**Quiz #1: (CORRECT ANSWERS)**

A model is evaluated on a test dataset containing 1000 images. For each image, the model predicts 5 labels (top-5 predictions). If the correct label for 50 images is not present in the top-5 predictions, what is the Top-5 error (%) of the model?

**Correct : 5%**

Deep Learning is

**Correct : AI, ML, Supervised & Unsupervised**

Training data: the images given for learning must come with labels

**Correct : False**

I must use 0 and 1 to denote the classes for binary classification.

**Correct : False**

Is the following statement True or False:

• Given training data 𝑥𝑖, 𝑦𝑖 : 1 ≤ 𝑖 ≤ 𝑛  
• Find 𝑦 = 𝑓(𝑥) using training data  
s.t. 𝑓 correct on test data.

statement: 𝑓 must be correct on all test data.

**Correct : False**

**Quiz #2 : (DON’T HAVE ANSWERS)**

**Select the correct statement/statements**

Classification tasks approximate a discrete function with discrete values or labels, such as indoor/outdoor or dog/cat.

Regression tasks approximate a discrete function and always have discrete values.

Regression tasks approximate a general function and can have continuous values.

**What is the primary purpose of data collection in a machine learning workflow?**

To minimize the empirical loss of the model.

To directly train the model without any preprocessing.

To choose the hypothesis class and loss function.

To extract meaningful features that represent the problem space.

**Why is *P(y|x)* considered a discriminative model?**Because it models the joint distribution P(x, y) to generate new data samples.

Because it only models P(x), ignoring the relationship between x and y.

Because it directly models the conditional probability P(y|x), focusing on decision boundaries between classes.

Because it uses the prior probability of P(y) to estimate the posterior probability.

**What is your verdict on the following statement:**

**The L2 loss function (squared error) is highly sensitive to outliers in the dataset.**

True

False

**Chose the best/closest answer:**

**Why do we use a step function in machine learning models like perceptrons?**

To introduce non-linearity into the model for better classification.

To ensure the output is a continuous value for regression tasks.

To normalize the input features for training the model.

To convert the linear model's output into discrete class labels (e.g., 0 or 1).

**The sigmoid function, denoted as \(\sigma(a)=\frac{1}{1+e^{-a}}\), is widely used in machine learning and statistics.**

**It has several important properties:**

* **It is *smooth* and differentiable, which makes it suitable for optimization algorithms.**
* **It maps input values to a range of [0,1], making it ideal for probabilistic interpretation in binary classification tasks.**
* **The sigmoid curve (S-shape) demonstrates a gradual transition from 0 to 1, providing a continuous and smooth decision boundary.**

**What is the output of the sigmoid function for a=2000,0,−5?**

0.7

0.5

0.0066

1

**Linear classification is called logistic regression because it models the conditional probability using the logistic function (sigmoid). The model predicts probabilities based on the linear combination of input features and uses the Maximum Likelihood Estimation (MLE) principle to optimize the weights . The sigmoid function ensures the output is bounded between 0 and 1, making it suitable for binary classification tasks. The loss function for logistic regression is derived from the negative log-likelihood, often referred to as the cross-entropy loss.**

**Having learned about logistic regression, what is your verdict about the following statement:**

**Logistic regression is a type of linear classification because it uses a linear model combined with the sigmoid function to predict probabilities.**

True

False

**The sigmoid function is symmetric about the origin(XY-plane), meaning that its value for a positive input a is related to its value for the corresponding negative input −a. Mathematically, this property is expressed as:**

**σ(−a)=1−σ(a)**

**This symmetry is a critical property used in optimization and probabilistic interpretations, as it ensures consistent behavior for positive and negative inputs.**

**True/False Question:**

**The sigmoid function satisfies the property σ(−a)=1−σ(a), demonstrating its symmetry about the origin.**

True

False

**The gradient of the sigmoid function, also known as its derivative, is given by:**

**σ′(a)=σ(a)⋅(1−σ(a))**

**This property is significant because it simplifies the computation of gradients during backpropagation in neural networks. The sigmoid's reliance on the exponential function ensures smoothness and a bounded gradient, which helps avoid extreme values during optimization.**

**True/False Question: Gradient  
The gradient of the sigmoid function is given by σ′(a)=σ(a)⋅(1−σ(a)), making it computationally efficient for backpropagation.**

True

False

**The exponential function exp⁡(−a) provides smooth and continuous transitions between 0 and 1. This smoothness ensures that the sigmoid function is differentiable everywhere, which is critical for gradient-based optimization methods like stochastic gradient descent.**

**True/False Question: Exponential Function Benefit  
The exponential function in the sigmoid ensures smoothness and differentiability, which are essential for gradient-based optimization methods.**

True

False

**Quiz #3 : (CORRECT ANSWERS)**

**In a linear classification model, the prediction is determined by the sign of(wTx, where y=+1 if wTx>0 and y=−1 if wTx<0).**

Correct : True

**In an online learning algorithm, data points are processed one by one, and the model is updated after receiving the correct classification for each data point**

Correct : True

**What happens in the Perceptron Algorithm when a misclassification occurs?**

Correct : The weight vector 𝑤 is updated by subtracting the input vector 𝑥 if the misclassification is on a negative example.

Correct : The weight vector 𝑤 is updated by adding the input vector 𝑥 if the misclassification is on a positive example.

Correct : The weight vector 𝑤 remains unchanged if the data is correctly classified.

**Gradient descent is called a stochastic process when it updates the weights w using a randomly selected data point from the dataset or misclassified set M.**

Correct : True

**Why is the loss function L(w) computed only over the set of misclassified data (M) instead of the entire dataset?**

Correct : Computing the loss over (M) reduces unnecessary computations for correctly classified data points.

Correct : Misclassified points are the only ones influencing the optimization process, as they indicate errors to be corrected.

Correct : Correctly classified data points contribute zero to the loss, so including them has no impact on the result.

**Quiz #4 : (DON’T HAVE ANSWERS)**

**Which of the following is true (or most appropriate) about Support Vector Machines (SVM)?**

SVM does not work well with high-dimensional data.

SVM maximizes the margin between different classes.

SVM always finds a unique solution for non-linearly separable data.

SVM is only applicable for binary classification problems.

**In Soft Margin SVM, what is the role of the slack variable ξi​?**

It is used only for linearly separable data.

It ensures that all points strictly follow the margin constraints.

It allows/tolerate some misclassification to improve model generalization.

 It is used to increase the margin beyond the optimal hyperplane.

**The Radial Basis Function (RBF) kernel is used in SVM to transform non-linearly separable data into a higher-dimensional space where it becomes linearly separable.**

True

False

**A diagram of a line and a line

AI-generated content may be incorrect.**

**In the given figure, the red line represents the decision boundary where y(x)=0.**

True

False

A diagram of a line with lines and numbers

AI-generated content may be incorrect.

**What is the role of the weight vector w in the SVM decision boundary?**

It determines the position of the hyperplane relative to the origin.

 It is perpendicular to the decision boundary and determines its orientation.

 It measures the distance between support vectors.

It only affects the bias term w0​.

**Quiz #5 : (CORRECT ANSWERS)**

**Which of the following statements about regularization in linear regression are true?**

**A) Ridge regression uses the L2​ norm of the weights as a penalty term.  
B) Lasso regression encourages sparsity by using the L1​ norm of the weights.  
C) A larger penalty term leads to a more complex model.  
D) Regularization helps balance the trade-off between bias and variance**

Correct : A, B, D

**Suppose you are performing 5-fold cross-validation on a dataset with 100 samples. During each fold, one group is used as the validation set, and the remaining groups are used as the training set.**

**What is the size of the training set and validation set in each fold?**

Training set: 80 samples, Validation set: 20 samples

**Which of the following statements about model selection is true?**

A validation dataset is used to approximate the test error.

**What is the primary trade-off when increasing the number of folds in cross-validation?**

Lower bias, higher variance, and increased computational cost

A graph with red and blue lines

AI-generated content may be incorrect.

**Increasing the polynomial degree M always reduces the error for both training and test data.**

Correct : False

**Quiz #6 : (CORRECT ANSWERS)**

**Neural networks require non-linearity in activation functions to model complex patterns; without it, they would behave like a single-layer linear model regardless of depth.**

Correct : True

**Deep learning differs from classic machine learning by automatically learning hierarchical features, rather than relying on hand-designed features.**

Correct : True

**Which of the following is a key characteristic of an activation function in a neural network?**

 Correct :  It introduces non-linearity to allow the network to learn complex patterns.

**Consider an image of size 32 × 32 with 3 color channels (RGB) as the input to a neural network. Each pixel has three values corresponding to the Red, Green, and Blue channels.**

**If this image is fully connected to the input layer of a neural network, how many nodes are required in the input layer?**

Correct : 3072

A math equations on a white background

AI-generated content may be incorrect.

Correct : 1

**Quiz #7 : (CORRECT ANSWERS)**

**Which of the following statements is true about Gradient-Based Learning in neural networks?**

Correct : Neural networks do not have a closed-form solution for optimization.



Correct : False

**During neural network inference, the backward pass takes place to update the model weights.**

Correct : False

**Why are partial derivatives used during backpropagation in neural networks?**

Correct : To compute how each weight affects the loss and update them accordingly.

**In the weight update equation below, what is the learning rate?**

A white background with black text

AI-generated content may be incorrect.

Correct : 1

**Quiz #8 : (CORRECT ANSWERS FOR 1, 2, 3, 5, STILL NEED #4)**

**What is the primary purpose of regularization in machine learning?**

Correct : To prevent overfitting and improve optimization

**A model with high bias is likely to underfit the data, while a model with high variance is likely to overfit the data.**

Correct : True

**Which of the following statements is true regarding L2 regularization in machine learning?**

Correct : L 2 ​ regularization corresponds to a normal (Gaussian) prior on the parameters.

Correct : L 2 ​ regularization helps reduce variance by penalizing large model weights.

**4. When applying L1 regularization, which loss function is typically used for regression and which for classification?**

Mean Absolute Error (MAE) for regression, Hinge Loss for classification

Mean Squared Error (MSE) for regression, Cross-Entropy Loss for classification

Hinge Loss for regression, Cross-Entropy Loss for classification

Mean Squared Error (MSE) for both regression and classification

(ORIGINALLY SELECTED FIRST TWO OPTIONS AND GOT 0 / 1 PTS)

**Which of the following statements correctly describe how Dropout works as a regularization technique in deep learning?**

Dropout applies a binary mask to input and hidden units, deciding which neurons to deactivate in each update step.

 Typical dropout probabilities are 0.2 for input layers and 0.5 for hidden layers.

Dropout randomly sets a subset of neurons to zero during training to prevent overfitting.

**Quiz #9 : (CORRECT ANSWERS)**

**Data augmentation involves synthetically generating data to expand the training set when the dataset is limited.**

Correct : True

**Which of the following statements about early stopping are correct? (Select all that apply)**

Correct : The validation loss is monitored to determine when to stop training.

Correct : If training continues after the validation loss flattens or increases, overfitting may occur.

Correct : Early stopping prevents overfitting by stopping training when validation loss stops improving.

**What is the typical dropout probability for input and hidden units?**

0.2 for input and 0.5 for hidden units

**In a convolutional neural network (CNN), the pooling layer is responsible for extracting features from the input data.**

False

**A 2D image of size M=10 and N=8 is convolved with a filter (kernel) of size m=3 and n=3, without padding and with a stride of 1.**

(8, 6)

**Quiz #10 : (CORRECT ANSWERS)**

A diagram of a diagram

AI-generated content may be incorrect.

Correct : 4

**Max pooling is an example of downsampling in convolutional neural networks (CNNs), reducing the spatial dimensions of the feature maps.**

Correct : True

**When flattening the feature maps before the fully connected layer in this CNN, how do we obtain 400 neurons? (Select all that apply)**

Correct : Each 5×5 feature map contains 25 values, and since we have 16 feature maps, flattening results in 400 neurons.

Correct : There are 16 feature maps, each of size 5×5, so the total number of values is 16×5×5=400.

Correct : Flattening means converting the 2D feature maps into a 1D vector.

**In the given LeNet architecture, the output layer currently has a dimension of 10. If the network were modified deal with 3 classes, what would be the new dimension of the output layer?**

Correct : 3

A graph with red and blue dots

AI-generated content may be incorrect.

**In the given PCA projection illustration, what does the projection of the blue data points onto the black line represent?**

Correct : The projection of the original data onto the first principal component, reducing dimensionality while retaining maximum variance.

**Principal Component Analysis (PCA) reduces data dimensionality and can help remove noise, but it might also lead to overfitting by ignoring information that could be important but is not present in the training set.**

Correct : True

**Which of the following statements about PCA, low-rank approximation, and eigen-decomposition/SVD is true for centered data?**

Correct : PCA, low-rank approximation, and eigen-decomposition/SVD are equivalent for centered data, meaning they provide the same transformation and decomposition results.