always have similar Word2Vec embeddings.
(e) Flattening is used in CNNs to convert 2D feature maps into 1D before feeding them into convolutional layers.
(f) All neurons in the body are located in the brain.
(g) The RGB color model used in digital displays is inspired by how rod cells in the human eye work.
(h) An image with sharp edges will have a Fourier spectrum with high-magnitude components at high frequencies.
(i) A band-pass filter in the frequency domain can isolate textures in specific orientation and scale ranges.
(j) A 3×3 averaging filter is a high-pass filter.

Question 2 [4+1+2+3=10 points]

You are given a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ containing n samples, where each sample $\mathbf{x}_i \in \mathbb{R}^N$.

(a) Describe the steps involved in the K-means clustering algorithm when used to segment a color image into foreground and background by searching for $K = 2$ clusters.
(b) Discuss whether the solution obtained by K-means is guaranteed to be a global optimum. Justify your answer.
(c) Propose a modification to the K-means algorithm that ensures the cluster centroids are selected from the existing data points, rather than being averages of the points in each cluster. This modification would be particularly useful in applications such as clustering face images, where it is desirable for the cluster representative to be an actual face image rather than an average face. Justify your choice of modification.
(d) Design a methodology that uses clustering to obtain a **sparse representation** of the dataset $\mathcal{D}$, and state the associated mathematical formulation. Here, a "sparse representation" means that each data point $\mathbf{x} \in \mathcal{D}$ should be approximated as a linear combination of only a few selected cluster centroids, with most of the coefficients being zero (or close to zero). The goal is to approximate $\mathbf{x} \approx \Psi\alpha$, where $\Psi \in \mathbb{R}^{N \times K}$ and $\alpha \in \mathbb{R}^{K \times 1}$ is a sparse coefficient vector with very few non-zero entries.

Question 3 [3+2+3+2=10 points]

You are provided with a dataset of samples $\mathcal{D} = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\}$ where each $x_i \in \mathbb{R}^N$ is a feature vector representing an image, y_i is a categorical label (e.g., 'pen', 'cup'), and z_i is a real-valued quantity (e.g., weight). You are given a new test sample x_{test}. Your task is to design a simple prediction method that uses the existing dataset to:

(a) Propose a method to predict the categorical label y_{test} for x_{test} using only the dataset and a suitable notion of similarity between vectors. Your method should not involve any additional training.
(b) Extend your method to estimate the real-valued quantity z_{test} for x_{test}, again using only the existing dataset.
(c) Compare it with models that involve parameter learning (e.g., neural networks) in terms of training cost, flexibility, and prediction time.
(d) In large-scale or real-time applications, comparing x_{test} with every x_i in the dataset can be computationally expensive. Suggest two methods to make the approach more efficient while maintaining reasonable prediction performance.

Question 4 [2+2=4 points]

(a) Design a spam email detector using the Bag of Words (BoW) approach. Describe how you will construct BoW-based feature vectors to represent each spam mail in this task.
(b) During testing, you encounter a words (e.g., "fr33", ofter) that were not present in your training vocabulary, and as a result, the model cannot represent these words and will ignore them. Suggest a modification to the BoW representation approach that helps overcome this limitation, particularly for handling unseen words.

Course Instructor: Neeraj Kumar Sharma (e-mail: neerajs@iitg.ac.in) | Mehta Family School of Data Science & AI

TEST DATE: [illegible]

QUESTION 5 [2+2+4 POINTS]

The Skip-gram model is part of the Word2Vec framework used for learning word embeddings. In this model, given a target word w_t, the goal is to predict its surrounding context words within a window of size c.

(a) Write the mathematical formulation for the loss function used in the Skip-gram model, focusing on the prediction of context words given a target word. Make sure you define all relevant symbols (e.g., target word w_t, context words w_{t+j}, vocabulary V).

(b) In the general context of embeddings, state two disadvantages of embeddings over one-hot encoding.

QUESTION 6 [1+1+2+4 POINTS]

You are given a 2-D dataset with two features, denoted as x_1 and x_2. The dataset consists of several data points drawn from a Gaussian distribution. You are to analyze how different choices of the covariance matrix affect the shape and orientation of the scatter plot of the data.

(a) What do the diagonal and off-diagonal elements of the covariance matrix represent?

(b) Draw scatter plots of data sampled from a bivariate Gaussian distribution for each of the following covariance matrices:

(i) Identity matrix: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

(ii) Diagonal matrix with unequal variances: $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$

(iii) Non-diagonal covariance matrix: $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

(c) State how the eigenvalues and eigenvectors of the covariance matrix relate to the shape and orientation of the scatter plot.

QUESTION 7 [2+2+2=6 POINTS]

Autoencoder is a neural network trained to reconstruct its input. Assume you have trained a deep neural network based auto-encoder on a face image dataset.

(a) How can you use this neural network to obtain compressed representations of images. State any assumptions, mathematical equations and neural network architecture, as required.

(b) How is the above approach similar and different from using PCA?

(c) State how a CNN-based autoencoder might be preferable over a DNN-based one?

QUESTION 8 [5 POINTS]

You are provided with a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where each data sample $\mathbf{x}_i \in \mathbb{R}^d$ and no additional labels are available.

(a) Design a methodology that combines a deep neural network with K-means clustering to learn meaningful representations of the data in an unsupervised manner.

(b) Clearly describe the architecture of the model you propose, and draw a schematic diagram to illustrate the components and data flow.

(c) Define all mathematical notations used, and provide the key equations governing the learning process.