KTH Royal Institute of Technology
Stockholm

School of Electrical Engineering and
Computer Science

Scalable Machine Learning and Deep Learning - ID2223

Review Questions 2

Author
Emil Ståhl

Author
Selemawit Fsha

Author
Erik Kongpachith

November 20, 2021

1 Which of the following is/are true about individual tree in Random Forest?

(a) Individual tree is built on a subset of the features.

True

(b) Individual tree is built on all the features.

(c) Individual tree is built on a subset of instances.

True

(d) Individual tree is built on full set of instances.

Answer: (a) and (c)
Random forest is based on bagging concept, that consider faction of sample and faction of feature for building the individual trees.

2 Ensemble model estimators (such as Random Forest) in Spark have a parameter called featureSubsetStrategy. What does it do?

featureSubsetStrategy specify the number of features to use as candidates for splitting at each tree node, as a fraction of the total number of features. Possible values of **featureSubsetStrategy** is ***auto, all, onethird, sqrt, log2, n***. If **featureSubsetStrategy** is chosen as ***auto***, the algorithm chooses the best feature subset strategy automatically. If the numTrees == 1, the **featureSubsetStrategy** is set to be all. However, if the numTrees >1 (that is, forest), the **featureSubsetStrategy** is set to be ***sqrt*** for classification. Example in Spark:

```
val gbt = new GBTRegressor().setLabelCol("label").setFeaturesCol("features")
                            .setMaxIter(10).setFeatureSubsetStrategy("auto")
```

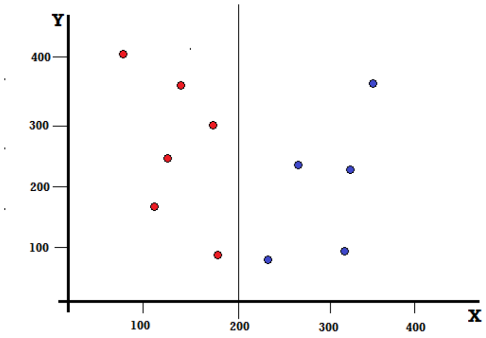
3 Explain why the entropy becomes zero when all class partitions are pure?

Entropy: the average information needed to identify the class label of an instance in the dataset.

It can also be seen as the degree of uncertainty, impurity or disorder of a random variable, or a measure of purity. Here, if all elements belong to a single class, then it is termed as “Pure”, and if not then the distribution is named as “Impurity”. It is computed between 0 and 1, however, heavily relying on the number of groups or classes present in the data set it can be more than 1 while depicting the same significance i.e. extreme level of disorder. In more simple terms, If a dataset contains homogeneous subsets of observations, then no impurity or randomness is there in the dataset, and if all the observations belong to one class, the entropy of that dataset becomes zero.

4 Explain why the Gini impurity becomes zero when all class partitions are pure?

When all the class partitions are pure, they all belong to the same class and the gini impurity becomes 0. Consider the following example where we have a so called "perfect split".



The gini impurity can be defined as

$$G = \sum_{i=1}^{C} p(i)(1 - p(i))$$

where C is the number of classes and p(i) is the probability of picking a datapoint with class i. For the formula we get the following:

$$G_Left = 1(1-1) + 0(1-0) = 0$$

$$G_Right = 0(1-0) + 1(1-1) = 0$$

5 Assume a feedforward neural network with one hidden layer, in which the output of the hidden units and output units are computed by functions h = f(x) and out = g(h), respectively. Show that if we use linear functions in f and g, e.g., h = f(x) = w1x and out = g(h) = w2h, then the feedforward network as a whole would remain a linear function of its input.

The function **z(x)** for the whole feedforward neural network of its input can be defined as

$$z(x) = g(f(x))$$

This can be written as

$$z(x) = xW$$

where

$$W = w_1^T w_2^T$$

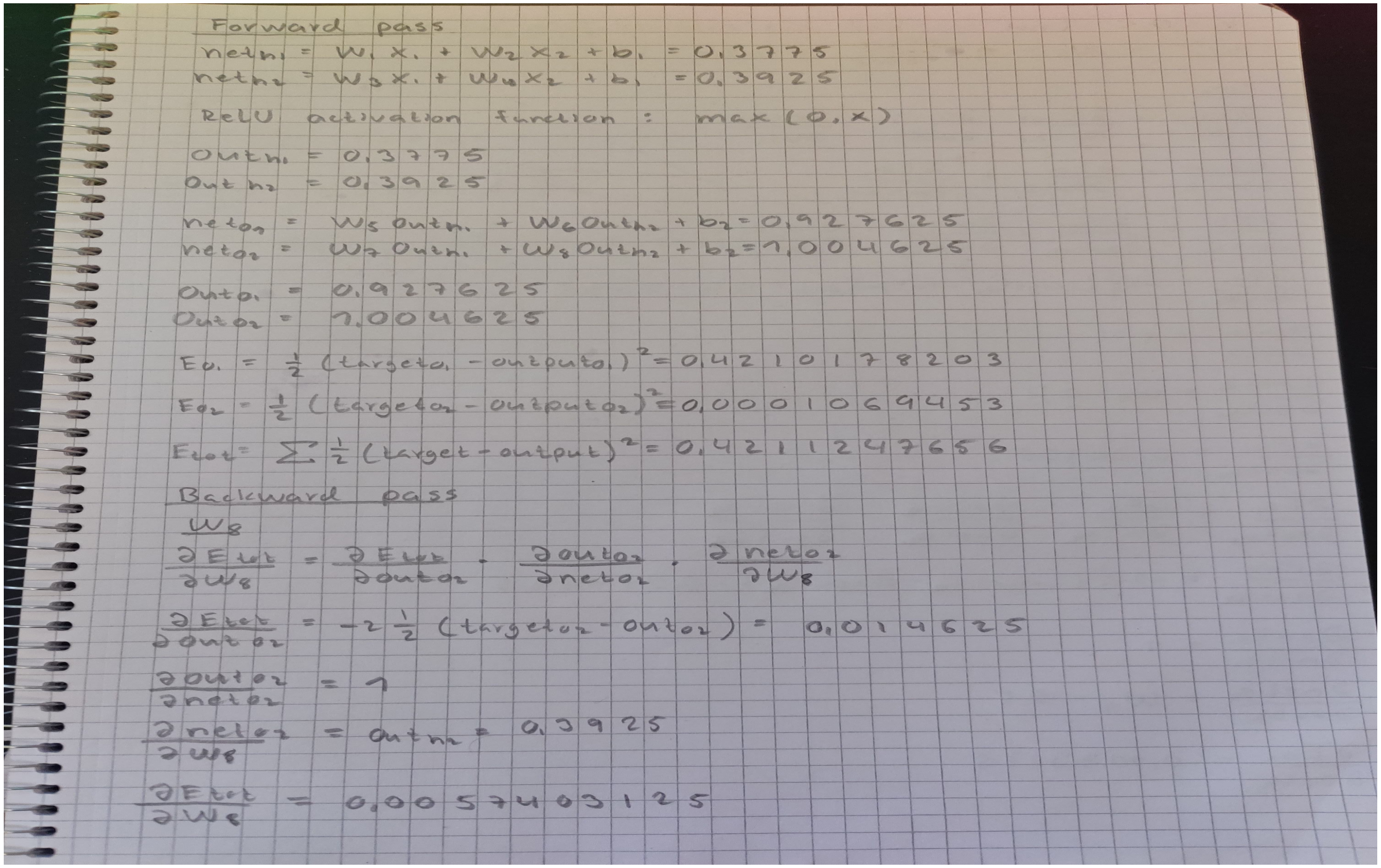
Since **f(x)** and **g(h)** are linear functions, **z(x)** would be a linear function.

6 What's the problem of using step function as an activation function in deep feedforward neural networks?

The main objective of the neural network is to learn the values of the weights and biases of the model. Therefore it is advantageous to have small changes in the weights or bias to cause only small change in the output from the network. This makes it easier to tweak the values of the model to make better predictions. A step function which only generates 0 or 1 will make the learning task hard. An efficient ways to train a multi-layer neural network is by using gradient descent with backpropagation. Backpropagation requires having a **differentiable** activation function. The step function is non-differentiable at x = 0 and it has 0 derivative elsewhere. Thus the gradient descent won't be able to make a progress in updating the weights of the neural network.

7 Compute the value of w2 and w8 after the first iteration of the backpropagation in the following figure. Assume all the neurons use the ReLU activation function and we use squared error function as the cost function. In this figure, red and orange colors indicate the initial values of the weights and biases, while the numbers in blue show the input and true output values.

Calculation shown on next page:



Forward pass

$net_{h_1} = W_1 x_1 + W_2 x_2 + b_1 = 0{,}3775$

$net_{h_2} = W_3 x_1 + W_4 x_2 + b_1 = 0{,}3925$

ReLU activation function : max (0, x)

$Out_{h_1} = 0{,}3775$

$Out_{h_2} = 0{,}3925$

$net_{o_1} = W_5\, Out_{h_1} + W_6\, Out_{h_2} + b_2 = 0{,}927625$

$net_{o_2} = W_7\, Out_{h_1} + W_8\, Out_{h_2} + b_2 = 1{,}004625$

$Out_{o_1} = 0{,}927625$

$Out_{o_2} = 1{,}004625$

$E_{o_1} = \frac{1}{2}(target_{o_1} - output_{o_1})^2 = 0{,}421017 8203$

$E_{o_2} = \frac{1}{2}(target_{o_2} - output_{o_2})^2 = 0{,}0001069453$

$E_{tot} = \sum \frac{1}{2}(target - output)^2 = 0{,}4211247656$

Backward pass

W_8

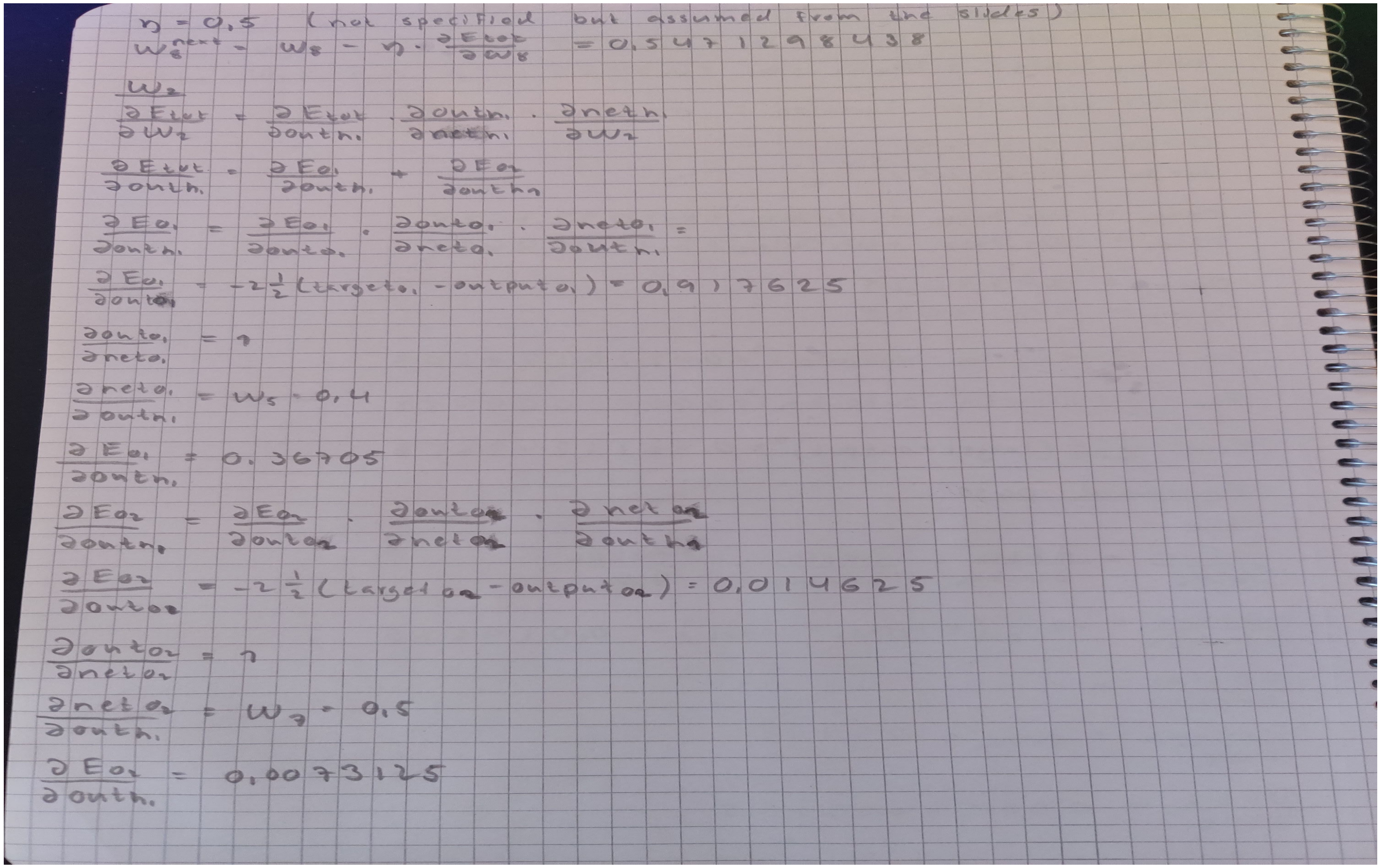
$$\frac{\partial E_{tot}}{\partial W_8} = \frac{\partial E_{tot}}{\partial out_{o_2}} \cdot \frac{\partial out_{o_2}}{\partial net_{o_2}} \cdot \frac{\partial net_{o_2}}{\partial W_8}$$

$$\frac{\partial E_{tot}}{\partial out_{o_2}} = -2\,\frac{1}{2}(target_{o_2} - out_{o_2}) = 0{,}014625$$

$$\frac{\partial out_{o_2}}{\partial net_{o_2}} = 1$$

$$\frac{\partial net_{o_2}}{\partial W_8} = out_{h_2} = 0{,}3925$$

$$\frac{\partial E_{tot}}{\partial W_8} = 0{,}00574403125$$



$\eta = 0,5$ (not specified but assumed from the slides)

$W_8^{next} = W_8 - \eta \cdot \frac{\partial E_{tot}}{\partial W_8} = 0,547129838$

<u>W_2</u>

$$\frac{\partial E_{tot}}{\partial W_2} = \frac{\partial E_{tot}}{\partial out_{h_1}} \cdot \frac{\partial out_{h_1}}{\partial net_{h_1}} \cdot \frac{\partial net_{h_1}}{\partial W_2}$$

$$\frac{\partial E_{tot}}{\partial out_{h_1}} = \frac{\partial E_{o_1}}{\partial out_{h_1}} + \frac{\partial E_{o_2}}{\partial out_{h_1}}$$

$$\frac{\partial E_{o_1}}{\partial out_{h_1}} = \frac{\partial E_{o_1}}{\partial out_{o_1}} \cdot \frac{\partial out_{o_1}}{\partial net_{o_1}} \cdot \frac{\partial net_{o_1}}{\partial out_{h_1}} =$$

$$\frac{\partial E_{o_1}}{\partial out_{o_1}} = -2\frac{1}{2}(target_{o_1} - output_{o_1}) = 0,917625$$

$$\frac{\partial out_{o_1}}{\partial net_{o_1}} = 1$$

$$\frac{\partial net_{o_1}}{\partial out_{h_1}} = W_5 = 0,4$$

$$\frac{\partial E_{o_1}}{\partial out_{h_1}} = 0.36705$$

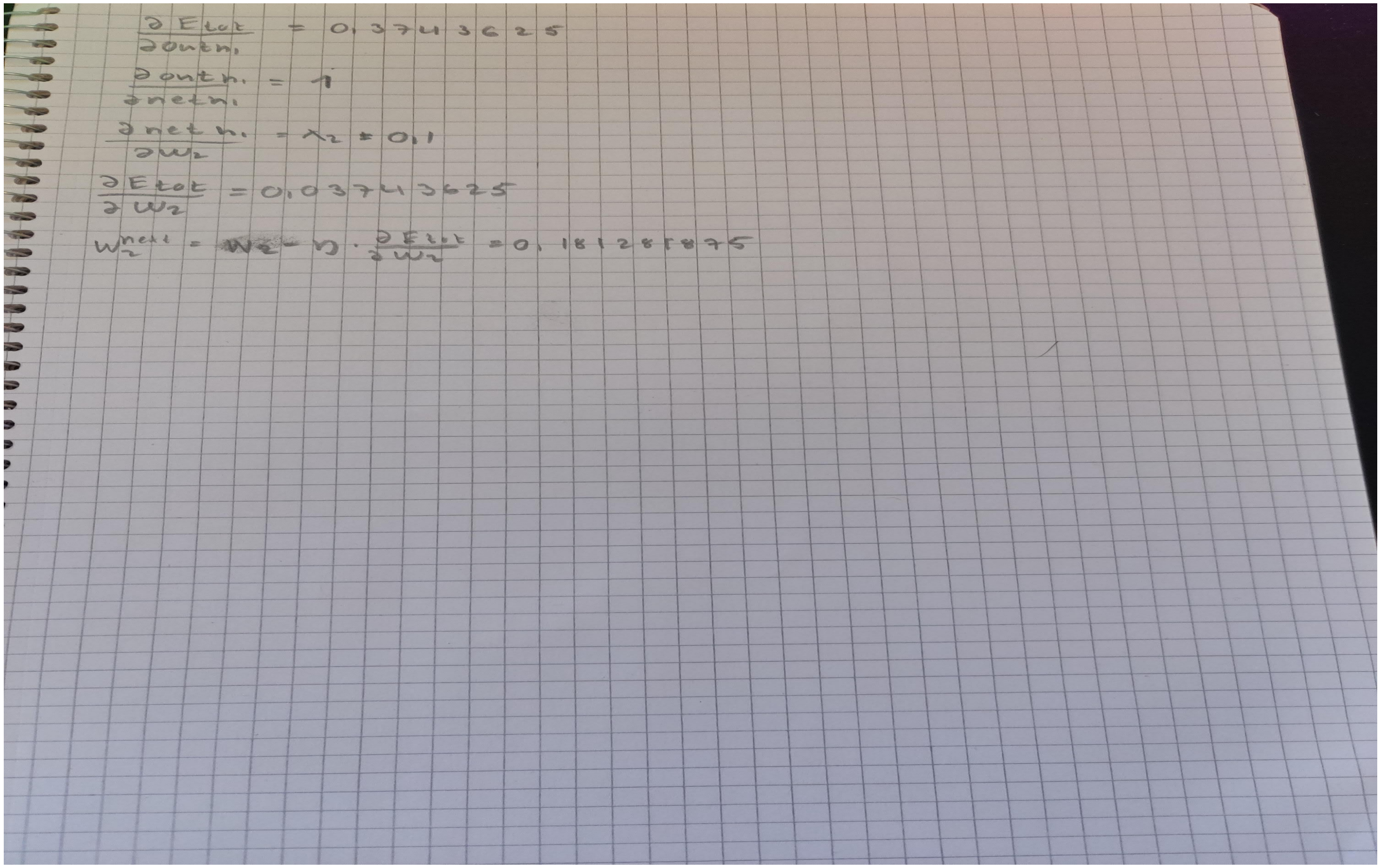
$$\frac{\partial E_{o_2}}{\partial out_{h_1}} = \frac{\partial E_{o_2}}{\partial out_{o_2}} \cdot \frac{\partial out_{o_2}}{\partial net_{o_2}} \cdot \frac{\partial net_{o_2}}{\partial out_{h_1}}$$

$$\frac{\partial E_{o_2}}{\partial out_{o_2}} = -2\frac{1}{2}(target_{o_2} - output_{o_2}) = 0,014625$$

$$\frac{\partial out_{o_2}}{\partial net_{o_2}} = 1$$

$$\frac{\partial net_{o_2}}{\partial out_{h_1}} = W_7 = 0,5$$

$$\frac{\partial E_{o_2}}{\partial out_{h_1}} = 0,0073125$$



$$\frac{\partial E_{tot}}{\partial out_{h_1}} = 0,3743625$$

$$\frac{\partial out_{h_1}}{\partial net_{h_1}} = 1$$

$$\frac{\partial net_{h_1}}{\partial w_2} = x_2 = 0,1$$

$$\frac{\partial E_{tot}}{\partial w_2} = 0,03743625$$

$$w_2^{next} = w_2 - \eta \cdot \frac{\partial E_{tot}}{\partial w_2} = 0,181281875$$