Review Final

1 Part A

Question 1: Consider a set of five training examples given as $((x_i, y_i), c_i)$ values, where x_i and y_i are the two attribute values (positive integers) and c_i is the binary class label:

$$\{((1,1),-1),((1,7),+1),((3,3),+1),((5,4),-1),((2,5),-1)\}.$$

Classify a test example at coordinates $(3,6)$ using a k-NN classifier with $k = 3$. Your answer should be either $+1$ or -1.

(A) $+1$

(B) -1

Question 2: A company is studying the impact of training on employee performance. Employees are categorized based on training levels (Basic or Advanced) and performance (High or Low). The probabilities from past data are summarized in the table below:

Training Level	**High Performance**	**Low Performance**
Basic	0.3	0.3
Advanced	0.3	0.1

Calculate the Information Gain $IG(\text{Performance}, \text{Training})$ for predicting performance based on training level.

(A) ≈ 0.05

(B) ≈ 0.06

(C) ≈ 0.07

(D) ≈ 0.04

Question 4: Which of the following functions is convex?

(A) $f(x) = x^3 - 3x$

(B) $f(x) = x^2 + 2x + 1$

(C) $f(x) = \ln(x) \quad \text{for } x > 0$

(D) $f(x) = -x^2 + 4x$

Question 5: Consider one layer of weights (edges) in a convolutional neural network (CNN) for grayscale images, connecting one layer of units to the next layer of units. Which type of layer has the fewest parameters to be learned during training?

(A) A convolutional layer with 10 3×3 filters

(B) A max-pooling layer that reduces a 10×10 image to 5×5

(C) A convolutional layer with 8 5×5 filters

(D) A fully-connected layer from 20 hidden units to 4 output units

Question 6: You are tasked with applying a 2-D convolution to an image using a given filter (kernel).

- **Input Image** I (4x4 matrix):

$$A = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 4 & 0 & 1 & 2 \\ 1 & 3 & 0 & 1 \\ 2 & 1 & 2 & 0 \end{bmatrix}$$

- **Filter** K:

$$K = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

Use the filter K to compute the output value at position $O(2, 2)$ of the resulting output matrix O. Assume a stride of 1 and no padding.

(A) -1

(B) 1

(C) 0

(D) 2

Question 7 : Suppose we have a $1000 \times 1000 \times 3$ dimension input image (width $\times$ height $\times$ channel). We apply a convolutional layer with 50 5×5 kernels. What is the dimension of the resulting tensor (width $\times$ height $\times$ channel) if we have stride = 1 and no padding?

(A) $995 \times 995 \times 3$

(B) $996 \times 996 \times 3$

(C) $995 \times 995 \times 50$

(D) $996 \times 996 \times 50$

Question 8: Suppose we are performing linear regression using a non-linear basis expansion. Which of the following statements is true about the learned predictor?

(a) It is a linear function of the inputs and a linear function of the weights.

(b) It is a linear function of the inputs and a non-linear function of the weights.

(c) It is a non-linear function of the inputs and a linear function of the weights.

(d) It is a non-linear function of the inputs and a non-linear function of the weights.

Question 8: In Gaussian mixture models (GMMs), which of the following statements is false?

(a) GMMs assume that the data points within each component follow a Gaussian distribution.

(b) GMMs can be used for clustering.

(c) The number of components in a GMM must be equal to the number of features in the dataset.

Question 8: In Gaussian mixture models (GMMs), which of the following statements is false?

(a) GMMs assume that the data points within each component follow a Gaussian distribution.

(b) GMMs can be used for clustering.

(c) The number of components in a GMM must be equal to the number of features in the dataset.

Question 9: What is the key reason why backpropagation is so important?

(a) Backpropogation allows us to compute the gradient of any diferentiable function.

(b) Backpropogation is the only algorithm that enables us to upate the weights of a Neural Network.

(c) Backpropogation is an efcient dynamic program that enables us to compute the gradient of a function at the same time-complexity it takes to compute the function.

(d) Backpropagation introduced Chain Rule into the world of mathematics, enabling signifcant advances in the feld.

Question 10 : [True/False] Suppose you set up and train a neural network on a classifcation task and converge to a fnal loss value. Keeping everything in the training process the exact same (e.g. learning rate, optimizer, epochs). It is possible to reach a lower loss value by ONLY changing the network initialization.

(a) True

(b) False

Question 11 : [True/False] The kernel density estimator is equivalent to performing kernel regression with the value Yi = 1/n at each point Xi in the original data set.

(a) True

(b) False

Question 12 : [True/False] The correspondence between logistic regression and Gaussian Naıve Bayes (with identity class covariances) means that there is a one-to-one correspondence between the parameters of the two classifiers.

(a) True

(b) False

Question 13 : [True/False] As the number of data points grows to infinity, the MAP estimate approaches the MLE estimate for all possible priors. In other words, given enough data, the choice of prior is irrelevant.

(a) True

(b) False

Question 14 : [True/False] Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

(a) True

(b) False

Question 15: Are these statements true or false?

- The i-th principal component is taken as the direction that is orthogonal to the $(i-1)$-th principal component and maximizes the remaining variability.
- Different individual principal components (directions) are linearly uncorrelated.

(A) True, True

(B) True, False

(C) False, True

(D) False, False

2 Part B

Question 1: A real estate agency aims to predict house prices based on three key features: the number of bedrooms (x_1), the size in square feet (x_2), and the number of bathrooms (x_3). The target variable (y) represents the price of the house in thousands of dollars. The dataset is given as follows:

x_1	x_2	x_3	y
1	5	3	11
4	9	6	20
7	8	2	14
9	3	1	7

We will initiate the weight vector with $\mathbf{w} = [0, 0, 0]$ and employ a step size $\eta = 0.01$ for the gradient descent algorithm.

Question 2: A telecommunications company wants to predict whether a customer will churn (leave) based on three key features:

- **Monthly Bill** (x_1): The customer's monthly billing amount.
- **Contract Length** (x_2): The length of the customer's contract in months.
- **Customer Service Calls** (x_3): The number of calls the customer made to customer service in the last month.

The target variable y indicates whether the customer has churned (1) or not (0). The dataset is given as follows:

x_1	x_2	x_3	y
7	12	3	1
5	6	1	0
9	4	5	1

- **Initial Weight Vector:** $\mathbf{w} = [-2, 1, 1]$
- **Learning Rate:** $\eta = 0.01$

Update the weight vector $\mathbf{w}$ using stochastic gradient descent based on the second training example only.

Question 3: A company conducted a survey in two cities to understand the preference for a new product. Each individual surveyed was asked to select one of three options: Strongly Prefer (SP), Somewhat Prefer (SWP), or Do Not Prefer (DNP). The observed counts for each preference category in each city are shown below:

City	Preference	Count (k)
1	Strongly Prefer (SP)	150
1	Somewhat Prefer (SWP)	200
1	Do Not Prefer (DNP)	100
2	Strongly Prefer (SP)	180
2	Somewhat Prefer (SWP)	250
2	Do Not Prefer (DNP)	120

Assume that:

- SP (Strongly Prefer): θ^2
- SWP (Somewhat Prefer): $2\theta(1-\theta)$

- DNP (Do Not Prefer): $(1-\theta)^2$

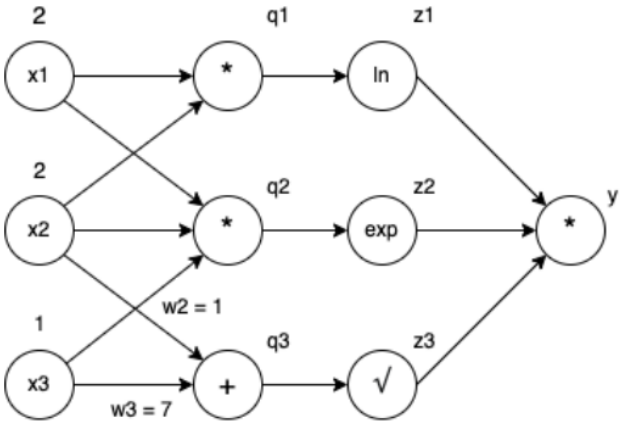
Find the Maximum Likelihood Estimate (MLE) of θ for each city separately.

Question 4: Suppose you are monitoring a machine and want to estimate the probability that it is in a "faulty" state based on past observations. You assume that the probability of the machine being faulty follows a Bernoulli distribution (either faulty or not faulty) with an unknown probability θ, where θ is the probability that the machine is faulty.
You are given the following information:

- **Prior Distribution:** Your prior belief about θ follows a Beta distribution with parameters $\alpha = 2$ and $\beta = 3$. This prior represents your initial assumption about the likelihood of the machine being faulty before any observations.
- **Data (observations):** Over the course of monitoring, you observe that the machine has been faulty in 5 out of 10 trials.

Using this data, find the Maximum A-Posteriori (MAP) estimate of θ.

Question 5: Consider a simple feed-forward neural network with three layers: an input layer, two hidden layers, and an output layer. Using the values provided in the graph, perform backpropagation to compute the following partial derivatives:



- $\frac{\partial y}{\partial z_2}$
- $\frac{\partial y}{\partial q_3}$
- $\frac{\partial y}{\partial x_1}$

Question 6: Let:

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

Write the Singular Value Decomposition (SVD) of A in matrix form.

Question 7: You are using a Gaussian Naive Bayes classifier to predict compatibility with potential boyfriends based on three personality traits: Sense of Humor, Kindness, and Ambition. You have collected the following data on four candidates:

Dataset: Compatibility with Potential Boyfriends

Candidate	Sense of Humor	Kindness	Ambition	Compatibility
Person 1	8	9	7	Compatible
Person 2	6	7	5	Incompatible
Person 3	7	8	6	Compatible
Person 4	5	6	4	Incompatible

For a new candidate with the following ratings:

- Sense of Humor = 7
- Kindness = 8
- Ambition = 6

Use the Gaussian probability density function to calculate the likelihood of this candidate belonging to each class (Compatible and Incompatible).

Question 8: Suppose we have a dataset with two features, represented by the matrix X of shape 3×2:

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

Calculate the coordinates of the projections of each data point in X onto the first principal component.