COMP3314 Homework 2

Wang Qifan 3035973452

2024.05.01

1

Given the data, we can calculate the true positive, false positive, true negative and false negative as follows:

- **True Positive**: The emails correctly classified as spam, which is 80.
- **False Positive**: The emails incorrectly classified as spam, which is 20.
- **True Negative**: The emails correctly classified as non-spam, which is 380.
- **False Negative**: The emails incorrectly classified as non-spam, which is 20.

With the above data, we can calculate the precision, recall, and F1 score as follows:

- **Precision**: $PRE = \frac{TP}{TP+FP} = \frac{80}{80+20} = 0.8$
- **Recall**: $REC = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{80}{80+20} = 0.8$
- **F1 Score**: $F1 = 2 \times \frac{PRE \times REC}{PRE+REC} = 2 \times \frac{0.8 \times 0.8}{0.8+0.8} = 0.8$

Thereofre, the required values are as follows:

$$\begin{cases} \texttt{True Positive} = 80 \\ \texttt{False Positive} = 20 \\ \texttt{Precision} = 0.8 \\ \texttt{Recall} = 0.8 \\ \texttt{F1 score} = 0.8 \end{cases}$$

2

2.1

The cost function can be mathmatically repreesented as:

$$J = \sum_{j=1}^{n} \sum_{i=1}^{k} r_{ij} ||X_j - \mu_i||^2$$

where

- n is the number of data points.
- k is the number of clusters.

- r_{ij} is the indicator function, which is 1 if the jth data point is in the ith cluster, and 0 otherwise.
- X_j is the jth sample point.
- μ_i is the centroid of the ith cluster.
- $||X_j - \mu_i||$ is the Euclidean distance between the X_j and the centroid of μ_i.

Cost function

$$J = \sum_{i=1}^{n} ||x_i - \mu_{y_i}||^2$$

which is the sum of squared Euclidean distance betweem x_i and the mean of the cluster which is belongs to.

2.2

Consider the cost function for k-means clustering:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n_i} ||X_j^{(i)} - \mu_i||^2$$

where

- n_i is the number of points in the ith cluster.
- $X_j^{(i)}$ is the jth sample point in the ith cluster.
- μ_i is the mean of the ith cluster.

To minimize J with respect to μ_i, take the derivative of J with respect to μ_i:

- Let's denote J_i as the portion of the cost function that relates to the ith cluster:

$$J_i = \sum_{j=1}^{n_i} ||X_j^{(i)} - \mu_i||^2$$

Now taking the derivative of J_i with respect to μ_i: (Using the chain rule and noting that x^2 has derivative $2x$)

$$\begin{aligned} \frac{\partial J_i}{\partial \mu_i} &= \sum_{j=1}^{n_i} \frac{\partial}{\partial \mu_i} ||X_j^{(i)} - \mu_i||^2 \\ &= -2 \sum_{j=1}^{n_i} (X_j^{(i)} - \mu_i) \end{aligned}$$

To find the value of μ_i that minimizes J_i, we set the derivative to zero:

$$\begin{aligned} & \sum_{j=1}^{n_i} -2(X_j^{(i)} - \mu_i) = 0 \\ \Rightarrow \quad & \sum_{j=1}^{n_i} X_j^{(i)} = n_i \mu_i \\ \Rightarrow \quad & \mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_j^{(i)} \end{aligned}$$

This shows that the mean μ_i that minimized the within-cluseter sum of squares, and therefore the cost function J, is the centroid of teh points in cluster i, which is the average of the points assigned to cluster i.

3

3.1

We apply the max pooling operation to the input matrix X to calculate Y, choosing the maximum value in each 3x3 window, moving the window by one element each time.

The output matrix Y is as follows:

$$Y = \begin{bmatrix} 9 & 5 & 8 \\ 9 & 5 & 8 \\ 8 & 8 & 8 \end{bmatrix}$$

3.2

We use the following steps to calculate the gradient of the input matrix X:

- For the corresponding position of y in x (using the max pooling indices), place the gradient value from $\frac{\partial L}{\partial y}$ in the same position in $\frac{\partial L}{\partial x}$. Because by **chain rule**, we have $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial x}$ and $\frac{\partial y}{\partial x}$ is a matrix with 1s in the positions of the max values in the pooling windows and 0s elsewhere.
- If multiple values in the max pooling window were equal to the max value and contributed to the output, only the gradient at the position with the smallest index gets the gradient (as per the given instruction).
- All other positions in $\frac{\partial L}{\partial x}$ taht do not corresspond to the max values in the pooling windows are set to 0 since they did not influence the forward pass.

$$\frac{\partial L}{\partial y} = \begin{bmatrix} 0.1111 & -0.0007 & 0 \\ 0 & 0.1104 & 0.1111 \\ 0.1111 & 0.1111 & 0.1111 \end{bmatrix}$$

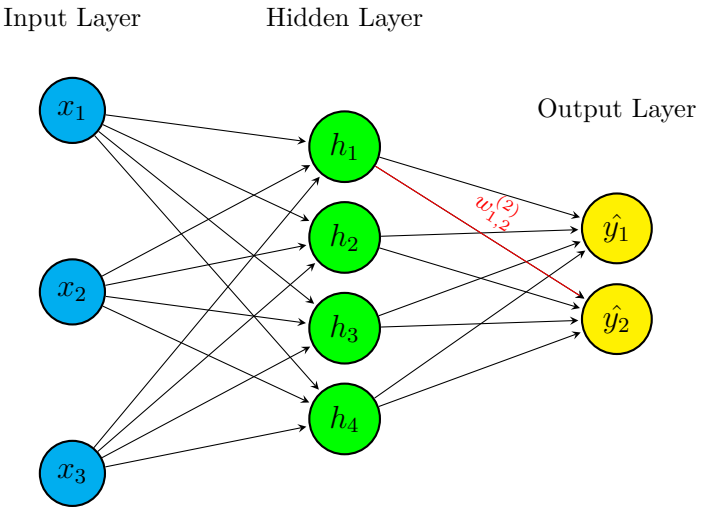
Therefore, the gradient of the input matrix X ($\frac{\partial L}{\partial x}$) is as follows:

$$\frac{\partial L}{\partial x} = \begin{bmatrix} 0 & 0 & -0.0007 & 0 & 0 \\ 0.1111 & 0 & 0.1104 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2222 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2222 & 0 & 0 \end{bmatrix}$$

4

The arhcitecture is as the following graph:

- Input Layer: 3 neurons
- Hidden Layer: 4 neurons
- Output Layer: 2 neurons



where $w_{1,2}^2$ connects the first neuron in the hidden layer and the second neuron in the output layer. (the red arrow in the graph)

$$\hat{y}_2 = w_{1,2}^{(2)} h_1 + w_{2,2}^{(2)} h_2 + w_{3,2}^{(2)} h_3 + w_{4,2}^{(2)} h_4$$

$$\frac{\partial \hat{y}_2}{\partial w_{1,2}^{(2)}} = \frac{\partial}{\partial w_{1,2}^{(2)}} (w_{1,2}^{(2)} h_1 + w_{2,2}^{(2)} h_2 + w_{3,2}^{(2)} h_3 + w_{4,2}^{(2)} h_4) = h_1$$

Because `ReLU` is used as the activation function for the hidden layer,

$$h_1 = \max(0, w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2 + w_{3,1}^{(1)} x_3)$$

The Loss function is written as:

$$L = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y_1 - \hat{y}_1)^2 + \frac{1}{2}(y_2 - \hat{y}_2)^2$$

Also, we have the derivative of the Loss function, which is

$$\frac{\partial L}{\partial \hat{y}_2} = \frac{1}{2} \cdot -1 \cdot 2(y_2 - \hat{y}_2) = \hat{y}_2 - y_2$$

Therefore, the partial derivative of the Loss function with respect to $w_{1,2}^{(2)}$ is:

$$\frac{\partial L}{\partial w_{1,2}^{(2)}} = \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial w_{1,2}^{(2)}} = (\hat{y}_2 - y_2) \cdot \max(0, w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2 + w_{3,1}^{(1)} x_3)$$

5

5.1

The mean value of the two festures are as follows:

- **Feature 1**: $\mu_1 = \frac{1}{3}(1 + 2 + 3) = 2$
- **Feature 2**: $\mu_2 = \frac{1}{3}(3 + 4 + 5) = 4$

Then the variance of teh two features are as follows:

- **Feature 1**: $\sigma_1^2 = \frac{1}{3-1}((1-2)^2 + (2-2)^2 + (3-2)^2) = 1$

- **Feature 2**: $\sigma_2^2 = \frac{1}{3-1}((3-4)^2 + (4-4)^2 + (5-4)^2) = 1$

Covariance:

$$\text{Cov}_{(1,2)} = \text{Cov}_{(2,1)} = \frac{1}{3-1}((1-2)(3-4) + (2-2)(4-4) + (3-2)(5-4)) = 1$$

Therefore, the covariance matrix is as follows:

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

5.2

Suppose the enginvalues of the matrix is λ_1 and λ_2. We have the following equation:

$$\det(\Sigma - \lambda I) = 0$$

where I is the identity matrix.

$$\begin{aligned} & \det \begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix} = 0 \\ \Rightarrow \quad & \lambda^2 - 2\lambda = 0 \\ \Rightarrow \quad & \lambda_1 = 0 \; and \; \lambda_2 = 2 \end{aligned}$$

To find the eigenvectors, we solve the following equation:

- For $\lambda_1 = 0$:

$$\begin{aligned} & (\Sigma - \lambda I)v = 0 \\ \Rightarrow \quad & \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \Rightarrow \quad & v_1 + v_2 = 0 \end{aligned}$$

Let $v_1 = \frac{\sqrt{2}}{2}$, then $v_2 = -\frac{\sqrt{2}}{2}$. We have the eigenvector for $\lambda_1 = 0$ as:

$$v_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$$

- For $\lambda_2 = \frac{4}{3}$:

$$\begin{aligned} & (\Sigma - \lambda I)v = 0 \\ \Rightarrow \quad & \begin{bmatrix} 1-2 & 1 \\ 1 & 1-2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \Rightarrow \quad & v_1 = v_2 \end{aligned}$$

Let $v_1 = \frac{\sqrt{2}}{2}$, then $v_2 = \frac{\sqrt{2}}{2}$. We have the eigenvector for $\lambda_2 = \frac{4}{3}$ as:

$$v_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

5.3

Since we use the first two principle components, we just choose $\lambda_1 = 0$ and $\lambda_2 = 2$. The projection matrix is as follows:

$$W = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

Standardize the original data:

$$X = \begin{bmatrix} \frac{1-\mu_1}{\sigma_1} & \frac{3-\mu_2}{\sigma_2} \\ \frac{2-\mu_1}{\sigma_1} & \frac{4-\mu_2}{\sigma_2} \\ \frac{3-\mu_1}{\sigma_1} & \frac{5-\mu_2}{\sigma_2} \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

Therefore, the projection of the original data onto the first two principle components is:

$$XW = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} 0 & -\sqrt{2} \\ 0 & 0 \\ 0 & \sqrt{2} \end{bmatrix}$$

5.4

Tov reconstruct the original data from the projected data, we use the following formula:

$$X_{\text{reconstructed}} = X_{means} + XW^T$$

$$X_{\text{reconstructed}} = \begin{bmatrix} 2 & 4 \\ 2 & 4 \\ 2 & 4 \end{bmatrix} + \begin{bmatrix} 0 & -\sqrt{2} \\ 0 & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 3 & 5 \end{bmatrix}$$