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)](https://xlinux.nist.gov/dads/HTML/manhattanDistance.html) 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 input is a vector of size 4, we can imagine it as a column matrix of size (4, 1).

Second layer (hidden layer of neurons) must have 6 activations (because it has 6 neurons).

Third layer has 2 activations.

Output layer has size 1.

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:



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**

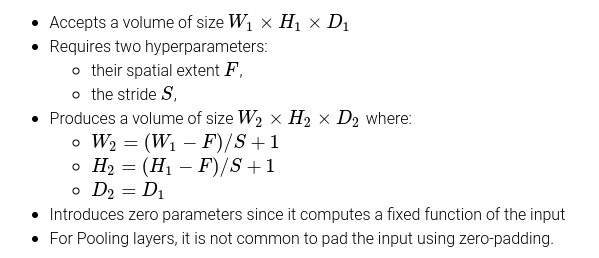
B. 32x32

C. 14x14

D. 28x28

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

Formulas:



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](https://en.wikipedia.org/wiki/Sign_function))?

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

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 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

# 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?

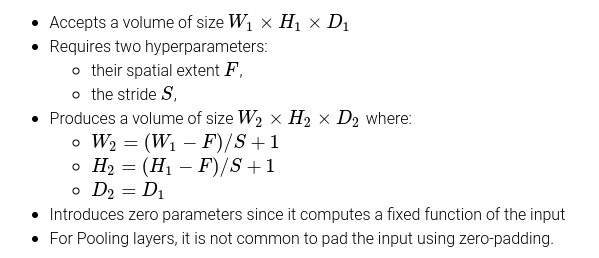
A. 14x14

**B. 12x12**

C. 18x18

D. 16x16

Formulas:



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%

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%**

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?

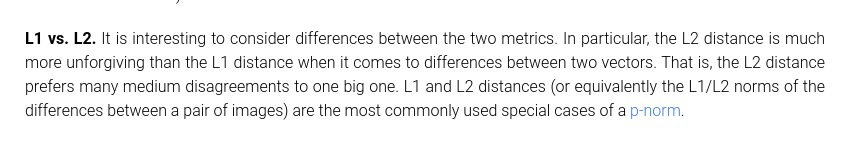
A. Using the L1 loss you can avoid getting stuck in a local minimum when using stochastic gradient descent in the case of neural networks.

**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/>:



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

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

RASPUNS: (N^2 \* (N^2 - 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

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