Assignment 2

Deadline: 23:55, December 4, 2022

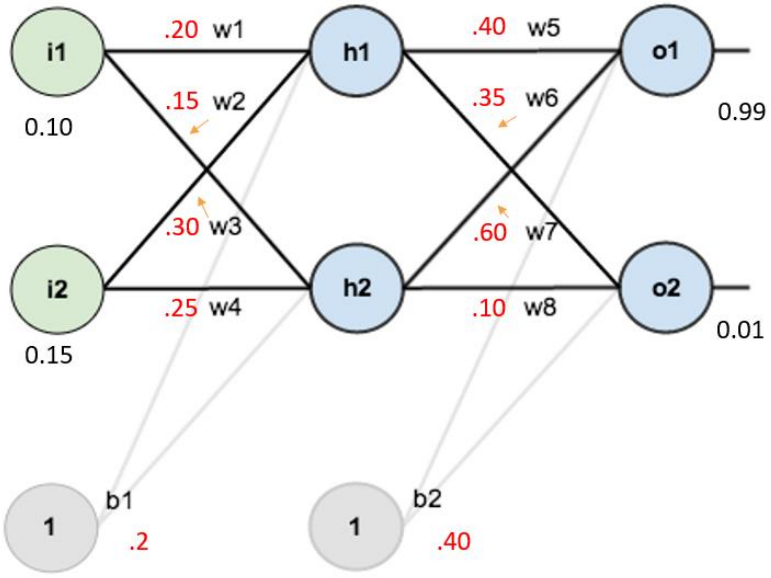
1. We can use logistic regression to do binary classification, since the output $h_\theta(\boldsymbol{x}) = \frac{1}{1+e^{-f_\theta(x)}}$ can be viewed as the probability that the input $\boldsymbol{x}$ belongs to the class 0. Then, we can formulate the objective function in the form of cross entropy loss like below:

$$\min_\theta L_\theta = \frac{1}{N}\sum_{n-1}^{N}[-y^{(n)}\ln\left(1 - h_\theta\left(\boldsymbol{x}^{(n)}\right)\right) - (1 - y^{(n)})\ln\left(h_\theta\left(\boldsymbol{x}^{(n)}\right)\right)]$$

(1) Relate L_θ to the probability $\prod_{n=1}^{N} P(y^{(n)}|\boldsymbol{x}^{(n)})$. **Show your steps clearly.** (10 marks)

(2) Compute the derivate of L_θ with respect to θ. **Show your steps clearly**. (10 marks)

2. There is a neural network with two inputs, two hidden neurons, two output neurons. Additionally, the hidden and output neurons will include a bias. Backpropagation is a common method for training a neural network. The goal of backpropagation is to optimize the weights so that the neural network can learn how to correctly map arbitrary inputs to outputs. The goal is: given inputs 0.1 and 0.15, the expected outputs of the neural network are 0.99 and 0.01. The initial values for the network parameters are shown in the following figure.



(1) Try to calculate two outputs, i.e., out_{o1} and out_{o2} in the first forward pass step. **Show your steps clearly.** (10 marks)
Hint: forward pass step includes *inputs to hidden layer neurons, squash the total input using an activation function (here we use the sigmoid function)* and *hidden layer outputs to output layer neurons.*

(2) Try to calculate the error for each output neuron using the *squared error function* and sum them to get the total error E_{total}. **Show your steps clearly.** (10 marks)

(3) Consider w_5. We want to know how much a change in w_5 would affect the total error, a.k.a. $\frac{\partial E_{total}}{\partial w_5}$, the gradient with respect to w_5. **Show your steps clearly.** (20 marks)
Hint: Firstly, you can apply the chain rule to get the equation of $\frac{\partial E_{total}}{\partial w_5}$. Then you can figure out each piece in the equation. Finally, to decrease the error, you can subtract this value from the current weight (optionally multiplied by some learning rate, η, which we'll set to 0.1) to get the new weight w_5^{new} in the spirit of the gradient descent method.

3. You are given the following data set, where

- "name" is not an attribute;
- attributes "eye-color" and "hair-color" are discrete;
- attributes "height" and "weight" are numeric;
- the class label "region" has three possible values: Asia, Europe, and America.

name	height	weight(kg)	eye-color	hair-color	region
Abel	183	82	hazel	black	Europe
Bob	176	83	brown	black	Europe
Carl	177	76	blue	blond	Europe
Dale	168	69	blue	blond	Asia
Eric	175	66	brown	blond	Asia
Felix	170	63	hazel	brown	Asia
George	186	79	blue	brown	America
Howard	184	84	hazel	black	America
Igor	183	72	brown	brown	America

Find the feature that should be used in the decision tree's root according to the information gain. **Show your steps clearly**. (40 marks)