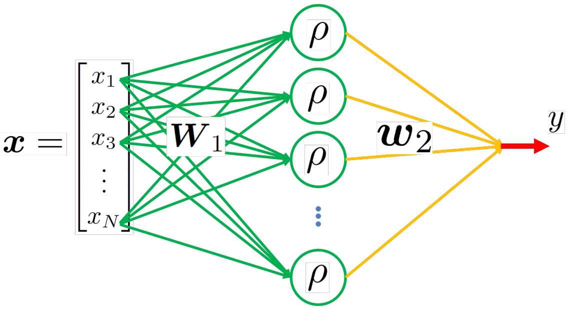
COMP 6721 Applied Artificial Intelligence (Winter 2023)

Assignment #2: Naive Baysian, Dicision Tree, Neural Netowrk

Due: 11:59PM, February 28th, 2023

Theoretical Part

Question 1 Gradient back-propagation technique is one of the fundamental algorithms for training feedforward neural networks. Using the chain rule, this algorithm calculates the gradient of the loss function at different layers of the network. In subsequent stages, computed gradients will be used to update weights with optimizers such as gradient descent or stochastic gradient descent to minimize a loss function. In this question, we want to drive an expression for the gradient of a cost function with respect to the weights and biases of a simple neural network. Consider a 1-hidden layer neural network as follows



where $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is the input feature vector and $\boldsymbol{y} \in \mathbb{R}$ is the network output. The network's weights are $\boldsymbol{W}_1 \in \mathbb{R}^{N\times M}$ and $\boldsymbol{w}_2 \in \mathbb{R}^{M}$, biases are $\mathbf{b}_1 \in \mathbb{R}^{M}$ and $\mathbf{b}_2 \in \mathbb{R}$ and activation function ρ.

(a) Following the Neural Network lecture from your course, derive the feed-forward equation that maps the input to the output i.e. $\boldsymbol{y} = f(\boldsymbol{x};\Theta)$, where $\Theta = \{\boldsymbol{W}_1, \boldsymbol{w}_2, \mathbf{b}_1, \mathbf{b}_2\}$ is the set of all learnable parameters.

(b) Consider a cost function

$$J = \frac{1}{B}\sum_{i=1}^{B} l\left(\hat{y}^i, y^i\right), \text{ where } l(\hat{y}, y) = (\hat{y} - y)^2,$$

where B is the batch size for optimization. Using the chain-rule, derive the expressions for the following gradients

$$\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial \boldsymbol{w}_2}, \frac{\partial J}{\partial \boldsymbol{b}_1}, \frac{\partial J}{\partial \boldsymbol{b}_2}.$$

Question 2 Consider the squared loss $\mathcal{L}(X,w,y) = \frac{1}{2}\|Xw - y\|^2$ for data matrix $X \in \mathbb{R}^{N\times D}$, weights $w \in R^{D\times 1}$, and outputs $y \in R^{N\times 1}$.

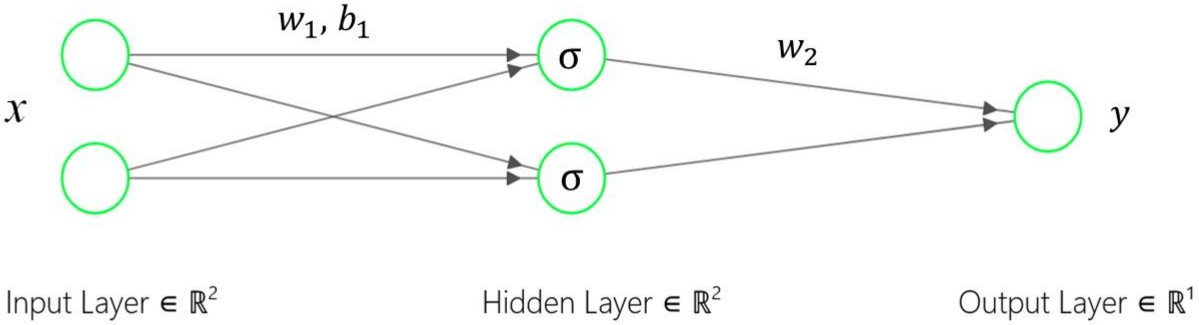
(a) Find the expression for gradient $\nabla_w \mathcal{L}(X,w,y)$ and minimizer of this loss, $\arg\min_w \mathcal{L}(X,w,y)$. (Hint: See the example on page 96 of Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press, Link .)

(b) Take w_0 as the initialization for gradient descent with step size α and show an expression for the first and second iterates w_1 and w_2 only in terms of α, w_0, X, y.

(c) Generalize this to show an expression for w_k in terms of α, w_0, X, y, k.

(d) Write a pseudo code for calculating the w_k in terms of α, w_0, X, y, k.

Question 3 Consider the following 1-hidden neural networks with 2 inputs and a single output:


w_1, b_1
w_2
x
σ
σ
y
Input Layer ∈ $\mathbb{R}^2$
Hidden Layer ∈ $\mathbb{R}^2$
Output Layer ∈ $\mathbb{R}^1$

We can write the below equation for the given neural network:

$$y = \mathbf{w}_2^T \sigma\left(\mathbf{w}_1^T \mathbf{x} + \mathbf{b}\right).$$

The loss function for training this neural network is:

$$l(y,t) = \frac{1}{3}(y-t)^3.$$

Consider the activation function σ as **Sigmoid** and values for parameters as:

$$\mathbf{w}_1 = \begin{bmatrix} -1 & 0.5 \\ 1 & 0.5 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix} \quad \mathbf{w}_2 = \begin{bmatrix} -1 \\ 1.0 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mathbf{t} = 1$$

Show the sequence of steps in backpropagation to get $\frac{\partial l}{\partial \mathbf{x}}, \frac{\partial l}{\partial \mathbf{b}}$. (Hint: $\sigma'(x) = \sigma(x)(1-\sigma(x))$, You may use intermediate variables in your answer.)

Implementation

Question 1 (**K-means**) K-means clustering can be used in image compression. It works on clustering specific (K) numbers of colors to represent the image color instead of actual number of colors and in this way, it reduces image size. Obviously, it clusters pixels with colors similar to each other and considers one value for them. Pleaes note that image *"bird.png"* is uploaded for this assignment.

It has dimension of *rows*columns* pixels and each pixel consists three channels of RGB showing color and intensity. Image data can be considered as arrays of [rows, columns, 3]. Using *'image_compression.py'*, find cluster centers for this image as [centroid, 3]. You ought to:

(a) Cluster image pixels using predefined python libraries for K-means.

(b) Find suitable value K, report accuracy and attach your *'compress_image.png'* by your report.

Question 2 (**Naive Bayes**) Consider the following table:

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Attributes are Color, Type, Origin, and the subject, stolen can be either "yes" or "no".

(a) design a Naive Bayes classifier by hand to determine the class of the Red Domestic SUV

(b) using the pre-defined functions of scikit-learn package, train Naive Bayes classifier to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set.

Question 3 (**Decision Tree**) Breast cancer is the most frequent reason for cancer mortality among women, which needs to be detected earlier in order to decrease the death rate. In this question, we aim to work on the Breast Cancer Wisconsin (Diagnostic) Data Set. You can download the data from the following link `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)` for diagnosis of breast cancer with Decision Tree (DT). There are 569 data points in the dataset: 212 – Malignant, and 357 –Benign. In this dataset, features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image.

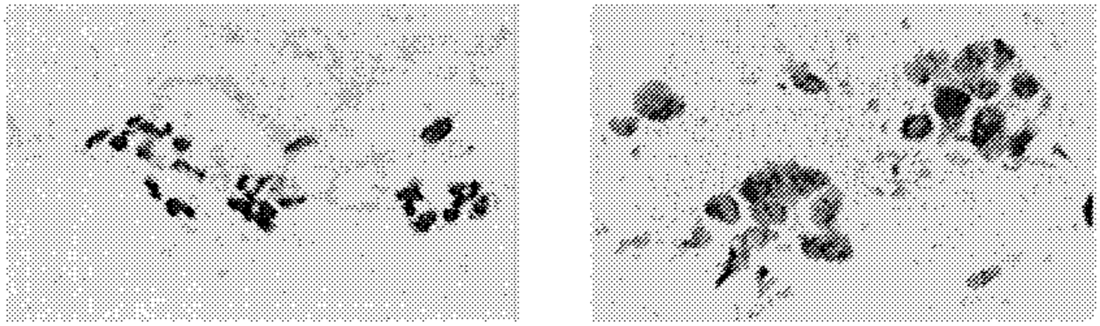


Figure 1: Images taken using the FNA test: (a) Benign, (b) Malign

(a) Using the NumPy or Pandas package, load the dataset. (Dataset "*breast_cancer_wisconsin.csv*" is uploaded for this assignment). Then split the dataset into train and test sets with a test ratio of 0.3.

(b) Using the scikit-learn package, define a DT classifier with custom hyperparameters and fit it to your train set. Measure the precision, recall, F-score, and accuracy on both train and test sets. Also, plot the confusion matrices of the model on train and test sets.

(c) Study how maximum tree depth and cost functions of the following can influence the efficiency of the Decision Tree on the delivered dataset. Describe your findings.

   i. three different cost functions: `['gini','entropy','log_loss']`

   ii. six different maximum tree depth: `[2,4,6,8,10,12]`

(d) Depict a plot of the decision boundary of the two mentioned hyperparameters. Comment on the fundamental features in short.