Name: _____ **Date:** _____

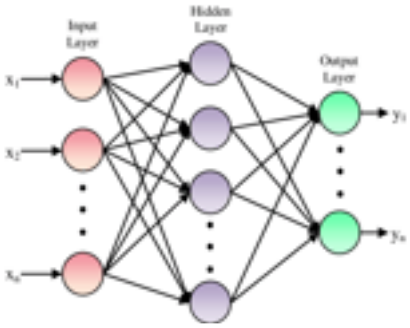
Intro to Deep Learning

(Circle the correct answer)

Q1: What is the minimum number of layers that should be present in a deep neural network?

A: Three - the input layer (i.e., data), a hidden layer, and the output layer.

Image File Name : https://factile-dynamic.s3.us-east-1.amazonaws.com/68cc12c3926a13dd3d11a925/img-1758206213210pf56116.png



Q2: What was Nvidia's original target market with GPUs?

A: Video games / gamers

Q3: True or False: Neural networks require that input data (predictors) be unstructured, e.g., images, audio, text, video.

A: False. Neural networks can be employed with structured data, too. Nothing about the algorithm(s) requires unstructured inputs.

Q4: What are the two things that TensorFlow is really good at?

A: Tensor manipulation and gradient calculations.

Q5: What is a key drawback of deep learning that makes it unsuitable for use in some industries / contexts?

A: Deep nets are notoriously 'black box' algorithms, which makes them difficult to interpret despite advancements in explainable AI. Some regulated industries require explainable decisions, e.g., loan application denials in the US must be explained.

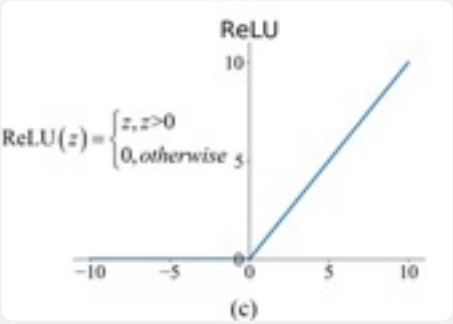
Q6: True or False. Inside a Perceptron, we activate the input and then calculate the sum-product of weights before adding the bias

term.

A: False. The activation comes last, *after* the WX+b.

Q7: What does ReLU stand for?

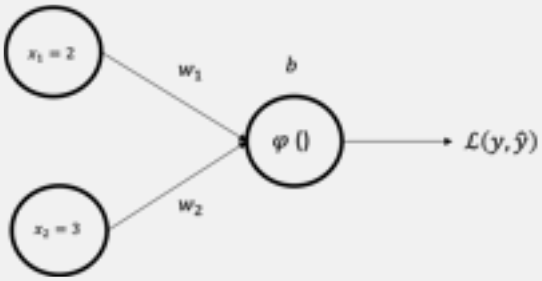
Image File Name : https://factile-dynamic.s3.us-east-1.amazonaws.com/68cc12c3926a13dd3d11a925/img-1758206473758pf89071.png



A: Rectified linear unit.

Q8: True or False. A 1-node neural network with a sigmoid activation and cross-entropy loss is mathematically equivalent to a Logistic regression, where the weights are the regression coefficients.

Image File Name : https://factile-dynamic.s3.us-east-1.amazonaws.com/68cc12c3926a13dd3d11a925/img-1758206136122pf9111.png



A: True!

Q9: True or False. A softmax activation should only ever appear in the output layer.

A: True. This is strictly used for multi-class classifications (nominal labels).

Q10: When does it make sense to use a linear activation function in a neural network?

A: In the output layer, specifically when you need to be able to predict any continuous numerical value.

Q11: What does TPU stand for?

A: Tensor Processing Unit. This is a Google-produced chip that is purpose built for tensor calculations (more efficient than a GPU in theory).

Q12: What do we mean when we refer to the 'rank' of a tensor?

A: The rank refers to the tensor's dimensionality (number of dimensions). For example, a rank-0 tensor is a scalar (a point), a rank-1 tensor is a vector (a line), a rank-2 tensor is a matrix (a square), a rank-3 tensor is a 3-dimensional array (a cube), and so on.

Q13: What rank tensor would be necessary to store a single grayscale image?

A: A rank-2 tensor (a matrix, or two-dimensional array).

Q14: What rank tensor would be required to store a batch of color videos?

A: We would require a rank-5 tensor (Vertical + Horizontal + Color + Time + Batch).

Q15: What shape matrix would I obtain if I multiply a 3x3 matrix with a 2x5 matrix?

A: Shape conformity error! We calculate a matrix product by taking the dot products of each row from matrix 1 with each column of matrix 2. For the math to work, the second dimension of the first matrix has to match the first dimension of the second matrix ($3 != 2$).

Q16: What does MSE stand for?

A: Mean squared error.

Q17: What is the name of the following loss function? (-1 / N) * Sum[log(p)*y + log(1-p)*(1-y)]

A: Cross-entropy loss or log-loss

Q18: True or False. Loss functions are a concept specific to neural networks.

A: False. Loss functions are employed in optimization and machine learning more generally. When you fit a linear regression in Excel or Stata or R, the

optimization algorithm employs an L2 or MSE loss to fit the line.

Q19: If you have a single observation in your data and this loss function, what would the resulting loss be in a scenario where the ground truth was 1, yet you predicted 0.0?

Image File Name : https://factile-dynamic.s3.us-east-1.amazonaws.com/68cc12c3926a13dd3d11a925/img-1758207613227pf85718.png

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

A: Infinity.

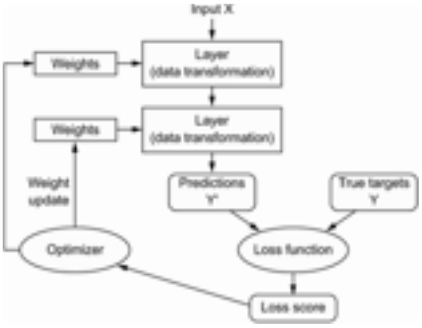
Q20: True or False: Employing MAE as the loss for a binary prediction problem will result in your model failing to converge.

A: False. Using MAE, your model may still converge, but it will likely take longer. Instead of penalizing log(1-abs(p-y)), MAE will penalize abs(p-y). Basically, the penalty for your prediction moving away from the label is increasing at a lower rate with MAE.

Q21: What is the name of the optimization approach we have discussed in class, i.e., forward-pass/inference/predict, gradient calculation, updating weights (rinse and repeat)?

A: Gradient descent

Image File Name : https://factile-dynamic.s3.us-east-1.amazonaws.com/68cc12c3926a13dd3d11a925/img-1758207980305pf80100.png



Q22: What is the name of TensorFlow's feature for implementing and evaluating computation graphs?

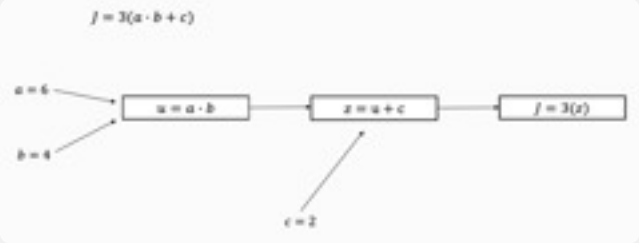
A: Gradient Tape.

Q23: What is the name of the derivative 'rule' that underpins backpropagation?

A: The chain rule. The basic idea with backprop is derivatives of nested functions. Ex: J = 3(U*2), U = a+1, then dJ/da can be calculated as dJ/dU*dU/da = 3*2 * dU/da = 3*2*1 = 6.

Q24: Considering this computation graph. What is the partial derivative of J with respect to b?

Image File Name : https://factile-dynamic.s3.us-east-1.amazonaws.com/68cc12c3926a13dd3d11a925/img-1758208055793pf72730.png



A: 18! If b increases by 1, J will increase by 3*a. The current value of a is 6, so J would increase by 3*6 = 18.

Q25: True or False. Backpropagation, as a method, guarentees that you arrive at a globally optimal solution.

A: False. You never know when you might have been able to do better, e.g., a different random initialization of weights might yield a better solution / result.