// // // // 1

1. What advantage does using a bias value bring in the context of the artificial neuron?

A. It significantly improves convergence time

*B. It prevents the neuron hyperplanes from being forced to go through the origin

C. It significantly helps in the context of imbalanced data sets by providing a bias towards the misrepresented class

D. It does not bring any advantage

2. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?

A. Cross Entropy

B. MSE

C. L2 Loss

*D. L1 Loss

3. The training data set contains the following examples [(3, PASS), (2, PASS), (2, PASS), (4, PASS), (0, FAIL), (1, FAIL), (3, FAIL), (1, FAIL)], the first component being the number of hours of study and the second denoting whether the student passed the exam. What is the probability of passing the exam with 2 hours of study - P(PASS|2)?

A. 25%

B. 50%

C. 75%

*D. 100%

4. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size; the other numbers represent the number of neurons in each layer)?

*A. 6x2

B. 6x1

C. 4x6

D. 2x1

5. What is the output of the perceptron if input= [2.4, 3.0], weights= [-0.5, 0.2], bias=1.0 (activation function - sign)?

*A. 1

B. 2.2

C. 0

D. -1

6. What is the MSE for the following predicted labels y_pred = [0.1, 0.4, 0.7, 0.3] and truth labels=[1, 0, 1, 0]?

A. 0.3315

B. 0.1430

C. 0.0715

*D. 0.2875

7. What is the difference between using an L1 loss and an L2 loss?

A. Using the L1 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

(x)B. The L2 loss generally favors having smaller errors instead of a having fewer but greater errors while the L1 loss does not differentiate between these cases.

C. The L1 loss generally favors having smaller errors instead of a having fewer but greater errors while the L2 loss does not differentiate between these cases.

D. Using the L2 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

8. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?

A. [0.0, 0.5, 1.0]

B. [10, 20, 30]

*C. [0.16, 0.33, 0.5]

D. [1, 2, 3]

9. What is the f1-score of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 1, 1] and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

A. 0.7

(x)B. 0.5

C. 0.6

D. 0.4

10. Which machine learning model can achieve the best performance in the context of an audio classification problem?

*A. Depends on problem details and should be determined by means of validation

B. An SVM classifer

C. A Neural Network with five layers

D. A Neural Network with two layers

// /// // 2

//1

1. What is the difference between using an L1 loss and an L2 loss?

A. Using the L1 loss you can avoid getting stuck in a local minima when using stochastic gradient descent in the case of neural networks.

(x)B. The L2 loss generally favors having smaller errors instead of a having fewer but greater errors while the L1 loss does not differentiate between these cases.

C. The L1 loss generally favors having smaller errors instead of a having fewer but greater errors while the L2 loss does not differentiate between these cases.

D. Using the L2 loss you can avoid getting stuck in a local minima when using stochastic gradient descent in the case of neural networks.

2. What advantage does using a bias value bring in the context of the artificial neuron?

A. It significantly improves convergence time

B. It prevents the neuron hyperplanes from being forced to go through the origin

C. It significantly helps in the context of imbalanced data sets by provinding a bias towards the misrepresented class

(x)D. It does not bring any advantage

3. How many neurons (including bias) has a neural network with the following configuration 6-6-8-1

(the first number is the input size, the other numbers represent the amount of neurons in each layer)?

6*6 + 6 + 6*8 + 8 + 8*1 + 1

(x)A. 107

B. 102

C. 92

D. 113

4. What is the purpose of the learning rate in the context of neural networks?

A. It specifies the rate for mini-batch selection, thus aiding the learning process

(x)B. It prevents divergence and regulates convergence speed

C. It acts as a much needed initialization for the weights

D. It specifies the maximum number of neurons for the output layer

5. What is the accuracy of the classifier if the ground-truth labels are y = [5, 7, 6, 1, 0, 6, 3, 2]

and the predicted labels are y_hat = [5, 6, 6, 1, 0, 2, 2, 1]?

(x)A. 0.5

B. 0.8

C. 0.55

D. 0.52

6. Which of the following is equivalent to a single artificial neuron without activation?

A. A KNN classifier with 3 neighbors

(x)B. A neural network with no activations

C. An SVM with polynomial kernel

D. A Naive Bayes classifier

7. Which of the following neuron activation is the result of the sigmoid activation function?

(x)A. [0.1, 0.11, 0.99]

B. [-0.1, 0.11, 0.2]

C. [0.1, 1.2, 0.2]

D. [1.01, 0.11, -0.2]

8. Calculate the cost for the Lasso Regression having weights=[3, 2], alpha=0.1, y_true=[10, 1, 9, 4], y_pred=[9, 3, 6, 7].

A. 36.50

(X)B. 23.50

C. 23.00

D. 0.10

9. What is the recall of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 0, 1]
and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

(x)A. 0.33

B. 0.45

C. 0.23

D. 0.99

10. Why does feature normalization help in the context of the KNN classifier?

A. Normalization does not help

(x)B. Normalization brings different features on the same scale so that the L1/L2 distances between samples are not affected by the different scale of different features

C. Normalization boosts performance when the data set is imbalanced

D. Normalization helps reduce noise and ensures a slower convergence

//2

3. Which of the following neuron activation is the result of the sigmoid activation function?

A. [0.1, 1.2, 0.2]

B. [1.01, 0.11, -0.2]

(x)C. [0.1, 0.11, 0.99]

D. [-0.1, 0.11, 0.2]

4.What is the output of the perceptron if input=[2.4, 3.0], weights=[-0.5, 0.2], bias=1.0 (activation function - sign)?

A. 0

(x)B. 1

C. -1

D. 2.2

4. What is the value of the Mean Absolute Error function if the ground-truth labels are y = [6, 8, -9, 5] and the predicted labels are y_hat = [6.5, 7.2, 1, 7]?

A. 3.5

B. 13.3

C. 13.5

(x)D. 3.325

3. How many neurons (including bias) has a neural network with the following configuration 6-6-8-1 (the first number is the input size, the other numbers

represent the amount of neurons in each layer)?

(x)A. 107

B. 102

C. 92

D. 113

9. Which of the following is a linear classifier?

A. A 3-NN classifier

(x)B. A neuron with no activation

C. A two layer neural network with ReLU activations

D. An SVM with polynomial kernel

10. What is the resulting data after applying L2 normalization to this vector [10, 20, 30]?

A. [1, 2, 3]

B. [0.16, 0.33, 0.5]

(x)C. [0.26, 0.53, 0.80]

D. [0.0, 0.5, 1.0]

4. What is the value of the Mean Absolute Error function if the ground-truth labels are y = [6, 8, -9, 5] and the predicted labels are y_hat = [6.5, 7.2, 1, 7]?

A. 3.5

B. 13.3

C. 13.5

(x)D. 3.325

//3

1. If the data is split into 9 classes, and we want to train a SVM for classification.

How many binary classifiers will be trained in the one-vs-one approach?

A. 18

(x)B. 36

C. 9

D. 81

2. Which of the following neuron activation is the result of the tanh activation function?

A. [1.01, 0.11, 0.2]

B. [-1.2, 0.11, 1.2]

C. [0.9, 0.11, -1.1]

(x)D. [0.99, 0.05, 0.99]

3. Calculate the cost for the Ridge Regression having weights=[3, 2], alpha=0.1, y_true=[10, 1, 9, 4], y_pred=[9, 3, 6, 7].

A. 36.23

(x)B. 23.36

C. 23.00

D. 0.10

4. Which of the following is equivalent to a single artificial neuron without activation?

A. A KNN classifier with 3 neighbors

B. A Naive Bayes classifier

(x)C. A neural network with no activations

D. An SVM with polynomial kernel

5. After training for 5 epochs, we have the following training losses for each epoch [0.60, 0.48, 0.30, 0.28, 0.26], and the following validation losses for each epoch [0.55, 0.43, 0.27, 0.27, 0.25]. Is the model overfitted, underfitted, both, or neither?

(x)A. Neither

B. Overffiting

C. Both

D. Underfitting

6. What is the output of neural network with 3 hidden units and 1 output unit having ReLU activations for the input x = [1, -2],

if the weights are W1 = [-0.5, 3, -2; 2, -1, 0], B1 = [0, 1, -1], W2 = [-1; -1; 2], B2 = [2]?

A. 1

B. 4.5

(x)C. 0

D. 8

7. What is the value of PReLU(x) - parametric ReLU, where alpha=0.1 and x=-0.2?

A. -1

B. 0

C. 0.002

(x)D. -0.02

8. If the current weights of a perceptron are X = [0.2, 0.4], their gradients are G = [-2.4, -1.2], and the learning rate is 0.1.

What are the weights after the weights update operation?

// X - G * learning_rate

A. [0.52, 0.44]

(x)B. [0.44, 0.52]

C. [0.44, 0.44]

D. [0.52, 0.52]

9. Which of the following is a linear classifier?

A. A two layer neural network with ReLU activations

B. A 3-NN classifier

(X)C. A neuron with no activation

D. An SVM with polynomial kernel

10. What is the output of SVM classifier for the input X = [0.1, -2, -5], if the weights are W = [-2, -1.2, -3] and the bias is b = 0.5?

// X * W + b

A. 2

B. 0

(x)C. 1

D. -1

//4

1. Which of the following is a technique for using an SVM as a multi-class classifier?

A. Split group classification

B. One versus all

C. All versus all

D. N-way split

(B)

2. What is the label of the test example t = [2, 3, 5] if you apply the k-nearest neighbors classifier with k = 1 and metric = L1 (Manhattan distance) given the training data X = [[1, 4, 1], [2, 4, 7], [2, 30, 5], [0, 1, 0]], Y = [1, 3, 2, 2]?

A. 2 B. 0 C. 1 D. 3

(D)

3. If we have the following probabilities for events P(A)=0.5 P(B)=0.9 P(A|B)=0.3, what is the value of P(B|A)?

A. 0.27 B. 0.75 C. 0.63 D. 0.54

(D)

4. Which of the following is a linear classifier?

A. A two layer neural network with ReLU activations

B. A 3-NN classifier

C. An SVM with polynomial kernel

D. A neuron with no activation

(D)

5. How many learned parameters (weights + biases) will a network with input size = 2, hidden layer size = 5, output layer size = 1, have?

A. 10 B. 8 C. 13 D. 21

(combinari de 9 luate cate 2) - 36

6. If the data is split into 9 classes, and we want to train a SVM for classification. How many binary classifiers will be trained in the one-vs-one approach?

A. 18 B. 9 C. 36 D. 81

(B)

7. In which scenario is measuring the accuracy of the model not enough to evaluate the model properly?

A. When the dataset is imbalanced

B. When the data set is balanced but the training set and test set come from different sources

C. When there are 3 classes in the dataset

D. When the data set is made out of audio samples

(A)

8. What is the recall of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 0, 1]

and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

(x)A. 0.33 B. 0.23 C. 0.99 D. 0.45

9. Given the following vocabulary {0 - dogs, 1 - cats, 2 - candies, 3 - likes, 4 - she, 5 - he}. What is the bag of words (BOW) representation of the sentence "she likes dogs and horses."?

A. [1, 0, 0, 1, 1, 0, 1, 1]

B. [1, 0, 0, 1, 1, 0]

C. [1, 0, 1, 1, 1, 0]

D. [2, 0, 0, 1, 1, 0]

(C)

10. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?

A. L1 Loss B. Cross Entropy C. MSE D. L2 Loss

(B)

11. What advantage does using a bias value bring in the context of the artificial neuron?

A. It significantly improves convergence time

B. It prevents the neuron hyperplanes from being forced to go through the origin

C. It significantly helps in the context of imbalanced data sets by provinding a bias towards the misrepresented class

D. It does not bring any advantage

(D)

12. The training data set contains the following examples [(3, PASS), (2, PASS), (2, PASS), (4, PASS), (0, FAIL), (1, FAIL), (3, FAIL), (1, FAIL)], the first component being the number of hours of study and the second denoting wether the student passed the exam. What is the probability of passing the exam with 2 hours of study - P(PASS|2)?

A. 25% B. 50% C. 75% D. 100%

(D)

13. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size, the other numbers represent the amount of neurons in each layer)?

A. 6x2 B. 6x1 C. 4x6 D. 2x1

(A)

14. What is the output of the perceptron if input=[2.4, 3.0], weights=[-0.5, 0.2], bias=1.0 (activation function - sign)?

A. 1 B. 2.2 C. 0 D. -1

(D)

15. What is the MSE for the following predicted labels y_pred = [0.1, 0.4, 0.7, 0.3] and truth labels=[1, 0, 1, 0]?

A. 0.3315 B. 0.1430 C. 0.0715 D. 0.2875

(B)

16. What is the difference between using an L1 loss and an L2 loss?

A. Using the L1 loss you can avoid getting stuck in a local minima when using stochastic gradient descent in the case of neural networks.

(x)B. The L2 loss generally favors having smaller errors instead of a having fewer but greater errors while the L1 loss does not differentiate between these cases.

C. The L1 loss generally favors having smaller errors instead of a having fewer but greater errors while the L2 loss does not differentiate between these cases.

D. Using the L2 loss you can avoid getting stuck in a local minima when using stochastic gradient descent in the case of neural networks.

(C)

17. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?

A. [0.0, 0.5, 1.0]

B. [10, 20, 30]

C. [0.16, 0.33, 0.5]

D. [1, 2, 3]

(C)

18. What is the f1-score of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 1, 1]

and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

A. 0.7 (x)B. 0.5 C. 0.6 D. 0.4

19. Which machine learning model can achieve the best performance in the context of an audio classification problem?

A. Depends on problem details and should be determined by means of validation

B. An SVM classifer

C. A Neural Network with five layers

D. A Neural Network with two layers

(A)

20. How many neurons should the hidden layer of a network with a single hidden layer and an output layer have in the context of a classification problem with 25 classes have?

A. Depends on the problem and should be determined by means of validation

B. 3

C. 10

D. 25

(A)

21. Which of the following neuron activation is the result of the tanh activation function?

A. [0.99, 0.05, 0.99]

B. [-1.2, 0.11, 1.2]

C. [1.01, 0.11, 0.2]

D. [0.9, 0.11, -1.1]

(A)

22. What is the value of the loss function of a Rigde regression model if the predicted values y_hat are [-2, -3, -1], the ground-truth values are [-2, -3, -2.5], the wights are W = [1, 0], bias = 5 and alpha = 0.1?

A. 0.85 B. 0.75 C. 0.22 D. 0.95

(A)

23. What is the label of the test example t = [5, 3, 8] if you apply the k-nearest neighbors classifier with k = 3 and metric = L1 (Manhattan distance) given the training data X = [[1, 4, 2], [5, 4, 8], [2, 6, 5], [1, 1, 1], [2, 9, 6]], Y = [2, 3, 3, 1, 2]?

A. 2 B. 3 C. 1 D. 0

(C)

24. Can an SVM be used to achieve 100% training accuracy on the following 2D data set [([0, 1], 1), ([1, 0], 1), ([0, 0], 1), ([-2, 2], 0), ([2, 2], 0), ([-2, -2], 0), ([2, -2], 0)]?

A. Yes, but only if the data is normalized

B. No, because the data is not linearly separable

C. Yes, by using the kernel trick

D. No, because the dataset is imbalanced

[0,1] Sum = 1 (A)

25. Which of the following neuron activation is the result of the softmax activation function?

A. [0.6, 0.2, 0.2]

B. [0.5, 0.2, 0.2]

C. [0.6, 0.2, 0.3]

D. [0.6, -0.2, 0.2]

(A)

26. How many neighbors should you consider in order to obtain the best result from a KNN classifier on the test set?

A. 1

B. 3

C. 7

D. It depends on the problem and should be determined by means of validation

(A)

27. What is the label of the test example t = [1, 2, 6] if you apply the k-nearest neighbors regressor with k = 3 and metric = L1 (Manhattan distance) given the training data X = [[1, 4, 2], [5, 4, 8], [2, 6, 5], [1, 1, 1], [2, 9, 6]], Y = [0.3, 0.6, 0.9, 0.6, 0.5]?

A. 0.6 B. 0.55 C. 0.65 D. 0.1

(F-N)/STRIDE + 1

(B)

28. What will be the shape of the activation maps if we apply a 5x5 convolutional filter with stride=1 and no padding to a 16x16 image?

A. 14x14 B. 12x12 C. 18x18 D. 16x16

(B)

29. Suppose our model has the following metrics TP (true positives)=30, FP (false positives)=10, FN (false negatives)=30. What is the precision (P) and recall (R)?

A. P=50%, R=75%

B. P=75%, R=50%

C. P=10%, R=50%

D. P=30%, R=75%

(B)

30. What type of metric can achieve 100% training accuracy on the following 2D data set [([1, 1], 1), ([5, 5], 1), ([10, 10], 1), ([5, 4], 0), ([6, 5], 0), ([6, 4], 0)] when considering a 1-NN classifier?

A. Cosine

B. None of the answers

C. L2

D. L1

(B)

31. What is the value of the Mean Absolute Error function if the ground-truth labels are y = [6, 8, -9, 5] and the predicted labels are y_hat = [6.5, 7.2, 1, 7]?

A. 13.3 B. 3.325 C. 3.5 D. 13.5

(B)

32. What is the output of neuron having sign activation for the input x = [1, -1], if the weights are W = [-1, 2], B = [1]?

A. 1 B. -1 C. 2 D. -2

(B)

33. What is the label of the test example x = [1, -1] with a 1-NN model based on the Euclidean distance having the training set S = {([2, -1], 1), ([1, 1], 2), ([-1, -1], 3)}?

A. 4 B. 3 C. 2S D. 1

(D)

34. What is the resulting data after applying min-max scaling to this data [[0.1, 0.4], [0.2, 0.5], [0.3, 0.6]] (3 examples, 2 features)?

A. [[0.0, 0.5], [0.25, 0.75], [0.5, 1.0]]

B. [[0.1, 0.4], [0.2, 0.5], [0.3, 0.6]]

C. [[0.0, 0.4], [0.25, 0.5], [0.5, 0.6]]

D. [[0.0, 0.0], [0.5, 0.5], [1.0, 1.0]]

(B)

35. Which classifier can achieve the best performance on a e-mail spam classification task?

A. A Neural Network with three layers

B. Depends on problem details and should be determined by means of validation

C. An SVM with RBF kernel

D. An SVM with linear kernel

(A)

36. What will be the shape of the activation maps if we apply a 2x2 max pooling with stride=2 to a 32x32 activation map?

(x)A. 16x16

B. 32x32

C. 14x14

D. 28x28

//5

1. What advantage does using a bias value bring in the context of the artificial neuron?

A. It significantly improves convergence time

*B. It prevents the neuron hyperplanes from being forced to go through the origin

C. It significantly helps in the context of imbalanced data sets by providing a bias towards the misrepresented class

D. It does not bring any advantage

2. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?

A. Cross Entropy

B. MSE

C. L2 Loss

*D. L1 Loss

3. The training data set contains the following examples [(3, PASS), (2, PASS), (2, PASS), (4, PASS), (0, FAIL), (1, FAIL), (3, FAIL), (1, FAIL)], the first component being the number of hours of study and the second denoting whether the student passed the exam. What is the probability of passing the exam with 2 hours of study - P(PASS|2)?

A. 25%

B. 50%

C. 75%

*D. 100%

4. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size; the other numbers represent the number of neurons in each layer)?

*A. 6x2

B. 6x1

C. 4x6

D. 2x1

5. What is the output of the perceptron if input= [2.4, 3.0], weights= [-0.5, 0.2], bias=1.0 (activation function - sign)?

*A. 1

B. 2.2

C. 0

D. -1

6. What is the MSE for the following predicted labels y_pred = [0.1, 0.4, 0.7, 0.3] and truth labels=[1, 0, 1, 0]?

A. 0.3315

B. 0.1430

C. 0.0715

*D. 0.2875

7. What is the difference between using an L1 loss and an L2 loss?

A. Using the L1 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

(x)B. The L2 loss generally favors having smaller errors instead of a having fewer but greater errors while the L1 loss does not differentiate between these cases.

C. The L1 loss generally favors having smaller errors instead of a having fewer but greater errors while the L2 loss does not differentiate between these cases.

D. Using the L2 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

8. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?

A. [0.0, 0.5, 1.0]

B. [10, 20, 30]

*C. [0.16, 0.33, 0.5]

D. [1, 2, 3]

9. What is the f1-score of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 1, 1] and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

A. 0.7

(x)B. 0.5

C. 0.6

D. 0.4

10. Which machine learning model can achieve the best performance in the context of an audio classification problem?

*A. Depends on problem details and should be determined by means of validation

B. An SVM classifer

C. A Neural Network with five layers

D. A Neural Network with two layers

// /// // 3

1. Cati neuroni ar trebui sa aiba stratul de iesire al unei retele neurale cu un singur strat ascuns si un strat de iesire in contextul unei probleme de clasificare cu 7 clase?

A. 14

B. 24

(x)C. 7

D. Depinde de problema si ar trebui determinat prin validare

2. Avand urmatoarele date de antrenare: X_train = ((1,3),(1,2),(3,2),(1,-1),(-1,1),(-1,-))

Y_train = (1,1,1,-1,-1,-1) si datele de test:

X_test = ((1,1), (2,-1), (-1,3), (1,-1)),

Y_test = (1,1,-1,-1), care este acuratetea metodei celui mai apropiat vecin pe multimea de antrenare?

A. 0.75

B. 0.5

(x)C. 1

D. 0.25

3. Ce avantaj aduce folosirea bias-ului in contextului neuronului artificial?

(x)A. Previne constragerea de a trece prin origine a hiperplanelor determinate de neuroni

B. Creste semnificativ viteza de convergenta

C. Ajuta semnificativ in contextul in care datele sunt debalansate inclinand (bias) decizia catre clasa defavorizata

D. Nu aduce niciun avantaj

4. Fie urmatoarele probabilitati pentru evenimentele A, B, P(A)=0.3 P(B)=0.5 P(A|B)=0.33, care este valoarea P(B|A)?

A. 0.75

B. 0.65

(x)C. 0.55

D. 0.45

5. Fiind date etichitele (1, 3, -1, 6, 4) si predictiile (1.5, 3, 0.5, 5, 4),

care este media patratelor erorilor pentru Regresia Lineara?

A. 0.5

B. 1

(x)C. 0.7

D. 0.3

6. Care dintre urmatoarele activari ale neuronilor este rezultatul functiei de activare softmax?

A. [0.51, 0.5, 0.0]

B. [0.54, 0.5, 0.0]

(x)C. [0.5, 0.5, 0.0]

D. [0.5, 0.3, 0.21]

7. Care dintre urmatoarele modele este echivalent cu un singur neuron artificial fara activare?

A. Un model 5-NN

(x)B. O retea neurala fara activari

C. Un clasificator Naive Bayes

D. O retea neurala cu activari ReLU

8. Care este outputul unei retele feed-forward cu 3 perceptroni grupati pe un strat ascuns,

avand urmatorii parametri: X = [1,2], W1 = [2, 3, 5; -1, 2, -3], B1 = [-1, 2, 3], W2 = [2, -2, -1], B2 = [5].

Functia de activare de pe stratul ascuns si cel de iesire este ReLU.

A. -15

B. 7

C. 21

(x)D. 0

9. Cati vecini ar trebui considerati pentru a obtine cel mai bun rezultat posibil din partea unui

clasificator bazat pe cei mai apropiati vecini?

(x)A. Depinde de datele problemei si ar trebui determinat prin validare

B. 3

C. 8

D. 5

10. Cati parametri antrenati (ponderi + bias) vom avea intr-o retea complet conectata cu dimesniunea de

intrare = 4, dimensiunea stratului ascuns = 8, dimensiunea stratului de iesire = 1?

(x)A. 49

B. 32

C. 8

D. 33

// // // 4

1. Care este recall-ul unui clasificator daca etichetele corecte sunt y = [1, 1, 1, 1, 0, 1, 0, 1] si cele prezise sunt y_hat = [1, 0, 0, 1, 0, 1, 1, 1]?

(x)A. 0.66

B. 0.44

C. 0.2

D. 0.1

2. Cati vecini ar trebui considerati pentru a obtine cel mai bun rezultat posibil din partea unui clasificator bazat pe cei mai apropiati vecini?

A. 3

B. 5

(x)C. Depinde de datele problemei si ar trebui determinat prin validare

D. 1

3. In urma antrenarii unui model, avem urmatoarele metrici TP (true positives)=120, FP (false positives)=12, FN (false negatives)=7. Care sunt valorile pentru precision (P) si recall (R)?

(x)A. P=90.09%, R=94.48%

B. P=72.32%, R=82.44%

C. P=80.08%, R=84.48%

D. P=94.48%, R=90.09%

4. Fiind date etichitele (1, 3, 1, 6, 4) si predictiile (1.5, 3, 0.5, 5, 4), care este media patratelor erorilor pentru Regresia Lineara?

A. 0.7

(x)B. 0.3

C. 0.5

D. 1

5. Poate obtine un model SVM 100% acuratete pe datele de antrenare pentru urmatorul set de puncte 2D: [([0, 1], 1), ([1, 0], 1), ([0, 0], 1), ([-2, 2], 0), ([2, 2], 0), ([-2, -2], 0), ([2, -2], 0)]?

(x)A. Da, folosind kernel trick.

B. Da, da doar daca datele vor fi normalizate.

C. Nu, pentru ca datele nu sunt liniar separabile.

D. Nu, pentru ca setul de date este debalansat.

6. Fiind data urmatoarea multime de antrenare, reprezentand inaltimea (in cm) a unei persoane si eticheta corespunzatoare (F-femeie, M-barbat): [(160, F), (175, M), (155, F), (172, F), (187, F), (180, M), (177, M), (190, M)], impartiti valorile continue (intaltimea) in urmatoarele intervale: (150-160, 161-170, 171-180, 181-190). Care este probabilitatea ca o persoana de 178cm inaltime sa fie femeie? Folositi Regula lui Bayes.

A. 1/2

(x)B. 1/4

C. 1/3

D. 3/4

7. Daca datele sunt impartite in 5 clase si folosim un SVM pentru antrenare, cate clasificatoare binare vor fi antrenate prin metoda one-vs-one?

A. 5

B. 25

(x)C. 10

D. 20

8. Care dintre punctele urmatoare, alaturi de etichetele corespunzatoare, pot fi discriminate corect de un perceptron?

(x)A. X = ((1,1),(1,2),(1,3),(-1,1),(-1,-1),(-2,-1)) Y = (1,1,1,1,-1,-1)

B. X = ((-1,1),(-1,2),(1,3),(1,-1),(-1,-1),(-2,-1)) Y = (1,1,-1,1,-1,-1)

C. X = ((1,1),(1,3),(2,3),(2,1),(-1,2),(3,2)) Y = (1,1,1,-1,-1,-1)

D. X = ((1,1),(1,2),(2,2),(1,-1),(-1,-1),(-2,-1)) Y = (1,1,-1,1,-1,-1)

9. Care dintre urmatoarele aduce valorile din setul de date in intervalul [0, 1]? X reprezinta setul de date, X_i represinta setul de trasaturi i al tuturor exemplelor din setul de date iar x reprezinta un exemplu din setul de date.

A. Standard Normalization (X_i / std(X_i) - mean(X_i) for each feature i)

(x)B. Min-Max Scaling ((X_i - min(X_i)) / (max(X_i) - min(X_i)) for each feature i)

C. L2 Normalization (x / sqrt(sum x_i^2) for each sample x)

D. L1 Normalization (x / sum |x_i| for each sample x)

10. Ce tip de metrica poate obtine 100% acuratete pe datele de antrenare pentru urmatorul set de puncte 2D [([1, 1], 1), ([5, 5], 1), ([10, 10], 1), ([5, 4], 0), ([6, 5], 0), ([6, 4], 0)] considerand un clasificator KNN cu un singur vecin?

A. L2

B. Niciunul dintre raspunsuri

C. Cosinus

D. L1

8. Calculate the cost for the Lasso Regression having weights=[3, 2], alpha=0.1, y_true=[10, 1, 9, 4], y_pred=[9, 3, 6, 7].

A. 36.50

(X)B. 23.50

C. 23.00

D. 0.10

4. What is the purpose of the learning rate in the context of neural networks?

A. It specifies the rate for mini-batch selection, thus aiding the learning process

(x)B. It prevents divergence and regulates convergence speed

C. It acts as a much needed initialization for the weights

D. It specifies the maximum number of neurons for the output layer

Exerciții Rezolvate ML

Test 48

1. What is the label of the test example t = [2, 3, 5] if you apply the k-nearest neighbors classifier with k = 1 and metric = L1 (Manhattan distance) given the training data

X = [[1, 4, 1], [2, 4, 7], [2, 30, 5], [0, 1, 0]],
Y = [1, 3, 2, 2]?

A. 2
B. 3
C. 0
D. 1

The L1 (Manhattan) distances are:

- [1, 4, 1] - [2, 3, 5] = |1 - 2| + |4 - 3| + |1 - 5| = 1 + 1 + 4 = 6
- [2, 4, 7] - [2, 3, 5] = |2 - 2| + |4 - 3| + |7 - 5| = 0 + 1 + 2 = 3
- [2, 30, 5] - [2, 3, 5] = |2 - 2| + |30 - 3| + |5 - 5| = 0 + 27 + 0 = 27
- [0, 1, 0] - [2, 3, 5] = 2 + 2 + 5 = 9

We need to pick the 1-nearest neighbor(s). That means the one neighbor with **minimum distance**. This is the **second** training example, which has **label 3**.

2. Given the following vocabulary {0 - dogs, 1 - cats, 2 - candies, 3 - likes, 4 - she, 5 - he}. What is the bag of words (BOW) representation of the sentence "she likes dogs and horses."?
A. [1, 0, 0, 1, 1, 0, 1, 1]
B. [1, 0, 1, 1, 1, 0]
C. [1, 0, 0, 1, 1, 0]
D. [2, 0, 0, 1, 1, 0]

The set of words in the sentence is { she, likes, dogs, and, horses }.
If we intersect this with the vocabulary, we have { she, likes, dogs }.
This means { 4, 3, 0 } so we need a vector where indices **0**, **3** and **4** are set to 1.
This means

v[0] = 1, v[3] = 1, v[4] = 1

which is

v = [1, 0, 0, 1, 1, 0]

3. What is the resulting data after applying min-max scaling to this data [[0.1, 0.4], [0.2, 0.5], [0.3, 0.6]] (3 examples, 2 features)?
A. [[0.0, 0.5], [0.25, 0.75], [0.5, 1.0]]
B. [[0.1, 0.4], [0.2, 0.5], [0.3, 0.6]]
C. [[0.0, 0.4], [0.25, 0.5], [0.5, 0.6]]
D. [[0.0, 0.0], [0.5, 0.5], [1.0, 1.0]]

Rescaling using min-max: https://en.wikipedia.org/wiki/Feature_scaling#Rescaling_(min-max_normalization)

Values on X axis: 0.1, 0.2, 0.3
Values on Y axis: 0.4, 0.5, 0.6

Minimum values on X, Y: [0.1, 0.4]
Maximum values on X, Y: [0.3, 0.6]
Difference between max and min values on each axis: [0.3 - 0.1, 0.6 - 0.4] = [0.2, 0.2]

Subtract minimum on each axis:
[[0.1 - 0.1, 0.4 - 0.4], [0.2 - 0.1, 0.5 - 0.4], [0.3 - 0.1, 0.6 - 0.4]]
= [[0, 0], [0.1, 0.1], [0.2, 0.2]]

Divide each axis by (max - min):
[[0 / 0.2, 0 / 0.2], [0.1 / 0.2, 0.1 / 0.2], [0.2 / 0.2, 0.2 / 0.2]
= [[0, 0], [0.5, 0.5], [1, 1]]

4. How many neurons should the hidden layer of a network with a single hidden layer and an output layer have in the context of a classification problem with 25 classes have?
A. 10
B. 25
C. 3
D. Depends on the problem and should be determined by means of validation

- Number of neurons in the **input layer** is the **number of features** in the input.
- Number of neurons in the **output layer** is the **number of classes** in the output.
- Hidden layers are hyperparameters that have to be determined by validation, they don't have a formula.

5. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size, the other numbers represent the amount of neurons in each layer)?
A. 6x1
B. 2x1
C. 4x6
D. 6x2

The second layer consists of 2 neurons, while the previous one has 6. Presuming they are fully connected, the weight dimension of that layer is 6x2.

6. Which classifier can achieve the best performance on a e-mail spam classification task?
A. A Neural Network with three layers
B. Depends on problem details and should be determined by means of validation
C. An SVM with RBF kernel
D. An SVM with linear kernel

We're not given enough information about the problem to pick a classifier.

7. Which of the following is a linear classifier?
A. A neuron with no activation
B. A 3-NN classifier
C. An SVM with polynomial kernel
D. A two layer neural network with ReLU activations

A neuron computes **f(Weight * input + bias)**, where f is the activation function.
With no activation function, this becomes a linear term: **Weight * input + bias**

8. What is the recall of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 0, 1] and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?
A. 0.23
B. 0.33
C. 0.99
D. 0.45

Formula:

$$\text{Recall} = \frac{tp}{tp + fn}$$

True positives are those with 1 in y and 1 in y_hat: 1 examples
False negatives are those with 1 in y and 0 in y_hat: 2 examples

Recall = 1/(1 + 2) = ⅓ = 0.33

9. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?
A. MSE
B. L2 Loss
C. Cross Entropy
D. L1 Loss

L1 loss uses absolute value function, which is not differentiable in 0, therefore cannot be used for gradient descent (at least theoretically).

10. What will be the shape of the activation maps if we apply a 2x2 max pooling with stride=2 to a 32x32 activation map?
A. 16x16
B. 32x32
C. 14x14
D. 28x28

With stride 2 and size 2, the pooling will halve the input's size.

Formulas:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires two hyperparameters:
  - their spatial extent F,
  - the stride S,
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F)/S + 1$
  - $H_2 = (H_1 - F)/S + 1$
  - $D_2 = D_1$
- Introduces zero parameters since it computes a fixed function of the input
- For Pooling layers, it is not common to pad the input using zero-padding.

Where W_1 = 32, H_1 = 32, D_1 = 1, F = 2, S = 2

Model 1

1. How many neurons should the hidden layer of a network with a single hidden layer and an output layer have in the context of a classification problem with 25 classes have?
A. Depends on the problem and should be determined by means of validation
B. 3
C. 10
D. 25

We don't know the parameters of the problem, therefore we cannot decide the best hidden layer size.

2. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?
A. [10, 20, 30]
B. [0.16, 0.33, 0.5]
C. [1, 2, 3]
D. [0.0, 0.5, 1.0]

To apply L1 normalization compute the L1 norm of the vector: |10| + |20| + |30| = 60 and divide each value by the norm: [10/60, 20/60, 30/60] = [0.16, 0.33, 0.5]

3. What advantage does using a bias value bring in the context of the artificial neuron?
A. It significantly improves convergence time
B. It does not bring any advantage
C. It prevents the neuron hyperplanes from being forced to go through the origin
D. It significantly helps in the context of imbalanced data sets by providing a bias towards

the misrepresented class

A neuron with ReLU activation can be seen as creating a hyperplane separating the points in the input space.
By adding a bias to the neuron, the separation hyperplane can be moved away from the origin.

4. Which of the following neuron activation is the result of the tanh activation function?
A. [0.99, 0.05, 0.99]
B. [-1.2, 0.11, 1.2]
C. [1.01, 0.11, 0.2]
D. [0.9, 0.11, -1.1]

The output of tanh is in the range [-1, 1].

5. What is the output of the perceptron if input=[2.4, 3.0], weights=[-0.5, 0.2], bias=1.0 (activation function - sign)?
A. 0
B. -1
C. 1
D. 2.2

weights * input + bias = [-0.5, 0.2] * [[2.4], [3.0]] + 1.0
= -0.5 * 2.4 + 0.2 * 3.0 + 1
= -1.2 + 0.6 + 1 = 0.4

Sign of the output is positive => output is +1

6. What is the value of the loss function of a Ridge regression model if the predicted values y_hat are [-2, -3, -1], the ground-truth values are [-2, -3, -2.5], the wights are W = [1, 0], bias = 5 and alpha = 0.1?
A. 0.85
B. 0.75
C. 0.22
D. 0.95

(L2 (y_hat, y))^2 = (-2 + 2)^2 + (-3 + 3)^2 + (-1 + 2.5)^2 = 1.5^2

We divide the square of the L2 distance by **n**, where n is the number of examples we are computing the loss for (3 in this case).

Loss = 1/n (L2(y_hat, y))^2 + alpha * (1^2 + 0^2)
= 1/3 * 2.25 + 0.1 * 1
= 0.75 + 0.1

= 0.85

7. If we have the following probabilities for events P(A)=0.5 P(B)=0.9 P(A|B)=0.3, what is the value of P(B|A)?
A. 0.54
B. 0.75
C. 0.63
D. 0.27

Apply Bayes' theorem: P(B|A) = P(A|B) * P(B) / P(A) = 0.3 * 0.9 / 0.5 = 0.54

8. What is the label of the test example t = [5, 3, 8] if you apply the k-nearest neighbors classifier with k = 3 and metric = L1 (Manhattan distance) given the training data
X = [[1, 4, 2], [5, 4, 8], [2, 6, 5], [1, 1, 1], [2, 9, 6]],
Y = [2, 3, 3, 1, 2]?
A. 2
B. 3
C. 1
D. 0

L1 distances:
- |1 - 5| + |4 - 3| + |2 - 8| = 4 + 1 + 6 = 11
- |5 - 5| + |4 - 3| + |8 - 8| = 0 + 1 + 0 = 1
- |2 - 5| + |6 - 3| + |5 - 8| = 3 + 3 + 3 = 9
- |1 - 5| + |1 - 3| + |1 - 8| = 4 + 2 + 7 = 13
- |2 - 5| + |9 - 3| + |6 - 8| = 3 + 6 + 2 = 11

The top 3 smallest distances are the second, third, and the first and fifth are tied.
The values would be 3, 3 and 2. By majority vote, the winner is 3.

9. In which scenario is measuring the accuracy of the model not enough to evaluate the model properly?
A. When the data set is made out of audio samples
B. When the dataset is imbalanced
C. When there are 3 classes in the dataset
D. When the data set is balanced but the training set and test set come from different sources

If the dataset is imbalanced, the model can just always predict the most common class, and get better accuracy than if it was picked at random.

10. Can an SVM be used to achieve 100% training accuracy on the following 2D data set
[([0, 1], 1), ([1, 0], 1), ([0, 0], 1), ([-2, 2], 0), ([2, 2], 0), ([-2, -2], 0), ([2, -2], 0)]?
A. Yes, but only if the data is normalized
B. No, because the data is not linearly separable
C. Yes, by using the kernel trick

D. No, because the dataset is imbalanced

In theory, you can get 100% training accuracy on *any* data set with the right kernel function.

Model 2

1. Which of the following neuron activation is the result of the softmax activation function?
A. [0.6, 0.2, 0.2]
B. [0.5, 0.2, 0.2]
C. [0.6, 0.2, 0.3]
D. [0.6, -0.2, 0.2]

The values after applying softmax should sum up to 1.

2. Given the following vocabulary {0 - dogs, 1 - cats, 2 - candies, 3 - likes, 4 - she, 5 - he}. What is the bag of words (BOW) representation of the sentence "she likes dogs and horses."?
A. [1, 0, 0, 1, 1, 0]
B. [2, 0, 0, 1, 1, 0]
C. [1, 0, 0, 1, 1, 0, 1, 1]
D. [1, 0, 1, 1, 1, 0]

Sentence = {she, likes, dogs, and, horses}
Intersection with vocabulary = {she, likes, dogs}
Indices of words = {0, 3, 4}
Result vector = [1, 0, 0, 1, 1, 0]

3. How many neighbors should you consider in order to obtain the best result from a KNN classifier on the test set?
A. 1
B. 3
C. It depends on the problem and should be determined by means of validation
D. 7

k is a hyperparameter, depends on the problem.

4. What is the label of the test example t = [1, 2, 6] if you apply the k-nearest neighbors regressor with k = 3 and metric = L1 (Manhattan distance) given the training data
X = [[1, 4, 2], [5, 4, 8], [2, 6, 5], [1, 1, 1], [2, 9, 6]],
Y = [0.3, 0.6, 0.9, 0.6, 0.5]?
A. 0.6
B. 0.55
C. 0.65
D. 0.1

L1 distances:

- |1 - 1| + |4 - 2| + |2 - 6| = 0 + 2 + 4 = 6
- |5 - 1| + |4 - 2| + |8 - 6| = 4 + 2 + 2 = 8
- |2 - 1| + |6 - 2| + |5 - 6| = 1 + 4 + 1 = 6
- |1 - 1| + |1 - 2| + |1 - 6| = 0 + 1 + 5 = 6
- |2 - 1| + |9 - 2| + |6 - 6| = 1 + 7 + 0 = 8

Pick top 3 smallest distances: first, third and fourth neighbor.
Their labels are 0.3, 0.9, 0.6.
Being a regressor, we average their output.
The result is (0.3 + 0.9 + 0.6)/3 = 0.6

5. What will be the shape of the activation maps if we apply a 5x5 convolutional filter with stride=1 and no padding to a 16x16 image?
A. 14x14
B. 12x12
C. 18x18
D. 16x16

Formulas:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires two hyperparameters:
  - their spatial extent F,
  - the stride S,
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F)/S + 1$
  - $H_2 = (H_1 - F)/S + 1$
  - $D_2 = D_1$
- Introduces zero parameters since it computes a fixed function of the input
- For Pooling layers, it is not common to pad the input using zero-padding.

W_1 = 16
H_1 = 16
F = 5
S = 1

W_2 = (16 - 5)/1 + 1 = 12
H_2 = (16 - 5)/1 + 1 = 12

6. Suppose our model has the following metrics TP (true positives)=30, FP (false positives)=10, FN (false negatives)=30. What is the precision (P) and recall (R)?
A. P=50%, R=75%
B. P=75%, R=50%

C. P=10%, R=50%
D. P=30%, R=75%

R=TP/(TP+FN) = 30/60 = 50% ; P = TP/(TP+FP) = 30/40 = 75%

7. How many learned parameters (weights + biases) will a network with input size = 2, hidden layer size = 5, output layer size = 1, have?
A. 10
B. 8
C. 21
D. 13

First weight matrix: 2 * 5 = 10
First bias vector: 5
Second matrix: 5 * 1 = 5
Second bias vector: 1
Total: 10 + 5 + 5 + 1 = 21

8. What type of metric can achieve 100% training accuracy on the following 2D data set [([1, 1], 1), ([5, 5], 1), ([10, 10], 1), ([5, 4], 0), ([6, 5], 0), ([6, 4], 0)] when considering a 1-NN classifier?
A. Cosine
B. None of the answers
C. L2
D. L1

9. Which of the following is a linear classifier?
A. A 3-NN classifier
B. A neuron with no activation
C. A two layer neural network with ReLU activations
D. An SVM with polynomial kernel

Neuron with no activation is just Weights * Input + Bias

10. What is the value of the Mean Absolute Error function if the ground-truth labels are y = [6, 8, -9, 5] and the predicted labels are y_hat = [6.5, 7.2, 1, 7]?
A. 13.3
B. 3.325
C. 3.5
D. 13.5

Absolute differences: [|6 - 6.5|, |8 - 7.2|, |-9 - 1|, |5 - 7|] = [0.5, 0.8, 10, 2].
Sum of absolute values: 0.5 + 0.8 + 10 + 2 = 13.3
Average of absolute values: 13.3 / 4 = 3.325

Model 3

1. What advantage does using a bias value bring in the context of the artificial neuron?
A. It significantly improves convergence time
B. It prevents the neuron hyperplanes from being forced to go through the origin
C. It significantly helps in the context of imbalanced data sets by providing a bias towards the misrepresented class
D. It does not bring any advantage

2. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?
A. Cross Entropy
B. MSE
C. L2 Loss
D. L1 Loss

L1 loss uses absolute value function, which is not differentiable in 0, therefore cannot be used for gradient descent (at least theoretically).

3. The training data set contains the following examples [(3, PASS), (2, PASS), (2, PASS), (4, PASS), (0, FAIL), (1, FAIL), (3, FAIL), (1, FAIL)], the first component being the number of hours of study and the second denoting whether the student passed the exam. What is the probability of passing the exam with 2 hours of study - P(PASS|2)?

A. 25%

B. 50%

C. 75%

D. 100%

$P(pass \mid 2) = P(pass, 2)P(2)=22=1$

4. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size; the other numbers represent the number of neurons in each layer)?

A. 6x2

B. 6x1

C. 4x6

D. 2x1

The second layer consists of 2 neurons, while the previous one has 6. Presuming they are fully connected, the weight dimension of that layer is 6x2.

5. What is the output of the perceptron if input= [2.4, 3.0], weights= [-0.5, 0.2], bias=1.0 (activation function - sign)?

A. 1

B. 2.2

C. 0

D. -1

Weights * Input + Bias = -0.5 * 2.4 + 0.2 * 3.0 + 1.0 = 0.4

0.4 is positive, therefore sign is +1

6. What is the MSE for the following predicted labels y_pred = [0.1, 0.4, 0.7, 0.3] and truth labels = [1, 0, 1, 0]?

A. 0.3315
B. 0.1430
C. 0.0715
D. 0.2875

The Mean Squared Error is ((0.1 - 1)^2 + (0.4 - 0)^2 + (0.7 - 1)^2 + (0.3 - 0)^2)/4 =
= (0.81 + 0.16 + 0.09 + 0.09)/4 =
= 0.2875

7. What is the difference between using an L1 loss and an L2 loss?

A. Using the L1 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

B. The L2 loss generally favors having smaller errors instead of having fewer but greater errors while the L1 loss does not differentiate between these cases.

C. The L1 loss generally favors having smaller errors instead of having fewer but greater errors while the L2 loss does not differentiate between these cases.

D. Using the L2 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

From https://cs231n.github.io/classification/:

L1 vs. L2. It is interesting to consider differences between the two metrics. In particular, the L2 distance is much more unforgiving than the L1 distance when it comes to differences between two vectors. That is, the L2 distance prefers many medium disagreements to one big one. L1 and L2 distances (or equivalently the L1/L2 norms of the differences between a pair of images) are the most commonly used special cases of a p-norm.

8. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?

A. [0.0, 0.5, 1.0]

B. [10, 20, 30]

C. [0.16, 0.33, 0.5]

D. [1, 2, 3]

The L1 norm of the vector is || || = |10| + |20| + |30| = 60. Therefore, the normalized values are: [10/60, 20/60, 30/60] = [0.16, 0.33, 0.5]

9. What is the f1-score of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 1, 1] and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

A. 0.7

B. 0.5

C. 0.6

D. 0.4

True Positives = y is 1 and y_hat is 1 = 2

False Positives = y is 0 and y_hat is 1 = 2

False Negatives = y is 1 and y_hat is 0 = 2

Precision = TP/(TP + FP) = 2/(2 + 2) = 1/2

Recall = TP / (TP + FN) = 2/(2 + 2) = 1/2

F1 = 2 precision recallprecision + recall=2TPTP +FPTPTP+FNTPTP +FP+TPTP+FN=222+222+222+2+22+2=241644=2416=0.5

10. Which machine learning model can achieve the best performance in the context of an audio classification problem?

A. Depends on problem details and should be determined by means of validation

B. An SVM classifier

C. A Neural Network with five layers

D. A Neural Network with two layers

Model 4

1. Which of the following is a technique for using an SVM as a multi-class classifier?
A. Split group classification
B. One versus all
C. All versus all
D. N-way split
RASPUNS : B (only approaches for multi-class SVM are one-versus-all and one-versus-one)

2. If the data is split into 9 classes, and we want to train a SVM for classification. How many binary classifiers will be trained in the one-vs-one approach?
A. 18
B. 9
C. 36
D. 81

For every one out of N classes, we'll train a binary classifier vs the other N - 1 classes. That means a total of **(N * (N - 1)) / 2** classifiers (we divide by 2 since a A-vs-B classifier can be used as a B-vs-A classifier).

The answer is (N * (N - 1)) / 2 = (9 * 8) / 2 = 36

3. Which of the following is equivalent to a single artificial neuron without activation?
A. A KNN classifier with 3 neighbors
B. A Naive Bayes classifier
C. A neural network with no activations
D. An SVM with polynomial kernel
RASPUNS : C (ca fara activari totul se reduce la a inmulti matrici)

4. What is the output of neural network with 3 hidden units and 1 output unit having ReLU activations for the input x = [1, -2], if the weights are W1 = [-0.5, 3, -2; 2, -1, 0], B1 = [0, 1, -1], W2 = [-1; -1; 2], B2 = [2]?
A. 1
B. 4.5
C. 0
D. 8
RASPUNS: C 0

Example Python code:

```
import numpy as np


def relu(x):
    return np.maximum(x, 0)


x = np.array([1, -2])
W1 = np.array([[-0.5, 3, -2], [2, -1, 0]])
B1 = np.array([0, 1, -1])


W2 = np.array([[-1], [-1], [2]])
B2 = np.array([2])


H1 = W1.T @ x + B1
H1 = relu(H1)


H2 = W2.T @ H1.T + B2
H2 = relu(H2)


print(H2.item()) # prints 0.0
```

5. What is the value of PReLU(x) - parametric ReLU, where alpha=0.1 and x=-0.2?
A. -1
B. 0
C. 0.002
D. -0.02

ReLU is
- x when x is positive
- 0 when x is negative.

PReLU is
- x when x is positive
- alpha * x when x is negative.

Since x = -0.2 is negative, PReLU(x) will be alpha * -0.2 = 0.1 * -0.2 = -0.02

6. If the current weights of a perceptron are [0.2, 0.4], their gradients are = [-2.4, -1.2], and the learning rate is 0.1. What are the weights after the weights update operation?
A. [0.52, 0.44]
B. [0.44, 0.52]
C. [0.44, 0.44]
D. [0.52, 0.52]

Weight update operation is
new weights = weights - (learning rate) * gradients

= [0.2, 0.4] - 0.1 * [-2.4, -1.2]
= [0.2 + 0.24, 0.4 + 0.12]
= [0.44, 0.52]

10. What is the output of a SVM classifier for the input X = [0.1, -2, -5], if the weights are W = [-2, -1.2, -3] and the bias is b = 0.5?
A. 2
B. 0
C. 1
D. -1

W * X + b = -2 * 0.1 + 1.2 * 2 + 3 * 5 + 0.5 =
= -0.2 + 2.4 + 15 + 0.5 =
= 17.2 + 0.5 =
= 17.7
If we assume the SVM's output goes through a sign activation function, the result is positive, therefore the output is 1.

Model 5

1. What is the label of the test example x = [1, -1] with a 1-NN model based on the Euclidean distance having the training set S = {([2, -1], 1), ([1, 1], 2), ([-1, -1], 3)}?
A. 4
B. 3
C. 2
D. 1

The Euclidean distance towards each element of S is {(1-2)2+(-1+1)2=1, (1-1)2+(-1-1)2=2, (1+1)2+(-1+1)2=2}, therefore the element with label 1 is the closest neighbour to x, therefore x is given the label 1.

2. Calculate the cost for the Ridge Regression having weights=[3, 2], alpha=0.1,
y_true=[10, 1, 9, 4],
y_pred=[9, 3, 6, 7].
A. 36.23
B. 23.36
C. 23.00
D. 0.10

(L2 (y_hat, y))^2 = (10 - 9)^2 + (1 - 3)^2 + (9 - 6)^2 + (4 - 7)^2
= 1^2 + 2^2 + 3^2 + 3^2
= 1 + 4 + 9 + 9
= 23
(L2 (weights))^2 = 3^2 + 2^2 = 9 + 4 = 13

Ridge regression loss can be calculated by dividing (L2 (y_hat, y))^2 by n or not, in this case it seems the answer works only if we don't divide.

Loss = 23 + alpha * 13 = 23 + 0.1 * 13 = 23 + 1.3 = 24.3 (I'm guessing B was meant to be 24.36 not 23.36)

3. After training for 5 epochs, we have the following training losses for each epoch [0.60, 0.48, 0.30, 0.28, 0.26], and the following validation losses for each epoch [0.55, 0.43, 0.27, 0.27, 0.25]. Is the model overfitted, underfitted, both, or neither?
A. Neither
B. Overfitting
C. Both
D. Underfitting

We'll compare the training and validation **losses**:
0.60, 0.48, 0.30, 0.28, 0.26
> > > > >
0.55, 0.43, 0.27, 0.27, 0.25

Observe that the training loss keeps getting smaller, meaning the model keeps performing better on the training set. This means it's **not underfitting**, it can learn just fine.
Observe that the validation loss is also getting smaller, and it's **smaller than the training loss**, therefore the model generalizes well and keeps improving. It's **not overfitting**.

Exerciții Rezolvate ML

Test 48

1. What is the label of the test example t = [2, 3, 5] if you apply the k-nearest neighbors classifier with k = 1 and metric = L1 (Manhattan distance) given the training data
X = [[1, 4, 1], [2, 4, 7], [2, 30, 5], [0, 1, 0]],
Y = [1, 3, 2, 2]?
A. 2
B. 3
C. 0
D. 1

The L1 (Manhattan) distances are:
- [1, 4, 1] - [2, 3, 5] = |1 - 2| + |4 - 3| + |1 - 5| = 1 + 1 + 4 = 6
- [2, 4, 7] - [2, 3, 5] = |2 - 2| + |4 - 3| + |7 - 5| = 0 + 1 + 2 = 3
- [2, 30, 5] - [2, 3, 5] = |2 - 2| + |30 - 3| + |5 - 5| = 0 + 27 + 0 = 27
- [0, 1, 0] - [2, 3, 5] = 2 + 2 + 5 = 9

We need to pick the 1-nearest neighbor(s). That means the one neighbor with **minimum distance**. This is the **second** training example, which has **label 3**.

2. Given the following vocabulary {0 - dogs, 1 - cats, 2 - candies, 3 - likes, 4 - she, 5 - he}. What is the bag of words (BOW) representation of the sentence "she likes dogs and horses."?
A. [1, 0, 0, 1, 1, 0, 1, 1]
B. [1, 0, 1, 1, 1, 0]
C. [1, 0, 0, 1, 1, 0]
D. [2, 0, 0, 1, 1, 0]

The set of words in the sentence is { she, likes, dogs, and, horses }.
If we intersect this with the vocabulary, we have { she, likes, dogs }.
This means { 4, 3, 0 } so we need a vector where indices **0**, **3** and **4** are set to 1.
This means

v[0] = 1, v[3] = 1, v[4] = 1

which is

v = [1, 0, 0, 1, 1, 0]

3. What is the resulting data after applying min-max scaling to this data [[0.1, 0.4], [0.2, 0.5], [0.3, 0.6]] (3 examples, 2 features)?
A. [[0.0, 0.5], [0.25, 0.75], [0.5, 1.0]]
B. [[0.1, 0.4], [0.2, 0.5], [0.3, 0.6]]
C. [[0.0, 0.4], [0.25, 0.5], [0.5, 0.6]]
D. [[0.0, 0.0], [0.5, 0.5], [1.0, 1.0]]

Rescaling using min-max: https://en.wikipedia.org/wiki/Feature_scaling#Rescaling_(min-max_normalization)

Values on X axis: 0.1, 0.2, 0.3
Values on Y axis: 0.4, 0.5, 0.6
Minimum values on X, Y: [0.1, 0.4]
Maximum values on X, Y: [0.3, 0.6]
Difference between max and min values on each axis: [0.3 - 0.1, 0.6 - 0.4] = [0.2, 0.2]

Subtract minimum on each axis:
[[0.1 - 0.1, 0.4 - 0.4], [0.2 - 0.1, 0.5 - 0.4], [0.3 - 0.1, 0.6 - 0.4]]
= [[0, 0], [0.1, 0.1], [0.2, 0.2]]

Divide each axis by (max - min):
[[0 / 0.2, 0 / 0.2], [0.1 / 0.2, 0.1 / 0.2], [0.2 / 0.2, 0.2 / 0.2]
= [[0, 0], [0.5, 0.5], [1, 1]]

4. How many neurons should the hidden layer of a network with a single hidden layer and an output layer have in the context of a classification problem with 25 classes have?
A. 10
B. 25
C. 3
D. Depends on the problem and should be determined by means of validation

- Number of neurons in the **input layer** is the **number of features** in the input.
- Number of neurons in the **output layer** is the **number of classes** in the output.
- Hidden layers are hyperparameters that have to be determined by validation, they don't have a formula.

5. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size, the other numbers represent the amount of neurons in each layer)?
A. 6x1
B. 2x1
C. 4x6
D. 6x2

The second layer consists of 2 neurons, while the previous one has 6. Presuming they are fully connected, the weight dimension of that layer is 6x2.

6. Which classifier can achieve the best performance on a e-mail spam classification task?
A. A Neural Network with three layers
B. Depends on problem details and should be determined by means of validation
C. An SVM with RBF kernel
D. An SVM with linear kernel

We're not given enough information about the problem to pick a classifier.

7. Which of the following is a linear classifier?
A. A neuron with no activation
B. A 3-NN classifier
C. An SVM with polynomial kernel
D. A two layer neural network with ReLU activations

A neuron computes **f(Weight * input + bias)**, where f is the activation function.
With no activation function, this becomes a linear term: **Weight * input + bias**

8. What is the recall of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 0, 1] and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?
A. 0.23
B. 0.33
C. 0.99
D. 0.45

Formula:

$$\text{Recall} = \frac{tp}{tp + fn}$$

True positives are those with 1 in y and 1 in y_hat: 1 examples
False negatives are those with 1 in y and 0 in y_hat: 2 examples

Recall = 1/(1 + 2) = ⅓ = 0.33

9. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?
A. MSE
B. L2 Loss
C. Cross Entropy
D. L1 Loss

L1 loss uses absolute value function, which is not differentiable in 0, therefore cannot be used for gradient descent (at least theoretically).

10. What will be the shape of the activation maps if we apply a 2x2 max pooling with stride=2 to a 32x32 activation map?
A. 16x16
B. 32x32
C. 14x14
D. 28x28

With stride 2 and size 2, the pooling will halve the input's size.

Formulas:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires two hyperparameters:
  - their spatial extent F,
  - the stride S,
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F)/S + 1$
  - $H_2 = (H_1 - F)/S + 1$
  - $D_2 = D_1$
- Introduces zero parameters since it computes a fixed function of the input
- For Pooling layers, it is not common to pad the input using zero-padding.

Where W_1 = 32, H_1 = 32, D_1 = 1, F = 2, S = 2

Model 1

1. How many neurons should the hidden layer of a network with a single hidden layer and an output layer have in the context of a classification problem with 25 classes have?
A. Depends on the problem and should be determined by means of validation
B. 3
C. 10
D. 25

We don't know the parameters of the problem, therefore we cannot decide the best hidden layer size.

2. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?
A. [10, 20, 30]
B. [0.16, 0.33, 0.5]
C. [1, 2, 3]
D. [0.0, 0.5, 1.0]

To apply L1 normalization compute the L1 norm of the vector: |10| + |20| + |30| = 60 and divide each value by the norm: [10/60, 20/60, 30/60] = [0.16, 0.33, 0.5]

3. What advantage does using a bias value bring in the context of the artificial neuron?
A. It significantly improves convergence time
B. It does not bring any advantage
C. It prevents the neuron hyperplanes from being forced to go through the origin
D. It significantly helps in the context of imbalanced data sets by providing a bias towards the misrepresented class

A neuron with ReLU activation can be seen as creating a hyperplane separating the points in the input space.
By adding a bias to the neuron, the separation hyperplane can be moved away from the origin.

4. Which of the following neuron activation is the result of the tanh activation function?
A. [0.99, 0.05, 0.99]
B. [-1.2, 0.11, 1.2]
C. [1.01, 0.11, 0.2]
D. [0.9, 0.11, -1.1]

The output of tanh is in the range [-1, 1].

5. What is the output of the perceptron if input=[2.4, 3.0], weights=[-0.5, 0.2], bias=1.0 (activation function - sign)?
A. 0
B. -1

C. 1
D. 2.2

weights * input + bias = [-0.5, 0.2] * [[2.4], [3.0]] + 1.0
= -0.5 * 2.4 + 0.2 * 3.0 + 1
= -1.2 + 0.6 + 1 = 0.4

Sign of the output is positive => output is +1

6. What is the value of the loss function of a Ridge regression model if the predicted values y_hat are [-2, -3, -1], the ground-truth values are [-2, -3, -2.5], the wights are W = [1, 0], bias = 5 and alpha = 0.1?
A. 0.85
B. 0.75
C. 0.22
D. 0.95

(L2 (y_hat, y))^2 = (-2 + 2)^2 + (-3 + 3)^2 + (-1 + 2.5)^2 = 1.5^2

We divide the square of the L2 distance by **n**, where n is the number of examples we are computing the loss for (3 in this case).

Loss = 1/n (L2(y_hat, y))^2 + alpha * (1^2 + 0^2)
= 1/3 * 2.25 + 0.1 * 1
= 0.75 + 0.1
= 0.85

7. If we have the following probabilities for events P(A)=0.5 P(B)=0.9 P(A|B)=0.3, what is the value of P(B|A)?
A. 0.54
B. 0.75
C. 0.63
D. 0.27

Apply Bayes' theorem: P(B|A) = P(A|B) * P(B) / P(A) = 0.3 * 0.9 / 0.5 = 0.54

8. What is the label of the test example t = [5, 3, 8] if you apply the k-nearest neighbors classifier with k = 3 and metric = L1 (Manhattan distance) given the training data
X = [[1, 4, 2], [5, 4, 8], [2, 6, 5], [1, 1, 1], [2, 9, 6]],
Y = [2, 3, 3, 1, 2]?
A. 2
B. 3
C. 1
D. 0

L1 distances:

- |1 - 5| + |4 - 3| + |2 - 8| = 4 + 1 + 6 = 11
- |5 - 5| + |4 - 3| + |8 - 8| = 0 + 1 + 0 = 1
- |2 - 5| + |6 - 3| + |5 - 8| = 3 + 3 + 3 = 9
- |1 - 5| + |1 - 3| + |1 - 8| = 4 + 2 + 7 = 13
- |2 - 5| + |9 - 3| + |6 - 8| = 3 + 6 + 2 = 11

The top 3 smallest distances are the second, third, and the first and fifth are tied.
The values would be 3, 3 and 2. By majority vote, the winner is 3.

9. In which scenario is measuring the accuracy of the model not enough to evaluate the model properly?
A. When the data set is made out of audio samples
B. When the dataset is imbalanced
C. When there are 3 classes in the dataset
D. When the data set is balanced but the training set and test set come from different sources

If the dataset is imbalanced, the model can just always predict the most common class, and get better accuracy than if it was picked at random.

10. Can an SVM be used to achieve 100% training accuracy on the following 2D data set [([0, 1], 1), ([1, 0], 1), ([0, 0], 1), ([-2, 2], 0), ([2, 2], 0), ([-2, -2], 0), ([2, -2], 0)]?
A. Yes, but only if the data is normalized
B. No, because the data is not linearly separable
C. Yes, by using the kernel trick
D. No, because the dataset is imbalanced

In theory, you can get 100% training accuracy on *any* data set with the right kernel function.

Model 2

1. Which of the following neuron activation is the result of the softmax activation function?
A. [0.6, 0.2, 0.2]
B. [0.5, 0.2, 0.2]
C. [0.6, 0.2, 0.3]
D. [0.6, -0.2, 0.2]

The values after applying softmax should sum up to 1.

2. Given the following vocabulary {0 - dogs, 1 - cats, 2 - candies, 3 - likes, 4 - she, 5 - he}. What is the bag of words (BOW) representation of the sentence "she likes dogs and horses."?
A. [1, 0, 0, 1, 1, 0]

B. [2, 0, 0, 1, 1, 0]
C. [1, 0, 0, 1, 1, 0, 1, 1]
D. [1, 0, 1, 1, 1, 0]

Sentence = {she, likes, dogs, and, horses}
Intersection with vocabulary = {she, likes, dogs}
Indices of words = {0, 3, 4}
Result vector = [1, 0, 0, 1, 1, 0]

3. How many neighbors should you consider in order to obtain the best result from a KNN classifier on the test set?
A. 1
B. 3
C. It depends on the problem and should be determined by means of validation
D. 7

k is a hyperparameter, depends on the problem.

4. What is the label of the test example t = [1, 2, 6] if you apply the k-nearest neighbors regressor with k = 3 and metric = L1 (Manhattan distance) given the training data
X = [[1, 4, 2], [5, 4, 8], [2, 6, 5], [1, 1, 1], [2, 9, 6]],
Y = [0.3, 0.6, 0.9, 0.6, 0.5]?
A. 0.6
B. 0.55
C. 0.65
D. 0.1

L1 distances:

- |1 - 1| + |4 - 2| + |2 - 6| = 0 + 2 + 4 = 6
- |5 - 1| + |4 - 2| + |8 - 6| = 4 + 2 + 2 = 8
- |2 - 1| + |6 - 2| + |5 - 6| = 1 + 4 + 1 = 6
- |1 - 1| + |1 - 2| + |1 - 6| = 0 + 1 + 5 = 6
- |2 - 1| + |9 - 2| + |6 - 6| = 1 + 7 + 0 = 8

Pick top 3 smallest distances: first, third and fourth neighbor.
Their labels are 0.3, 0.9, 0.6.
Being a regressor, we average their output.
The result is (0.3 + 0.9 + 0.6)/3 = 0.6

5. What will be the shape of the activation maps if we apply a 5x5 convolutional filter with stride=1 and no padding to a 16x16 image?
A. 14x14
B. 12x12
C. 18x18
D. 16x16

Formulas:

- Accepts a volume of size $W_1 \times H_1 \times D_1$
- Requires two hyperparameters:
  - their spatial extent F,
  - the stride S,
- Produces a volume of size $W_2 \times H_2 \times D_2$ where:
  - $W_2 = (W_1 - F)/S + 1$
  - $H_2 = (H_1 - F)/S + 1$
  - $D_2 = D_1$
- Introduces zero parameters since it computes a fixed function of the input
- For Pooling layers, it is not common to pad the input using zero-padding.

W_1 = 16
H_1 = 16
F = 5
S = 1

W_2 = (16 - 5)/1 + 1 = 12
H_2 = (16 - 5)/1 + 1 = 12

6. Suppose our model has the following metrics TP (true positives)=30, FP (false positives)=10, FN (false negatives)=30. What is the precision (P) and recall (R)?
A. P=50%, R=75%
B. P=75%, R=50%
C. P=10%, R=50%
D. P=30%, R=75%

R=TP/(TP+FN) = 30/60 = 50% ; P = TP/(TP+FP) = 30/40 = 75%

7. How many learned parameters (weights + biases) will a network with input size = 2, hidden layer size = 5, output layer size = 1, have?
A. 10
B. 8
C. 21
D. 13

First weight matrix: 2 * 5 = 10
First bias vector: 5
Second matrix: 5 * 1 = 5
Second bias vector: 1
Total: 10 + 5 + 5 + 1 = 21

8. What type of metric can achieve 100% training accuracy on the following 2D data set [([1, 1], 1), ([5, 5], 1), ([10, 10], 1), ([5, 4], 0), ([6, 5], 0), ([6, 4], 0)] when considering a 1-NN classifier?
A. Cosine
B. None of the answers
C. L2
D. L1

9. Which of the following is a linear classifier?
A. A 3-NN classifier
B. A neuron with no activation
C. A two layer neural network with ReLU activations
D. An SVM with polynomial kernel

Neuron with no activation is just Weights * Input + Bias

10. What is the value of the Mean Absolute Error function if the ground-truth labels are y = [6, 8, -9, 5] and the predicted labels are y_hat = [6.5, 7.2, 1, 7]?
A. 13.3
B. 3.325
C. 3.5
D. 13.5

Absolute differences: [|6 - 6.5|, |8 - 7.2|, |-9 - 1|, |5 - 7|] = [0.5, 0.8, 10, 2].
Sum of absolute values: 0.5 + 0.8 + 10 + 2 = 13.3
Average of absolute values: 13.3 / 4 = 3.325

Model 3

1. What advantage does using a bias value bring in the context of the artificial neuron?
A. It significantly improves convergence time
B. It prevents the neuron hyperplanes from being forced to go through the origin
C. It significantly helps in the context of imbalanced data sets by providing a bias towards the misrepresented class
D. It does not bring any advantage

2. Which of the following does not constitute a valid loss for a neural network trained with gradient descent?
A. Cross Entropy
B. MSE
C. L2 Loss
D. L1 Loss

L1 loss uses absolute value function, which is not differentiable in 0, therefore cannot be used for gradient descent (at least theoretically).

3. The training data set contains the following examples [(3, PASS), (2, PASS), (2, PASS), (4, PASS), (0, FAIL), (1, FAIL), (3, FAIL), (1, FAIL)], the first component being the number of hours of study and the second denoting whether the student passed the exam. What is the probability of passing the exam with 2 hours of study - P(PASS|2)?

A. 25%

B. 50%

C. 75%

D. 100%

$P(pass \mid 2) = P(pass, 2)P(2)=22=1$

4. What is the dimension of the weights from the second layer of a neural network with the following configuration 4-6-2-1 (the first number is the input size; the other numbers represent the number of neurons in each layer)?

A. 6x2

B. 6x1

C. 4x6

D. 2x1

The second layer consists of 2 neurons, while the previous one has 6. Presuming they are fully connected, the weight dimension of that layer is 6x2.

5. What is the output of the perceptron if input= [2.4, 3.0], weights= [-0.5, 0.2], bias=1.0 (activation function - sign)?

A. 1

B. 2.2

C. 0

D. -1

Weights * Input + Bias = -0.5 * 2.4 + 0.2 * 3.0 + 1.0 = 0.4

0.4 is positive, therefore sign is +1

6. What is the MSE for the following predicted labels y_pred = [0.1, 0.4, 0.7, 0.3] and truth labels = [1, 0, 1, 0]?

A. 0.3315
B. 0.1430

C. 0.0715
D. 0.2875

The Mean Squared Error is ((0.1 - 1)^2 + (0.4 - 0)^2 + (0.7 - 1)^2 + (0.3 - 0)^2)/4 =
= (0.81 + 0.16 + 0.09 + 0.09)/4 =
= 0.2875

7. What is the difference between using an L1 loss and an L2 loss?

A. Using the L1 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

B. The L2 loss generally favors having smaller errors instead of having fewer but greater errors while the L1 loss does not differentiate between these cases.

C. The L1 loss generally favors having smaller errors instead of having fewer but greater errors while the L2 loss does not differentiate between these cases.

D. Using the L2 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

From https://cs231n.github.io/classification/:

> **L1 vs. L2.** It is interesting to consider differences between the two metrics. In particular, the L2 distance is much more unforgiving than the L1 distance when it comes to differences between two vectors. That is, the L2 distance prefers many medium disagreements to one big one. L1 and L2 distances (or equivalently the L1/L2 norms of the differences between a pair of images) are the most commonly used special cases of a p-norm.

8. What is the resulting data after applying L1 normalization to this vector [10, 20, 30]?

A. [0.0, 0.5, 1.0]

B. [10, 20, 30]

C. [0.16, 0.33, 0.5]

D. [1, 2, 3]

The L1 norm of the vector is || || = |10| + |20| + |30| = 60. Therefore, the normalized values are: [10/60, 20/60, 30/60] = [0.16, 0.33, 0.5]

9. What is the f1-score of the classifier if the ground-truth labels are y = [0, 1, 1, 0, 0, 0, 1, 1] and the predicted labels are y_hat = [1, 0, 0, 0, 0, 1, 1, 1]?

A. 0.7

B. 0.5

C. 0.6

D. 0.4

True Positives = y is 1 and y_hat is 1 = 2

False Positives = y is 0 and y_hat is 1 = 2

False Negatives = y is 1 and y_hat is 0 = 2

Precision = TP/(TP + FP) = 2/(2 + 2) = 1/2

Recall = TP / (TP + FN) = 2/(2 + 2) = 1/2

F1 = 2 precision recallprecision + recall=2TPTP +FPTPTP+FNTPTP +FP+TPTP+FN=222+222+222+2+22+2=241644=2416=0.5

10. Which machine learning model can achieve the best performance in the context of an audio classification problem?

A. Depends on problem details and should be determined by means of validation

B. An SVM classifier

C. A Neural Network with five layers

D. A Neural Network with two layers

Model 4

1. Which of the following is a technique for using an SVM as a multi-class classifier?
A. Split group classification
B. One versus all
C. All versus all
D. N-way split
RASPUNS : B (only approaches for multi-class SVM are one-versus-all and one-versus-one)

2. If the data is split into 9 classes, and we want to train a SVM for classification. How many binary classifiers will be trained in the one-vs-one approach?
A. 18
B. 9
C. 36
D. 81

For every one out of N classes, we'll train a binary classifier vs the other N - 1 classes. That means a total of **(N * (N - 1)) / 2** classifiers (we divide by 2 since a A-vs-B classifier can be used as a B-vs-A classifier).

The answer is (N * (N - 1)) / 2 = (9 * 8) / 2 = 36

3. Which of the following is equivalent to a single artificial neuron without activation?
A. A KNN classifier with 3 neighbors
B. A Naive Bayes classifier
C. A neural network with no activations
D. An SVM with polynomial kernel
RASPUNS : C (ca fara activari totul se reduce la a inmulti matrici)

4. What is the output of neural network with 3 hidden units and 1 output unit having ReLU activations for the input x = [1, -2], if the weights are W1 = [-0.5, 3, -2; 2, -1, 0], B1 = [0, 1, -1], W2 = [-1; -1; 2], B2 = [2]?
A. 1
B. 4.5
C. 0
D. 8
RASPUNS: C 0

Example Python code:

```
import numpy as np


def relu(x):
    return np.maximum(x, 0)


x = np.array([1, -2])
W1 = np.array([[-0.5, 3, -2], [2, -1, 0]])
B1 = np.array([0, 1, -1])


W2 = np.array([[-1], [-1], [2]])
B2 = np.array([2])


H1 = W1.T @ x + B1
H1 = relu(H1)


H2 = W2.T @ H1.T + B2
H2 = relu(H2)


print(H2.item()) # prints 0.0
```

5. What is the value of PReLU(x) - parametric ReLU, where alpha=0.1 and x=-0.2?
A. -1

B. 0
C. 0.002
D. -0.02

ReLU is
- x when x is positive
- 0 when x is negative.

PReLU is
- x when x is positive
- alpha * x when x is negative.

Since x = -0.2 is negative, PReLU(x) will be alpha * -0.2 = 0.1 * -0.2 = -0.02

6. If the current weights of a perceptron are [0.2, 0.4], their gradients are = [-2.4, -1.2], and the learning rate is 0.1. What are the weights after the weights update operation?
A. [0.52, 0.44]
B. [0.44, 0.52]
C. [0.44, 0.44]
D. [0.52, 0.52]

Weight update operation is
new weights = weights - (learning rate) * gradients
= [0.2, 0.4] - 0.1 * [-2.4, -1.2]
= [0.2 + 0.24, 0.4 + 0.12]
= [0.44, 0.52]

10. What is the output of a SVM classifier for the input X = [0.1, -2, -5], if the weights are W = [-2, -1.2, -3] and the bias is b = 0.5?
A. 2
B. 0
C. 1
D. -1

W * X + b = -2 * 0.1 + 1.2 * 2 + 3 * 5 + 0.5 =
= -0.2 + 2.4 + 15 + 0.5 =
= 17.2 + 0.5 =
= 17.7

If we assume the SVM's output goes through a sign activation function, the result is positive, therefore the output is 1.

Model 5

1. What is the label of the test example x = [1, -1] with a 1-NN model based on the Euclidean distance having the training set S = {([2, -1], 1), ([1, 1], 2), ([-1, -1], 3)}?
A. 4
B. 3
C. 2

D. 1

The Euclidean distance towards each element of S is {(1-2)2+(-1+1)2=1, (1-1)2+(-1-1)2=2, (1+1)2+(-1+1)2=2}, therefore the element with label 1 is the closest neighbour to x, therefore x is given the label 1.

2. Calculate the cost for the Ridge Regression having weights=[3, 2], alpha=0.1,
y_true=[10, 1, 9, 4],
y_pred=[9, 3, 6, 7].
A. 36.23
B. 23.36
C. 23.00
D. 0.10

(L2 (y_hat, y))^2 = (10 - 9)^2 + (1 - 3)^2 + (9 - 6)^2 + (4 - 7)^2
= 1^2 + 2^2 + 3^2 + 3^2
= 1 + 4 + 9 + 9
= 23
(L2 (weights))^2 = 3^2 + 2^2 = 9 + 4 = 13

Ridge regression loss can be calculated by dividing (L2 (y_hat, y))^2 by n or not, in this case it seems the answer works only if we don't divide.

Loss = 23 + alpha * 13 = 23 + 0.1 * 13 = 23 + 1.3 = 24.3 (I'm guessing B was meant to be 24.36 not 23.36)

3. After training for 5 epochs, we have the following training losses for each epoch [0.60, 0.48, 0.30, 0.28, 0.26], and the following validation losses for each epoch [0.55, 0.43, 0.27, 0.27, 0.25]. Is the model overfitted, underfitted, both, or neither?
A. Neither
B. Overfitting
C. Both
D. Underfitting

We'll compare the training and validation **losses**:
0.60, 0.48, 0.30, 0.28, 0.26
> > > > >
0.55, 0.43, 0.27, 0.27, 0.25

Observe that the training loss keeps getting smaller, meaning the model keeps performing better on the training set. This means it's **not underfitting**, it can learn just fine.
Observe that the validation loss is also getting smaller, and it's **smaller than the training loss**, therefore the model generalizes well and keeps improving. It's **not overfitting**.