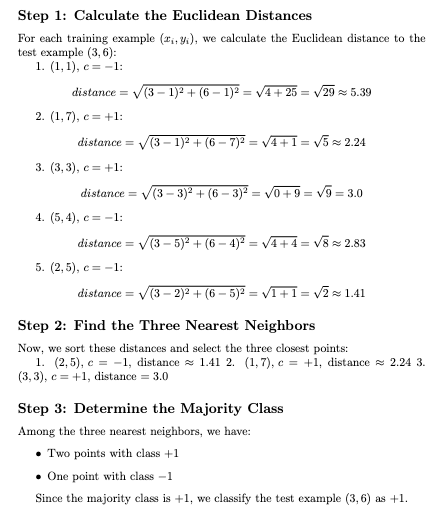
Question 1 : Consider a set of five training examples given as ((xi, yi), ci) values, where xi and yi are the two attribute values (positive integers) and c\_i is the binary class label:

{((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.

1. +1
2. -1



Choose 3 points which is the nearest : (2,5), (1,7), (5,4) => Class -1

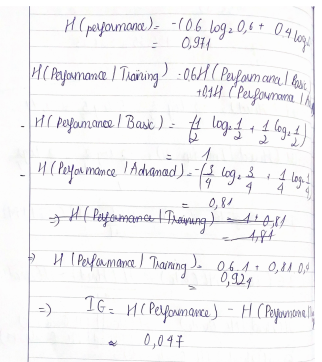
Question 2 :

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(Performance,Training) for predicting performance based on training level. Which of the following steps correctly outlines the calculation?

1. ~0.05
2. ~0.06
3. ~0.07
4. ~0.04



|  |  |  |  |
| --- | --- | --- | --- |
| x1 | x2 | x3 | y |
| 1 | 5 | 3 | 11 |
| 4 | 9 | 6 | 20 |
| 7 | 8 | 2 | 14 |
| 9 | 3 | 1 | 7 |

Question 3 : A real estate agency aims to predict house prices based on three key features: the number of bedrooms (x1), the size in square feet (x2), and the number of bathrooms (x3). The target variable (y) represents the price of the house in thousands of dollars. The dataset is given as follows:

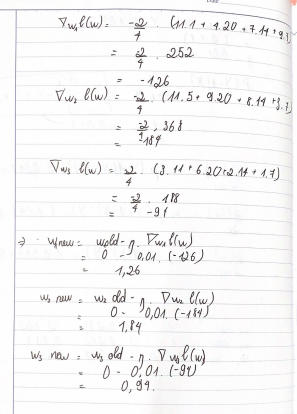
We will initiate the weight vector with w = [0,0,0] and employ a step size η=0.01 for the gradient descent algorithm.

Hint :

Using the update rule:

W\_new = w − η.∇wℓ(w)

Compute the gradient ∇(w) and subsequently update the weight vector to obtain w\_new.

****

Question 4 :

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

- Monthly Bill (x1): The customer's monthly billing amount.

- Contract Length (x2): The length of the customer's contract in months.

| x1​ | x2 | x3 | Y |
| --- | --- | --- | --- |
| 7 | 12 | 3 | 1 |
| 5 | 6 | 1 | 0 |
| 9 | 4 | 5 | 1 |

- Customer Service Calls (x3): 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).

* Initial Weight Vector: w=[-2,1,1]
* Learning Rate: η=0.01

Update the weight vector w using **stochastic gradient descent** based on the **second training example** only.

**Hint :**

1. Calculate Prediction Using Sigmoid Function

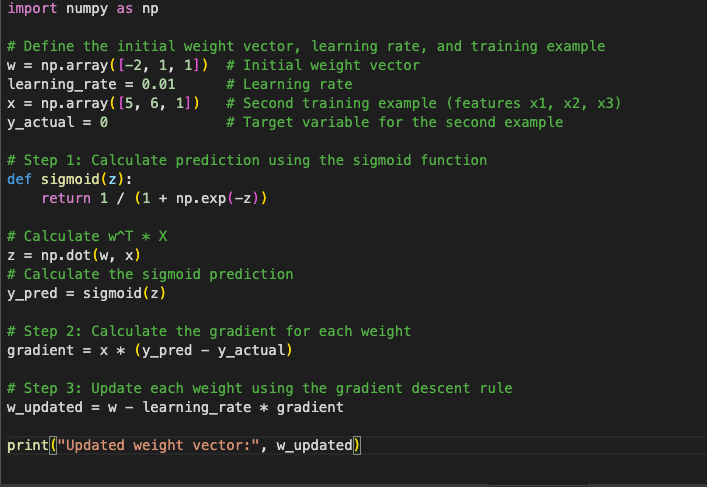
σ(w^T. X) = 1+e^-(w1.x1+ w2.x2+ w3.x3)​

1. Calculate the Gradient

∂ℓ(w)/∂wj =xj(σ(w^T. X) − Y)

1. Update Weights Using Gradient Descent Rule

wj\_new = wj\_old − η. ∂ℓ(w)/∂wj​



Ảnh chụp Màn hình 2024-11-10 lúc 13.02.25

Question 5 : 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*θ*.(1−*θ*)

- DNP (Do Not Prefer): (1−*θ*)^2

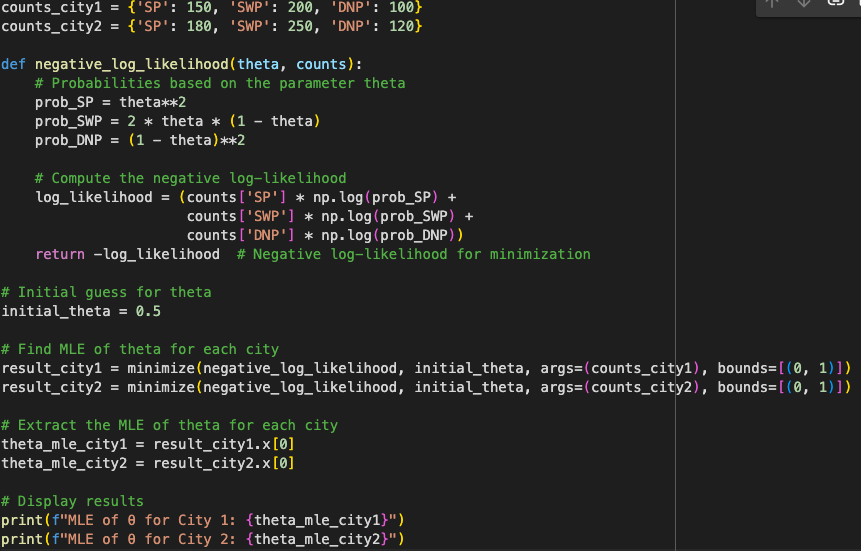
Find the MLE of θ for each city separately. And compare the preference rate θ between the two cities.

\*)Step 1: Set up Likelihood Function

\*)Step 2 : Take the log-likelihood

\*)Step 3 : Differentiate Log-Likelihood

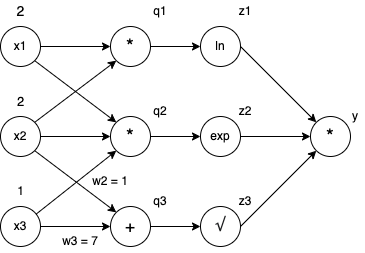
\*)Step 4 : Set derivative to 0 and solve for parameter



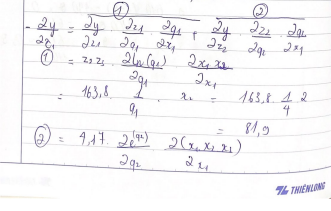
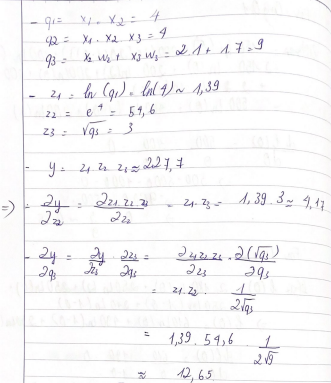
Ảnh chụp Màn hình 2024-11-10 lúc 13.07.47

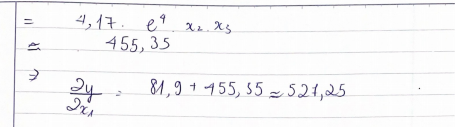
Question 6 :

Consider a simple feed-forward neural network with three layers: an input layer, two hidden layers, and an output layer.



Use the values provided in the graph, perform backpropagation to compute the following partial derivatives: ∂y/∂z2, ∂y/∂q3, ∂y/∂x1.



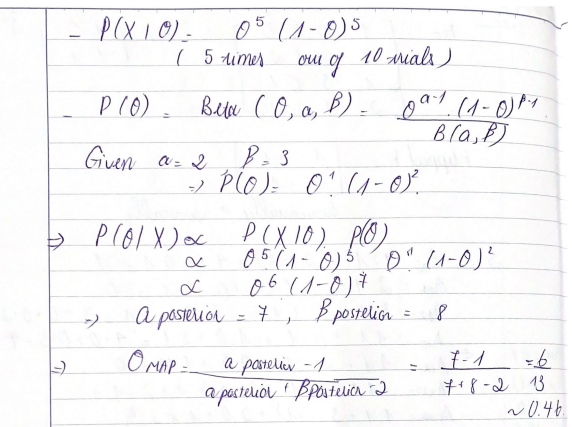


Question 7 : Suppose you are monitoring a machine and want to estimate the probability that it is in a "faulty" state based on past observations. You assume that the probability of the machine being faulty follows a Bernoulli distribution (either faulty or not faulty) with an unknown probability θ, where θ is the probability that the machine is faulty.

You are given the following information:

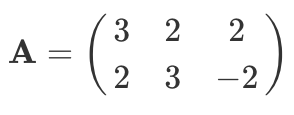
* Prior distribution: Your prior belief about θ follows a Beta distribution with parameters α = 2 and β = 3. This prior represents your initial assumption about the likelihood of the machine being faulty before any observations
* Data (obsevations): Over the course of monitoring, you observe that the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

Using this data, find the Maximum A-Posteriori (MAP) estimate of θ (i.e., the value of θ that maximizes the posterior distribution).

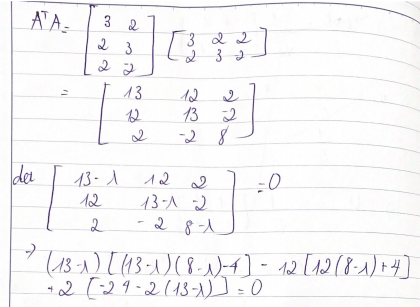


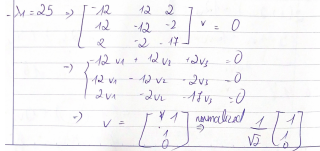
Question 8 :

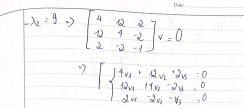
Let:

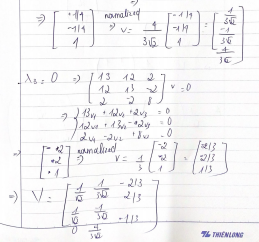


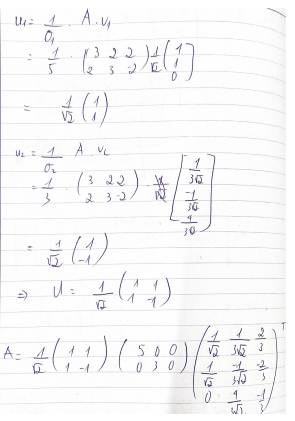
Write the singular value decomposition of S in matrix form.











Question 9:

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 |

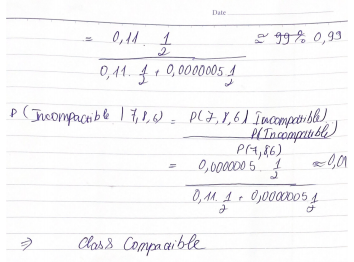
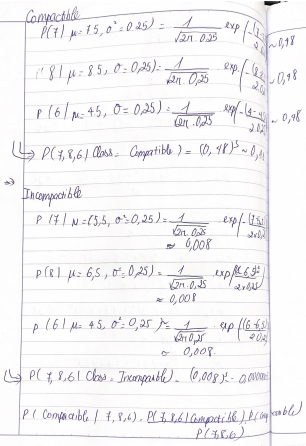
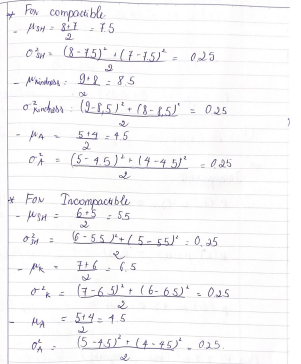
For a new candidate with the following ratings:

- Sense of Humor = 7

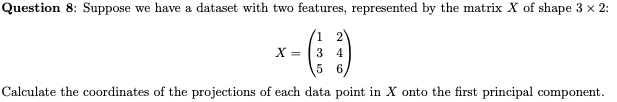
- Kindness = 8

- Ambition = 6

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

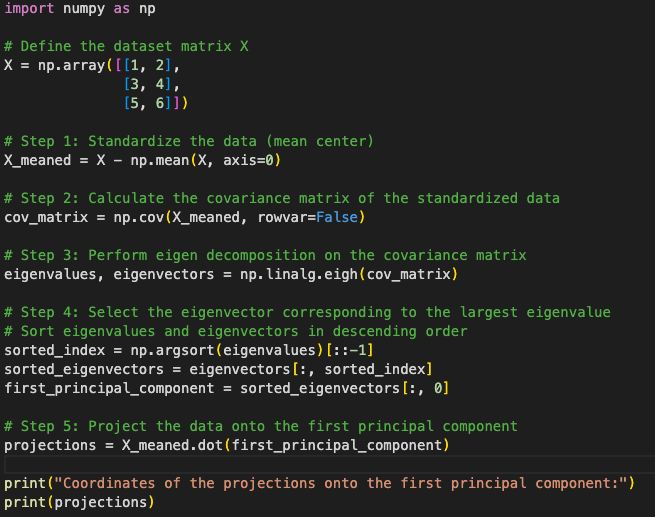


Question 10 :



1. Compute the mean of each feature (for X1 and X2) and subtract it from the dataset to center the data.
2. After centering the data, calculate the covariance matrix for the dataset.
3. Compute the eigenvalues and eigenvectors of the covariance matrix. These eigenvectors represent the principal components, and the eigenvalues show the variance explained by each component.
4. Identify the first principal component as the eigenvector corresponding to the largest eigenvalue.
5. Calculate the projection of X onto the first principal component

X\_pc1​= X. v1



Ảnh chụp Màn hình 2024-11-10 lúc 13.15.34

Question 11: Suppose we are performing linear regression using a non-linear basis expansion . Which of the following statements is true about the learned predictor?

1. 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 11 :Which of the following functions is convex?

1. f(x)= x^3−3x
2. **f(x)= x^2+2x+1**
3. f(x)= −ln(x) for x>0
4. f(x)= −x^2+4x

Question 12 : In Gaussian mixture models (GMMs), which of the following statements is false?

1. GMMs assume that the data points within each component follow a Gaussian distribution.
2. GMMs can be used for clustering.
3. **The number of components in a GMM must be equal to the number of features in the dataset.**

Question 13 : What is the key reason why backpropagation is so important?

1. Backpropogation allows us to compute the gradient of any diferentiable function.
2. Backpropogation is the only algorithm that enables us to upate the weights of a Neural Network.
3. **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.**
4. Backpropagation introduced Chain Rule into the world of mathematics, enabling signifcant advances in the feld.

Question 14 : In neural networks, nonlinear activation functions such as sigmoid, tanh, and ReLU

1. speed up the gradient calculation in backpropagation, as compared to linear units
2. **help to learn nonlinear decision boundaries**
3. always output values between 0 and 1
4. are applied only to the output units

Question 15 : 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?

1. A convolutional layer with 10 3 × 3 filters
2. **A max-pooling layer that reduces a 10 × 10 image to 5 × 5**
3. A convolutional layer with 8 5 × 5 filters
4. A fully-connected layer from 20 hidden units to 4 output units

Question 16 :

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

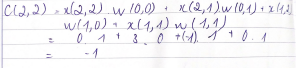
Input Image I (4x4 matrix):

A=

K = ]

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.

1. **-1**
2. 1
3. 0
4. 2



Question 17 : Suppose we have a 1000x1000x3 dimension input image (width x height x channel). We apply a convolutional layer with 50 5x5 kernels. What is the dimension of the resulting tensor (width x height x channel) if we have stride=1 and no padding?

A. 995x995x3

B. 996x996x3

C. 995x995x50

D. 996x996x50

Output Dimension= (Input Dimension−Filter Size)/Stride+1

Question 18 : 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.

=> True. Changing the initialization of a Neural Network can result in lower loss. This is because Neural Networks are non-convex, meaning that a change in initialization may converge better local minimum with lower loss.

Question 19 : 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.

=> False: Kernel regression predicts the value of a point as the weighted average of the values at nearby points, therefore if all of the points have the same value, then kernel regression will predict a constant (in this case, 1 n ) for all values.

Question 20 : 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.

F=> alse: Each LR model parameter corresponds to a whole set of possible GNB classifier parameters, there is no one-to-one correspondence because logistic regression is discriminative and therefore doesn’t model P(X), while GNB does model P(X).

Question 21 : 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.

=> False: A simple counterexample is the prior which assigns probability 1 to a single choice of parameter θ.

Question 22 : Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

=> True: The number of iterations in boosting controls the complexity of the model, therefore, a model selection procedure like cross validation can be used to select the appropriate model complexity and reduce the possibility of overfitting