A neural network is being trained for a regression problem in which the target is in $\mathbb{R}$. Which of the following is the *most appropriate* activation function at the output layer?

OPTIONS :

- ◯ Linear (or identity function)
- ◯ ReLU
- ◯ Tanh
- ◯ Sigmoid
- ◯ Either Tanh or Sigmoid

Your score : 0

Consider applying dropout to a hidden layer with p = 0.5. Which of the following are true?

OPTIONS :

- ☐ During training, a randomly chosen set of neurons in the layer are droppedout. The neurons to be dropped are determined dynamically during each iteration.
- ☐ During inference (testing), no neurons are dropped out. Instead, the activationof each neuron in the layer is scaled by 0.5.
- ☐ During training, a fixed set of neurons in the layer are dropped out. This setremains the same in every iteration and the neurons to be dropped are determined before the training begins.
- ☐ During inference (testing), a randomly chosen set of neurons in the layer aredropped out.

Your score : 0

Consider a problem in which we have images of digits, similar to the MNIST dataset. The image size is 100 × 100. Each pixel value lies in the range [0, 255]. In the context of data augmentation, which of the following transformations will help a neural network trained on this dataset to generalize better?

OPTIONS :

- ☐ Translate the digits by at most 5 pixels, horizontally or vertically
- ☐ Flip the image about the vertical or horizontal axis.
- ☐ Add a random noisy image of size 100×100 with each pixel sampled from a normal distribution with mean $\mu = 200$ and standard deviation $\sigma = 5$
- ☐ Rotate the digits by any angle, clockwise or anti-clockwise.

The input volume to a convolutional layer is 30 × 30 × 5. If ten kernels, each of size 7 × 7, with unit stride are applied over this volume, what should be the padding so that the output volume has dimensions 30 × 30 × 10? Note that you should enter the value of P as per the convention we have been following.

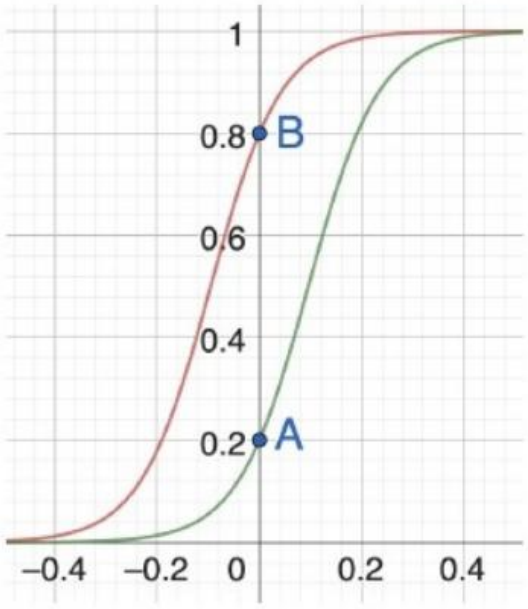
Answer (Numeric):

Answer

Accepted Answer : 3

Based on the above data, answer the given subquestions.

Consider two sigmoid neurons with the same weight, w, but different biases b_1 and b_2. Their outputs are denoted by the functions $\sigma_1(x)$ and $\sigma_2(x)$ respectively. The graph of $\sigma_1(x)$ crosses the y-axis at the point A with a value of 0.2, while $\sigma_2(x)$ crosses the y-axis at the point B with a value of 0.8. These two sigmoid neurons are combined to produce a tower whose function is given by $\sigma_2(x) - \sigma_1(x)$.



Find the value of x at which the tower attains its maximum value.

Answer (Numeric):

Answer

Accepted Answer : 0

Your score : 0

Find the maximum value that the tower attains.

Answer (Numeric):

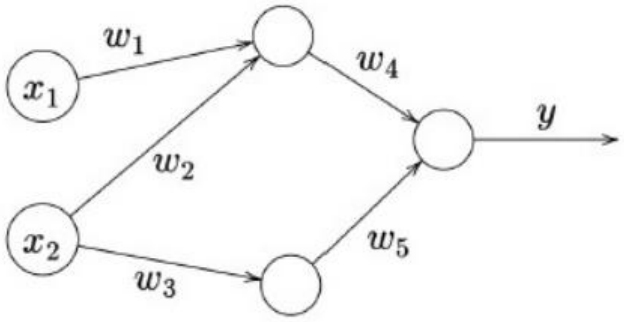
Answer

Accepted Answer : 0.6

Your score : 0

Based on the above data, answer the given subquestions.

Consider the network given below with one hidden layer. Note that unlike a typical fully connected layer, the first neuron in the input is connected to only the first neuron in the hidden layer:



x_1 and x_2 correspond to input neurons. The other three neurons are ReLU neurons. Ignore the biases. Given the following information:

- $w_1 = 2, w_2 = -1, w_3 = -5, w_4 = 2, w_5 = 3$
- $x_1 = 2, x_2 = 1$

Find the value y at the end of the forward pass.

Answer (Numeric):

Answer

Accepted Answer : 4

Your score : 0

In the backward pass, find the value of $\frac{\partial y}{\partial w_1}$ evaluated at the given configuration.

Answer (Numeric):

Answer

Accepted Answer : 4

Your score : 0

In the backward pass, find the value of $\frac{\partial y}{\partial w_3}$ evaluated at the given configuration.

Answer (Numeric):

Answer

Accepted Answer : 0

Your score : 0

Based on the above data, answer the given subquestions.

Consider a sigmoid neuron with three inputs and weight vector $w = (1, 2, -2)$. Ignore the bias. Among all inputs with unit norm, let $x^* = (a, b, c)$ be the unit-norm vector that maximizes the neuron's output. use L2 norm. Enter your answer correct to two decimal places for all three given sub-questions.

Find a.

Answer (Numeric):

Answer

Accepted Answer : 0.30 to 0.36

Find b.

Answer (Numeric):

Answer

Accepted Answer : 0.64 to 0.70

Your score : 0

Find c.

Answer (Numeric):

Answer

Accepted Answer : -0.70 to -0.64

For a multi-class classification problem with three classes, consider the following CNN architecture:

Layer	Specs	Volume
Input	NA	$32 \times 32 \times 3$
Convolution-1	$F = 3, S = 1, P = 1, K = 6$	V_1
MaxPooling-1	$F = 2, S = 2, P = 0$	V_2
Convolution-2	$F = 3, S = 1, P = 1, K = 12$	V_3
MaxPooling-2	$F = 2, S = 2, P = 0$	V_4
Convolution-3	$F = 1, S = 1, P = 0, K = 8$	V_5
FC-1	10	NA
Output	3	NA

- The first column is the type of layer.
- The second column is the layer specification. If it is an FC layer, it is the number of neurons. If it is a convolution or pooling layer, it is the information pertaining to kernels. NA refers to "Not Applicable" wherever this information is not needed. In a convolutional/pooling layer, K is the number of filters, F is the spatial dimension of the filter, P is the padding and S is the stride.
- The third column corresponds to the activation volumes output by the non-FC layers. For example, the input layer passes on a volume of size $32 \times 32 \times 3$ to "Convolution-1", which outputs a volume of size V_1. Each volume is of the type $W \times H \times D$.

Note that V_5 is flattened before it is passed on to "FC-1". Output layer is also considered as an FC layer with a softmax activation function.

Find the number of parameters associated with the layer “Convolution-2” that are required to transform V2 to V3. Ignore biases.

Answer (Numeric):

Answer

Accepted Answer : 648

Which of the following corresponds to V4, the activation volume output by the layer “MaxPooling-2”?

OPTIONS :

- ◯ 8 × 8 × 12
- ◯ 16 × 16 × 12
- ◯ 8 × 8 × 8
- ◯ 32 × 32 × 12

Your score : 0

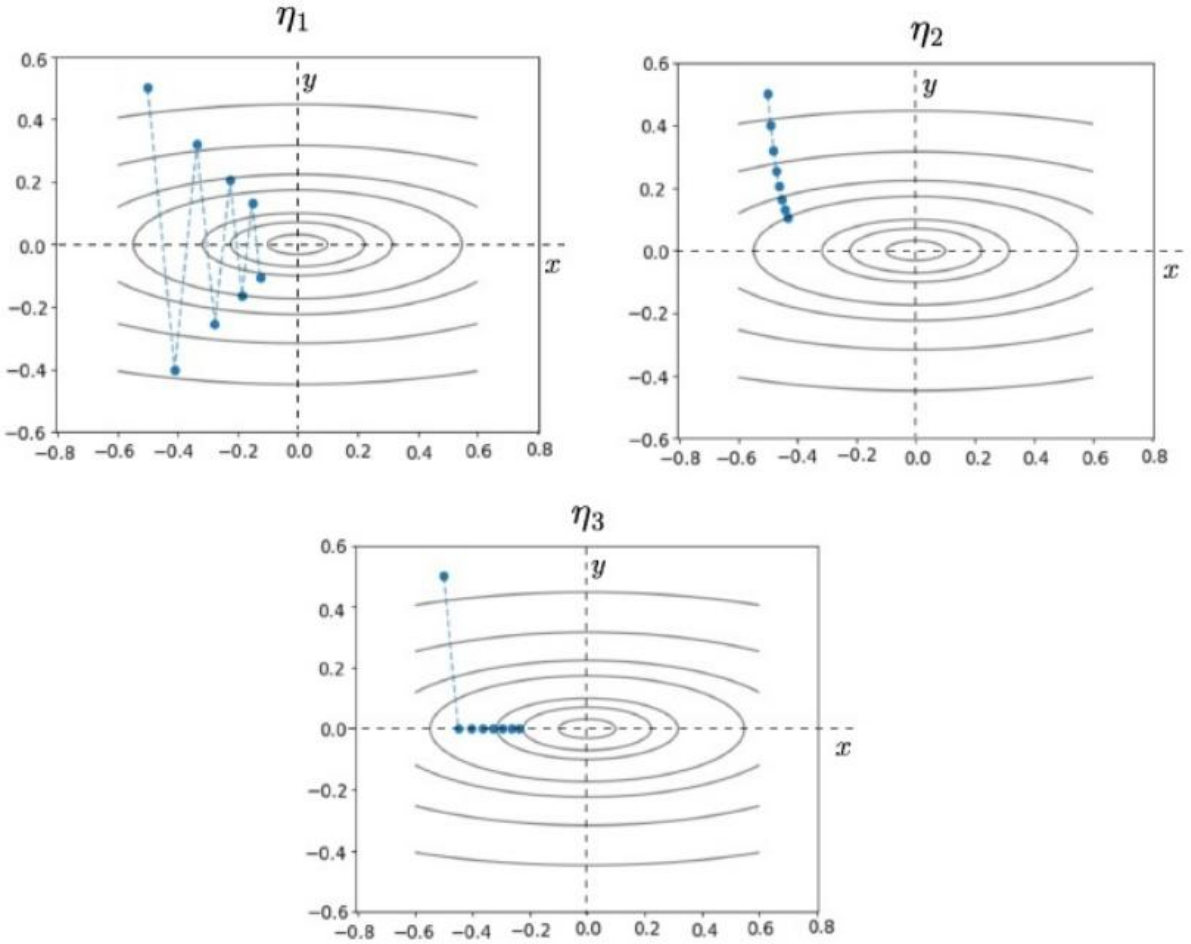
The parameters associated with the fully connected layers, namely “FC- 1” and “Output”, represent what percentage of the total number of parameters in the network? Ignore biases.

OPTIONS :

- ◯ 85%
- ◯ 70%
- ◯ 20%
- ◯ 50%

Your score : 0

Consider optimizing a function $f(x, y) = bx^2 + y^2$ with $b = 0.1$ using gradient descent. Starting with $(x_0, y_0) = (-0.5, 0.5)$, three separate runs of gradient descent are done with three distinct learning rates, η_1, η_2, η_3. The algorithm is run for seven iterations in each case. The contours of the objective function along with the individual iterates are shown below. Assume that the images are drawn to scale:



Which of these is the correct ordering among the learning rates?

OPTIONS :

- ◯ $\eta_2 < \eta_3 < \eta_1$
- ◯ $\eta_3 < \eta_2 < \eta_1$
- ◯ $\eta_1 < \eta_3 < \eta_2$
- ◯ $\eta_1 < \eta_2 < \eta_3$

Which of the following is η_3?

OPTIONS :

- ◯ 0.5
- ◯ 0.9
- ◯ 0.1
- ◯ 1

Your score : 0

Starting with $(x_0, y_0) = (1, 1)$, run 5 iterations of gradient descent with $\eta = \eta_3$ (as identified in the previous question). Enter the value of x_5 correct to two places after the decimal. Note that you will obtain the iterates, $(x_1, y_1), \cdots, (x_5, y_5)$, in this process.

Answer (Numeric):

Answer

Accepted Answer : 0.54 to 0.64

Your score : 0

Which of the following threshold θ values for an MP neuron implements the AND Boolean function, denoted as $f(\mathbf{x})$? Assume that the number of inputs x_i to the neuron is five and the neuron does not have any inhibitory inputs.

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{i=1}^{5} x_i \geq \theta \\ 0, & otherwise \end{cases}$$

OPTIONS :

- ◯ 5
- ◯ 3
- ◯ 1
- ◯ 0

Your score : 0

You are given a dataset of 10×10 *grayscale* images. Your goal is to build a 5-class classifier. You have to adopt one of the following two options:

- **Model A**: the input is flattened into a 100-dimensional vector, followed by a fully-connected layer with 5 neurons without bias
- **Model B**: the input is directly given to a convolutional layer with five 10×10 filters

Suppose you make your choice on the basis of the number of parameters in the models, p_A = number of parameters in **ModelA**, and similarly let p_B = number of parameters in **ModelB**.

OPTIONS :

- ○ $p_A > p_B$
- ○ $p_A = p_B$
- ○ $p_A < p_B$
- ○ None of these

Your score : 0

As per gradient descent, we should move towards 180 degrees with respect to gradient direction. What will happen if we move between 90 and 180 degrees? Consider the loss function to be convex.

OPTIONS :

- ○ The loss function will increase.
- ○ The loss function will decrease, although not to the maximum possible extent.
- ○ The loss function will remain the same.
- ○ Can't say. It depends on other parameters of the convex function.

Your score : 0

Which of the following statements about the parameter **β** in RMSProp is correct?

OPTIONS :

- ○ A higher value of **β** means that past gradients retain more influence on themoving average.
- ○ A lower value of **β** means past gradients have a lesser influence on the movingaverage.
- ○ A higher value of **β** means the current gradient will have lesser influence thanpast gradients.
- ○ A lower value of **β** means the current gradient will have higher influence thanpast gradients.
- ○ All of these

Your score : 0

Consider a feed-forward neural network containing three inputs and one output. It consists of 50 hidden layers, each with two neurons. The activation function used in all the hidden layers is ReLU, the output layer uses the Sigmoid activation function for binary classification. The network is trained with a binary cross-entropy loss function. The network architecture is as follows:

- Input layer: 3 nodes - Hidden layer: 2 nodes each - Output layer: 1 node

Assume all weights are initialized to 1 and all biases to 0. For an input of [-10, 5, -20], what will be the output of the neural network?

OPTIONS :

- ◯ 0
- ◯ 1
- ◯ 0.5
- ◯ 0.25
- ◯ -25
- ◯ Insufficient information

Your score : 0

Based on the above data, answer the given subquestions.

Consider a sigmoid neuron that takes in an input vector $\mathbf{x} = \begin{bmatrix} 1 \\ 0.5 \\ 1 \end{bmatrix}$. The weight vector $\mathbf{w}$ is initialized to $\begin{bmatrix} 0.5 \\ 1 \\ 0.5 \end{bmatrix}$ and $\mathbf{b} = 1$. The output from the sigmoid neuron is

$$z = \mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathbf{b}$$

$$\hat{y} = \frac{1}{1 + exp^{-(z)}}$$

Suppose we use the following loss function

$$\mathcal{L} = \frac{1}{2}(y - \hat{y})^2$$

What is the value of z
(write correct upto 2 decimals).

Answer (Numeric):

Answer

Accepted Answer : 2.47 to 2.53

What is the value of $\hat{y}$
(write correct upto 2 decimals).

Answer (Numeric):

Answer

Accepted Answer : 0.89 to 0.95

Your score : 0

Update the weight vector once by running the vanilla Gradient Descent algorithm with $\eta = 2$. Assume the true label is $\mathbf{y} = \mathbf{0}$. What is the new loss value (that is, the loss computed after updating the weights and biases)? (consider upto two digits after the decimal for all the calculations)

Answer (Numeric):

Answer

Accepted Answer : 0.36 to 0.42

Your score : 0

Which of the following techniques does NOT help prevent a model from overfitting?

OPTIONS :

- ◯ Data augmentation
- ◯ Dropout
- ◯ Early stopping
- ◯ None of these

Your score : 0

Given a neuron with three binary inputs x_1, x_2, and x_3 (all take values either 0 or 1). The neuron gives the output as follows:

$$\hat{y} = \begin{cases} 1 & \text{if } x_1 - 2x_2 + 4x_3 > \theta \\ 0 & \text{if } x_1 - 2x_2 + 4x_3 \leq \theta \end{cases}$$

What is the minimum threshold value θ for which the neuron outputs 0 for all possible input vectors (x_1, x_2, x_3)?

Answer (Numeric):

Answer

Accepted Answer : 5

Consider a dataset with 150 samples and a batch size of 15. If each minibatch iteration contributes an average loss of 0.4, what will be the total loss after 15 epochs?

Answer (Numeric):

Answer

Accepted Answer : 60

Your score : 0

Consider a feedforward neural network with the following structure:

One input layer with 2 nodes

One hidden layer with 2 nodes

One output layer with 1 node

All weights and biases are initialized to zero. The activation function used in the hidden layer is the Rectified Linear Unit (ReLU), and the output layer uses the Sigmoid activation for binary classification. The network is trained with a binary cross-entropy loss function.

Two training examples are given: 1. Input vector: [2, -3], true label: 1 2. Input vector: [-1, 1], true label: 0

What will be the value of the total binary cross-entropy loss given these two training examples?

Answer (Numeric):

Answer

Accepted Answer : 1 to 1.5

A neural network has the following structure:

- **Input Layer:** $\mathbf{h_0} = \mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^{100}$
- **Hidden Layers:** Two hidden layers ($\mathbf{h_1}$ and $\mathbf{h_2}$), each with 120 neurons, using the sigmoid activation function.
- **Output Layer:** $\mathbf{O}$ with 8 neurons, using the softmax activation function.

Assuming that all weights between layers $\mathbf{h_2}$ and $\mathbf{O}$ are initialized to 0.2, with no bias associated with any neuron, what would be the computed cross-entropy loss for a given single data point? If the provided information is insufficient, please enter -1.

Answer (Numeric):

Answer

Accepted Answer : 2 to 2.2

Given an input array X and a kernel/filter K as follows:

$$X = \begin{bmatrix} -1 & -1 & 0 & 2 \\ -2 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$K = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

- Convolve the kernel K over the input X with a stride $s = 1$ and no padding to obtain matrix A.
- Apply average pooling on A to produce matrix B.
- Pass B through the sigmoid (logistic) function to get the final output $\hat{y}$.

Given that $\frac{\partial L}{\partial \hat{y}} = -1$, determine the value of $\frac{\partial L}{\partial K_{00}}$, where K_{00} is the element of K at index $(0, 0)$.

Answer (Numeric):

Answer

Accepted Answer : 0.02 to 0.22

What is the derivative of the ReLU activation function at x = 10?

Answer (Numeric):

Answer

Accepted Answer : 1

In terms of convergence speed, which gradient descent method can show the most rapid progress initially but may suffer from high variance in updates?

OPTIONS :

- ◯ Batch Gradient Descent
- ◯ Stochastic Gradient Descent
- ◯ Mini-batch Gradient Descent
- ◯ None of these

Your score : 0

How does the use of early stopping in training a neural network affect the model's performance on unseen data?

OPTIONS :

- ◯ It usually leads to better performance on unseen data by preventingoverfitting
- ◯ It generally worsens the performance on unseen data by halting training tooearly
- ◯ It does not affect the performance on unseen data
- ◯ It increases the risk of overfitting by allowing more epochs of training

Which of the following statements is/are not true with respect to a dropout rate of 0.2?

OPTIONS :

- ◯ The exact number of neurons dropped in each iteration will always be exactly20%.
- ◯ The exact number of neurons dropped and retained can vary slightly from oneiteration to another due to the probabilistic nature of dropout.
- ◯ Each neuron has a 20% chance of being dropped during any given trainingiteration.
- ◯ Over many training iterations, the average percentage of retained neuronswill approximate 80%.

In the context of unsupervised pretraining of artificial neural networks, which of the following statements accurately describes the role and benefits of using unsupervised pretraining techniques for initializing a neural network?

OPTIONS :

- Unsupervised pretraining methods help in identifying patterns in unlabeleddata, which can be used to initialize weights and reduce the risk of overfitting in the subsequent supervised training phase.
- The primary purpose of unsupervised pretraining is to generate synthetic datathat can be used to expand the training dataset for the neural network, leading to more robust performance.
- Unsupervised pretraining enables the network to learn a hierarchicalrepresentation of data, which can be fine-tuned with supervised learning, enhancing the model's generalization capabilities.
- Using unsupervised pretraining techniques ensures that the neural networkcan skip the initial training phase, directly achieving high accuracy on test data without further training.

What are the maximum values of the derivatives of sigmoid and tanh?

OPTIONS :

- 1, 1
- 0.5, 0.5
- 0, 0.5
- 0.5, 0
- 0.25, 1
- 0.25, 0.5

Your score : 0

Consider a scenario where you have a dataset with overlapping classes (that is instances from different classes share similar or identical feature values), and you decide to train a perceptron model for classification. Assertion (A): The perceptron model may struggle to classify instances accurately when classes overlap in the feature space. Reason (R): The perceptron learning algorithm aims to find a linear decision boundary that separates the classes, and in the presence of overlapping classes, it may not be able to capture the underlying patterns effectively. Select the correct option:

OPTIONS :

- ◯ Both A and R are true, and R is the correct explanation of A.
- ◯ Both A and R are true, but R is not the correct explanation of A.
- ◯ A is true, but R is false.
- ◯ A is false, but R is true.

Consider a feedforward neural network with one hidden layer trained using backpropagation for a binary classification task. The network has the following architecture:

- Input layer with 15 neurons
- Hidden layer with 25 neurons
- Output layer with 1 neuron

During the backpropagation process, the derivative of the sigmoid activation function $\sigma(z)$ with respect to its argument z is given by:

$$\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$$

If the loss function used for binary classification is the binary cross-entropy loss, and the activation fuction at hidden layer and output layer is sigmoid. The output of the neural network is denoted as $\hat{y}$, and the true label is denoted as y, what is the expression for $\frac{\partial L}{\partial w_j}$, where w_j represents the weights connecting the jth neuron of hidden layer to the output layer? Assume that the output of jth neuron of hidden layer is h_j and no biases in the network.

OPTIONS :

- $\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma'(w_j h_j) \cdot h_j$
- $\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma'\left(\sum_{i=1}^{25} w_i\right) \cdot h_j$
- $\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma'(h_j) \cdot h_j$
- $\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma'\left(\sum_{i=1}^{25} w_i h_i\right) \cdot h_j$

Your score : 0

In the context of mini-batch gradient descent, if doubling the size of the mini-batch makes your model take twice as many epochs to reach convergence, how does this affect the total number of parameter updates compared to using the original mini-batch size? Assume everything else remains constant.

OPTIONS :

- The number of updates required is doubled.
- The number of updates required is four times.
- The number of updates required is halved.
- The number of updates remains the same.

Your score : 0

Suppose you are working with a sparse dataset and using Momentum-based Gradient Descent (GD) and AdaGrad algorithms for optimization. Which of the following statements accurately describes the behavior of weight and bias term updates?

OPTIONS :

- Bias vector will update frequently only in the case of the AdaGrad algorithm.
- Weight vector will have very few updates in the case of Momentum-based GD.
- Weight vector will have frequent updates in both cases.
- Bias vector will have very few updates in the case of Momentum-based GD.

Your score : 0

State True or False. Sigmoid activation function helps mitigating vanishing gradient problem better than ReLU activation function.

OPTIONS :

◯ FALSE

◯ TRUE

Your score : 0

Consider an ensemble consisting of 3 models, where each model has an individual error rate of 50% on the test set. Assuming a simple majority voting scheme, what is the probability that a given instance is misclassified by the ensemble? Assume that the models are independent and their errors are uncorrelated.

Answer (Numeric):

Answer

Accepted Answer : 0.45 to 0.55

Your score : 0

Given the true outputs for a dataset, y = [1,2,3,4], and the predictions from three different models (A, B, and C) trained on different subsets of the training data, as follows: Model A predictions: A = [1.1,1.9,3.1,3.9] Model B predictions: B = [0.9,2.1,2.9,4.1] Model C predictions: C = [2,2,3,4] Calculate the squared bias (bias2) for these models evaluated in a regression task. Assume that each model uses the same parameters but is trained on different subsets of the training data. (Enter your answer up to 3 decimal places.)

Answer (Numeric):

Answer

Accepted Answer : 0.02 to 0.03

Your score : 0

Consider a Convolutional Neural Network (CNN) architecture for image classification with the following layers: 1. Convolutional layer with 32 filters of size 3 × 3, with a stride of 1 and no padding. 2. Max pooling layer with a pool size of 2 × 2 and a stride of 2. 3. Convolutional layer with 64 filters of size 4 × 4, with a stride of 1 and no padding. 4. Max pooling layer with a pool size of 2 × 2 and a stride of 2. 5. Fully connected layer with 128 neurons. 6. Output layer with 10 neurons (for 10 classes) using softmax activation. If the input image size is 64 × 64 × 3, and the network has no bias term, how many parameters are there in the CNN?

Answer (Numeric):

Answer

Accepted Answer : 28256

Which of the following threshold θ values of MP neuron implements OR Boolean function denoted by $f(\mathbf{x})$? Assume that the number of inputs x_i to the neuron is five and the neuron does not have any inhibitory inputs.

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{i=0}^{4} x_i > \theta \\ 0, & otherwise \end{cases}$$

OPTIONS :

- [ ] 4
- [ ] 3
- [ ] 2
- [ ] 0
- [ ] 1

Consider two data points $\mathbf{x_1} = \begin{bmatrix} a \\ a \end{bmatrix}$ and $\mathbf{x_2} = -1 * \mathbf{x_1}$, where $a > 0$. The data point $\mathbf{x_1}$ belongs to positive class (denoted as 1) and the datapoint $\mathbf{x_2}$ belongs to negative class (denoted by 0). Suppose that the perceptron learning algorithm is used to find the decision boundary that separates these data points with the following rule,

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{w^T x} \geq 0, \\ 0 & \mathbf{w^T x} < 0 \end{cases}$$

The algorithm checks $\mathbf{x_1}$ in the first iteration and $\mathbf{x_2}$ in the second iteration and so on. How many times the weights get updated until convergence (That is, the algorithm classifies both the points correctly)? The weights do not include bias.

Assume the weights are initialized to zero

OPTIONS :

- ◯ 1
- ◯ 2
- ◯ 4
- ◯ It oscillates and never converges

Your score : 0

Consider a sigmoid neuron. Suppose that input $x = 10$, output $y = 1$ and the parameters, $w = 0.1, b = 0.1$. The loss function is $L = \frac{1}{2}(y - \hat{y})^2$. Update the parameters **once** using GD with η=1. Enter the sum of updated parameter values. (that is, if $w = 0.01$ and $b = 0.09$ after updating them, then you need to enter the sum 0.1)

Answer (Numeric):

Answer

Accepted Answer : 0.68 to 0.72

Your score : 0

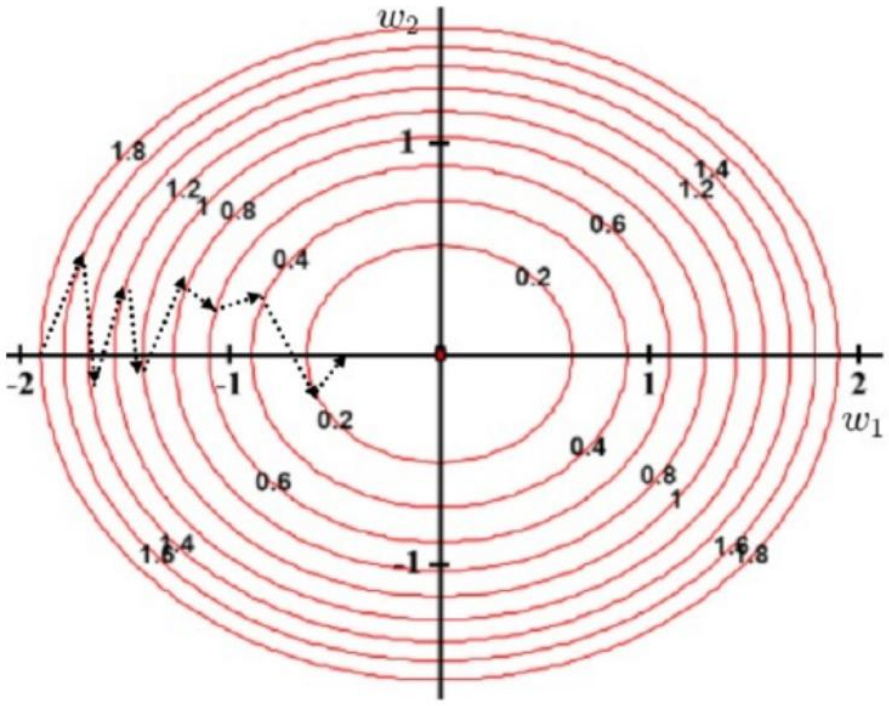
Consider a neural network with three hidden layers and one output layer. The hidden layers contain 100 neurons each. Suppose we have a square image of size 30×30 containing either a cat (class-1) or a dog (class-2). The neural network is designed to recognize it by outputting a probability distribution over two classes using softmax activation at the output layer. The input image is flattened into an array of size 900. Assume that all neurons in the network have bias associated with them and use the sigmoid activation function in the hidden layers. How many parameters are there in the network?

Answer (Numeric):

Answer

Accepted Answer : 110502

Consider the contours of a loss surface as shown in the figure below. The parameters (w_1, w_2) were initialized to $(w_1 = -2, w_2 = 0)$. Suppose we run an optimization algorithm (not necessarily gradient-based) for a few iterations. The dotted arrows in the figure show the trajectory of updated parameters in each iteration.



Choose the correct statements that are inferred from the figure

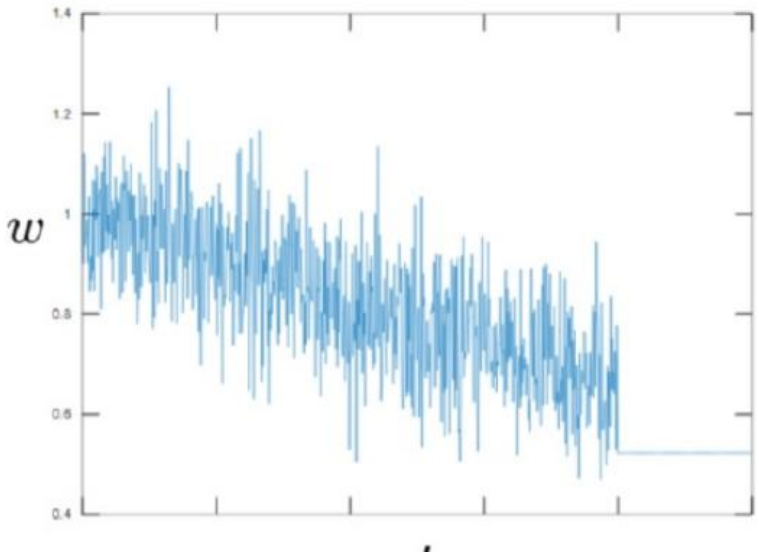
Choose the correct statements that are inferred from the figure

OPTIONS :

- [ ] The loss surface is convex (with a global minimum)
- [ ] The loss value oscillates over iterations
- [ ] The loss value decreases consistently over iterations
- [ ] The loss value increases consistently

Your score : 0

Suppose we have a fully connected neural network with three hidden layers containing 10 neurons each. All the neurons in the network use the standard ReLU activation function. Suppose we use cross entropy loss with Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$ to update the parameters. The figure below shows the change in the value of one of the weights w in the first hidden layer over the entire training duration.



Based on the given information and from the figure, it is implied that.

OPTIONS :

- [ ] Definitely, the loss value at the end of the training is zero
- [ ] The neuron in the hidden layer might be experiencing a vanishing gradientproblem
- [ ] The loss could have converged to the local minimum after a finite number ofiterations
- [ ] The loss value at the end of the training may not necessarily be zero

Your score : 0

Suppose that a neural network has millions of parameters (weights and biases). A team decides to use an optimization algorithm with a learning rate scheme that is local to each parameter in the network. Moreover, the learning rate changes in each iteration such that it should decrease on the steep surface and increase on the gentle surface. Which of the following optimization algorithms satisfy the team's requirements?

OPTIONS :

- [ ] GD with an exponentially decaying learning rate scheduler
- [ ] AdaGrad
- [ ] AdaM
- [ ] NADAM
- [ ] RMSProp
- [ ] SGD with line search

The diagram below shows contours of a loss function $\hat{L}(\theta)$ where, $\hat{L}(\theta) =$

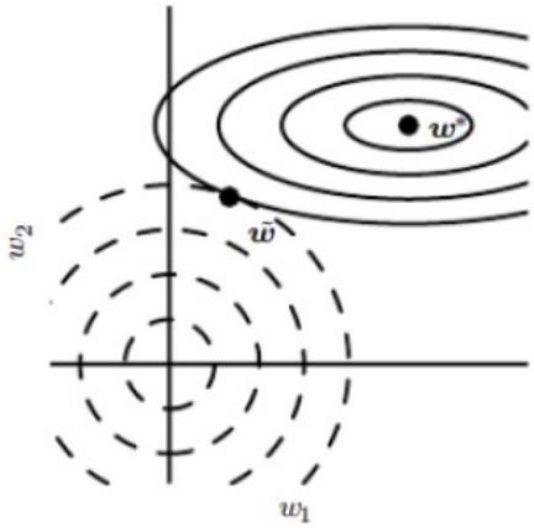


Figure 1: Contours

$L(\theta) + \alpha\Omega(\theta)$. Here, $L(\theta)$ is an un-regularied loss function and $\Omega(\theta)$ is a regularization erm. Suppose we use L_2 regularization, then choose the correct statements from the ollowing statements (with respect to this diagram)

OPTIONS :

- [ ] The dotted circles represent contours of regularization term $\Omega(\theta)$
- [ ] The dotted circles represent contours of loss term $L(\theta)$
- [ ] The solid circles represent the contours of loss term $L(\theta)$
- [ ] The solid circles represent the contours of regularization term $\Omega(\theta)$
- [ ] Increasing the value of α make the parameters sparse
- [ ] Decreasing the value of α makes the parameter sparse
- [ ] The L_2 regularization is independent of the input samples used to train the network

Your score : 0

Consider an intermediate feature map H obtained after applying convolution operation on the input X using kernel F.

$$H = \begin{bmatrix} 1 & 2 & 0 & 0 & -1 & 1 \\ 3 & 0 & 2 & 1 & 3 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 1 & 2 & 0 & 0 \\ 4 & 0 & 0 & 3 & -2 & 1 \end{bmatrix}$$

Apply the max-pooling operation with stride $s = 2$ and no padding ($p = 0$) and store the resultant output in matrix H_m. The prediction $\hat{y}$ is simply the sum of elements in H_m. Suppose $\frac{\partial L}{\partial \hat{y}} = 2.5$ (that is, the gradient of loss with respect to the prediction). What is the gradient $\frac{\partial L}{\partial H_{10}}$ where H_{10} is the element at the 1-st row and 0-th column? If you think the given info is insufficient, enter -1 as the answer.

Answer (Numeric):

Answer

Accepted Answer : 2.5

Which of the following threshold θ value of MP neuron implements AND Boolean function denoted by $f(\mathbf{x})$? Assume that the number of inputs x_i to the neuron is seven and the neuron does not have any inhibitory inputs.

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{i=0}^{6} x_i > \theta \\ 0, & otherwise \end{cases}$$

OPTIONS :

- ◯ 0
- ◯ -6
- ◯ 6
- ◯ 7
- ◯ -7

Your score : 0

Suppose that MP neuron takes in 7 Boolean inputs $(x_0, \cdots, x_6)$ and produces the Boolean output y. Assume none of the inputs is inhibitory. Select all true statements

OPTIONS :

- [ ] There are 2^{2^7} possible Boolean functions
- [ ] There are 2^7 possible Boolean functions
- [ ] The function $y = min(x_0, \cdots, x_6)$ is linearly separable
- [ ] The function $y = min(x_0, \cdots, x_6)$ is not linearly separable

Your score : 0

Suppose that we have a deep Feed Forward Fully Connected Neural Network. The network is observed to have a high variance. Then, which of the following techniques regularize the parameter of the network to reduce the high variance ?

OPTIONS :

- [ ] Adding L2 norm of weights to the loss function
- [ ] Adding a noise to the input samples
- [ ] Adding a noise to the output prediction
- [ ] Adding more samples to the dataset by augmenting existing samples usingsome augmentation techniques
- [ ] Dropping hidden layers in a neural network randomly during training

The logistic sigmoid function is defined as follows,

$$f(x) = \frac{1}{1 + exp(-(wx + b))}$$

The parameters are initialized to $w = 1$ $b = 1$. Suppose the loss in defined as

$$L = \frac{1}{2}(f(x) - y)^2$$

where y is the true value. Compute the gradient of b for the following sample $x = 0.2, y = 0$.

Answer (Numeric):

Answer

Accepted Answer : 0.024 to 0.029

Your score : 0

Consider a training set that contains 10 samples to train a neural network. Further, mini-batch GD algorithm has been chosen to update the parameters of the network with a batch size of 2. Suppose that we use an exponentially decaying learning rate scheme $\eta_t = 2\exp(-\frac{t}{4})$ and train the model for 2 epochs. What will be the value of the learning rate η_t at the end of the training? Assume, t starts from zero.
Enter the answer to 3 decimal points (that is, if your answer is -0.12145, then enter it as -0.121)

Answer (Numeric):

Answer

Accepted Answer : 0.19 to 0.23

Your score : 0

Consider an input image of size $256 \times 256 \times 3$, where 3 is the number of channels. Suppose we apply a set of convolution kernels on the input image that generates the output feature maps of size $248 \times 248 \times 32$. How many parameters (including bias) do the kernels have? Assume stride $(s = 1)$ and padding $p = 1$.

Answer (Numeric):

Answer

Accepted Answer : 11648

Your score : 0

Based on the above data, answer the given subquestions.

Consider a fully connected feed forward neural network with 3 hidden layers. The weight matrix W_1 connecting the input layer to the first hidden layer is of shape 20×150, similarly the shape of other weight matrices are as follows $W_2 : 150 \times 100$, $W_3 : 100 \times 10$, and the weight W_4 connecting the final hidden layer and the output layer is of shape 10×3. The network solves the multi-class classification problem by using the cross entropy loss function. Moreover, the labels are one hot encoded.

Your score : 0

How many neurons are there in the network. Every neuron in the network has bias associated with it? Note: A neuron is a computation unit that takes in some inputs and produce an output.

Answer (Numeric):

Answer

Accepted Answer : 263

Your score : 0

How many learnable parameters (including bias) does the network have? Assume dropout regularization is applied.

Answer (Numeric):

Answer

Accepted Answer : 19293

Your score : 0

For the given neural network configuration, we can replace the output layer with softmax activation by logistic sigmoid and still use cross entropy loss. The statement is

OPTIONS :

- ◯ True
- ◯ False

Your score : 0

The update rule for the ADAM (Adaptive Moments) optimization algorithm is given below,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla w_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla w_t)^2$$
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}}\hat{m}_t$$

Here, $0 \leq \beta_1 < 1$ and $0 \leq \beta_2 < 1$ and t starts from zero (that is, $t = 0, 1, 2, \cdots$). Both m_t and v_t are initialized to zero. However, the update rule uses the bias corrected version of m_t and v_t. Which of the following is the bias corrected version of m_t?

Helper:

$$m_t = (1-\beta_1)\sum_{\tau=0}^{t} \beta_1^{t-\tau}\nabla w_\tau$$

and assume that $E[\nabla w_\tau] = E[\nabla w] \quad \forall \tau$, if required.

OPTIONS :

- $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
- $\hat{m}_t = \frac{m_t}{1-\beta_1^{t+1}}$
- $\hat{m}_t = m_t$
- $\hat{m}_t = \frac{m_t}{1-t\beta_1^t}$

Consider a target variable $y = f(x) + \epsilon$ that is related to x, where ϵ is a random variable (noise) distributed normally. About 1000 points are sampled from the true function $f(x)$ to form a training set. Suppose that a prediction model $\hat{f}(x)$ is sufficiently complex in that $f(x) \subset \hat{f}(x)$. Then, the statement that the training error is lower bounded by σ^2 (that is, the variance of the noise) is

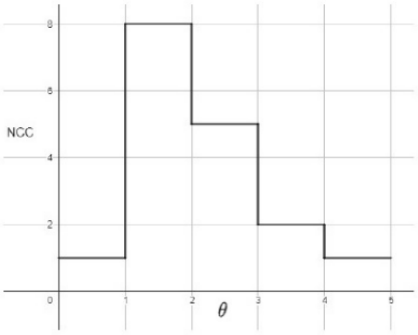
OPTIONS :

- True, due to the presence of noise in the target
- True, due to the high variance of the prediction model
- False, due to the high variance of the prediction model
- False, due to zero mean of the noise added to the target

Your score : 0

Suppose that we implement a three input Boolean function using the Mc-Culloch Pitts (MP) neuron. The graph below shows the Number of Correctly Classified (NCC) data points for various values of threshold **θ**. The threshold **θ** is incremented by 1 from 0 to 5. Assume that the neuron does not have any inhibitory input. This graph represents which of the following Boolean functions?

$$\hat{y} = \begin{cases} 1, & \text{if } \sum x_i \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

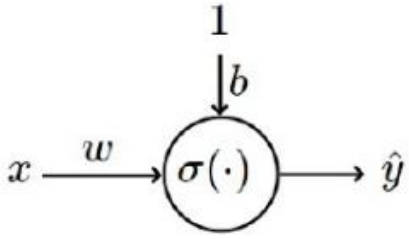


OPTIONS :

- ○ NOR
- ○ AND
- ○ OR
- ○ NAND
- ○ None of these

Your score : 0

Consider a sigmoid neuron shown below. The input to the neuron is a real number. Suppose that input $x = 10$, output $y = 1$ and the parameters, $w = 0.1, b = 0.1$. Update the parameters once using GD with η=1. Enter the sum of updated parameter values. The loss function is $L = \frac{1}{2}(y - \hat{y})^2$.



Answer (Numeric):

Answer

Accepted Answer : 0.68 to 0.72

Your score : 0

A CNN model takes an input image of size 100×100. The first convolution layer uses 30 kernels(filters) each of size 3×3, the output from the first convolution layer is passed as an input to the second convolution layer. The second convolution layer uses 20 kernels each of size 5×5. How many parameters (including bias) are there in the network? Assume stride=1,padding=0 if required.

Answer (Numeric):

Answer

Accepted Answer : 15320

Your score : 0

Based on the above data, answer the given subquestions.

A neural network contains an input layer $\mathbf{h_0} = \mathbf{x}$, five hidden layers $(\mathbf{h_1}, \mathbf{h_2}, \cdots, \mathbf{h_5})$ and an output layer $\mathbf{O}$. All the hidden layers use *Relu* activation and the output layer uses softmax activation.

Your score : 0

Suppose the input $\mathbf{x} \in \mathbb{R}^{900}$ and all the hidden layers contains 10 neurons each. The output layer contains 20 neurons. How many parameters are there in the entire network?

Answer (Numeric):

Answer

Accepted Answer : 9670

Your score : 0

Suppose that all the elements in the input vector are zero and the corresponding true label is also 0. Further, suppose that all the parameters are initialized to zero. What is the loss value if cross entropy loss is used? Use natural logarithm ln.

Answer (Numeric):

Answer

Accepted Answer : 2.9 to 3.1

Your score : 0

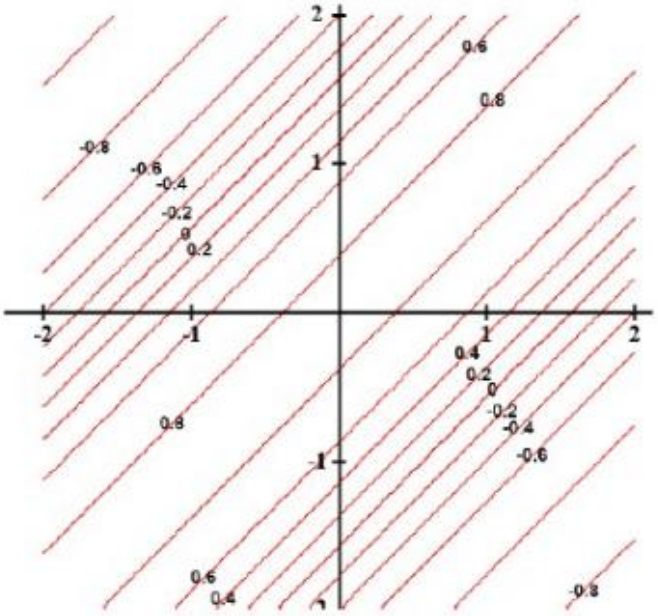
Assume that we use backpropagation to compute the gradient of the loss w.r.t. all the parameters. Then the statement that the gradients w.r.t. all the parameters are zero is

OPTIONS :

- ◯ TRUE
- ◯ FALSE
- ◯ Need more information to validate the statement

Your score : 0

The plot below shows contours of a function f(w, b). Choose the correct statements



OPTIONS :

- [ ] There might be two flat minima
- [ ] There might be two flat maxima
- [ ] There is one flat maxima
- [ ] There is one flat minima

Your score : 0

Suppose that a model produces zero training error without using any regularization technique. What happens if we use L2 regularization and retrain the model, in general?

OPTIONS :

- [ ] This might increase training error
- [ ] This might decrease test error
- [ ] Reduce the complexity of the model by driving less important weights to close to zero
- [ ] This might decrease training error

Your score : 0