5. (a) Suppose 10,000 patients get tested for flu; out of them, 9,000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8,820. Construct a confusion matrix for the data and compute the precision and recall for the data. **8**

(b) What is Matrix Factorization and Matrix Completion ? Explain. **7**

6. (a) Draw and explain architecture of neural network. **7**

(b) What is Bayesian learning and Inference ? Explain. **8**

7. Write short notes on the following : **15**

(a) Scalable Machine Learning

(b) Feature Representation Learning

(c) Naïve Baye's Classification.

016601

May 2024
B.Tech. (CE(DS)) (Sixth Semester)
Machine Learning Principles (PCC-DS-601)

Time : 3 Hours]   [*Maximum Marks :* 75

Note : It is compulsory to answer all the questions (1.5 marks each) of Part A in short. Answer any *four* questions from Part B in detail. Different sub-parts of a question are to be attempted adjacent to each other.

Part A

1. (a) What are various stages involved in designing a learning system ? **1.5**

(b) Describe LMS weight update rule. **1.5**

(c) What are outliers ? List any two methods to deal with outliers. **1.5**

(d) What is the pruning in Decision tree ? **1.5**

(e) What do you mean by Concept Learning ? **1.5**

(f) Define the following :   1.5
  (i)   Prior Probability
  (ii) Conditional Probability
  (iii) Posterior Probability.

(g) Differentiate bagging, boosting and voting.   1.5

(h) Differentiate between Gradient Descent and Stochastic Gradient Descent.   1.5

(i) What is Artificial Neural Network ?   1.5

(j) How entropy and information gain are related ?   1.5

Part B

2. (a) The sales of a company (in million dollars) for each year are shown in the table below.   8

x (year)	y (sales)
2005	12
2006	19
2007	29
2008	37
2009	45

(i) Find the least square regression line $y = ax + b$.

(ii) Use the least squares regression line as a model to estimate the sales of the company in 2012.

(b) Use K Means clustering to cluster the following data into two groups. Assume cluster centroid are $m_1 = 2$ and $m_2 = 4$. The distance function used is Euclidean distance.

$\{2, 4, 10, 12, 3, 20, 30, 11, 25 \}$ 7

3. State the mathematical formulation of the soft margin SVM as a convex optimization problem. Describe, how is dual form useful in applying kernel trick ? Given the hyperplane defined by the line $y = x_1 - 2x_2$. Are these points correctly predicted :

(a) $y = 1$, $x = (1, 0)$?

(b) $y = 1$, $x = (1, 1)$?

4. Explain PCA with calculation of Eigen vectors and derivation of covariance matrix.   15