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 hidden layer consists of 2 neurons, while the first hidden layer 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

If we assume that accuracy is computed by leaving out the point when doing the prediction (otherwise each point would be nearest to itself), the points with label 1 are on a line, and the points with label 0 are on different lines.

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%

$$P(pass \mid 2) = \frac{P(pass, 2)}{P(2)} = \frac{2}{2} = 1$$

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

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 2 \cdot \frac{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} = 2 \cdot \frac{\frac{2}{2+2} \cdot \frac{2}{2+2}}{\frac{2}{2+2} + \frac{2}{2+2}} = 2 \cdot \frac{\frac{4}{16}}{\frac{4}{4}} = 2 \cdot \frac{4}{16} = 0.5$$

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 $\{\sqrt{(1-2)^2+(-1+1)^2}=1, \sqrt{(1-1)^2+(-1-1)^2}=2, \sqrt{(1+1)^2+(-1+1)^2}=2\}$, 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

In this case, we don't scare the L2 of the weights:
L2 (weights) = sqrt(3^2 + 2^2) = sqrt(9 + 4) = sqrt(13) = 3.6

Ridge regression loss can be calculated by multiplying (L2 (y_hat, y))^2 by 1/n or not (depends on formulation), in this case it seems the answer works only if we don't divide.

Loss = 23 + alpha * 3.6 = 23 + 0.1 * 3.6 = 23 + 0.36 = 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

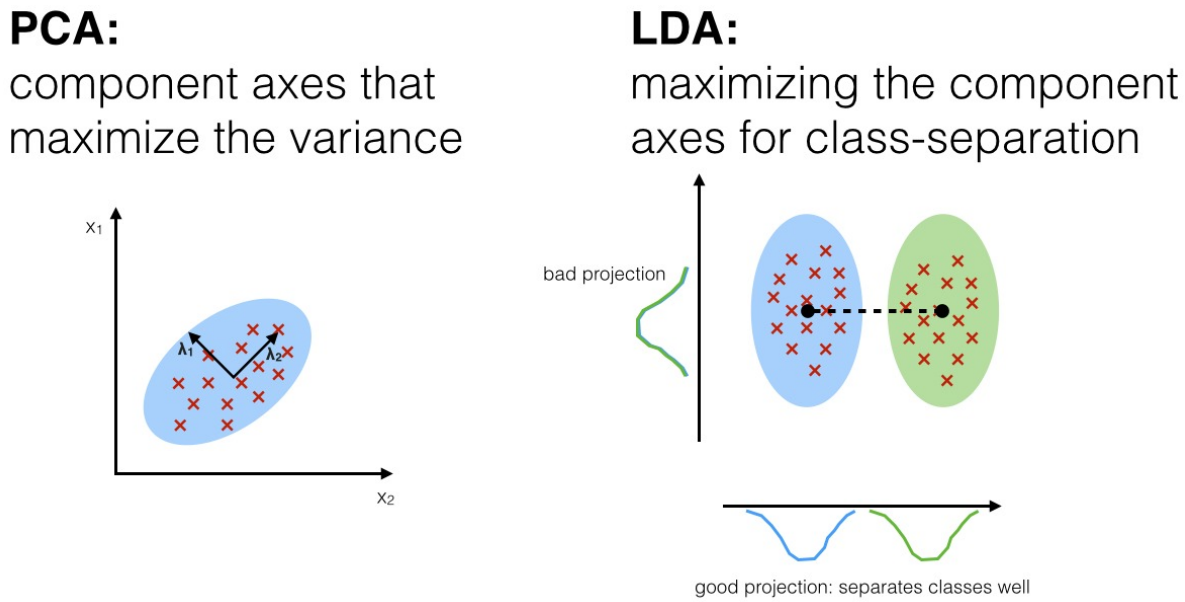
We'll compare the training and validation **losses**:

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**.

Culese

7. In cazul analizei liniar discriminante, hiperplanul pe care se proiecteaza punctele este:
A. Paralel cu hiperplanul de separare, distanta fiind controlata prin bias
B. Perpendicular pe hiperplanul de separare
C. Orienta astfel incat punctele sa fie cat mai departate
D. Orientat astfel incat punctele sa fie cat mai apropriate



3. Cat este valoarea functiei de pierdere data de media patratelor erorilor daca etichetele corecte sunt y = [100, -25, 0.5] si etichetele prezise sunt y_hat = [101, -23, 0]?
A. 1.16
B. 1.1
C. 1.75
D. 1.7

Calculăm diferențele pentru fiecare element, le ridicăm la pătrat, și facem media:
MSE = 1/3 ((100 - 101)^2 + (-25 + 23)^2 + (0.5 - 0)^2)
= 1/3 (1 + 4 + 0.25)
= 1/3 * 5.25

= 1.75

9. Ce activare aduce cea mai mare performanta in contextul unei probleme de clasificare cu doua clase si un clasificator format dintr-un singur neuron?
A. ReLU
B. Sigmoid
C. Leaky ReLU
D. Liniara

Output-ul unui astfel de neuron ar trebui să fie o probabilitate, ca să poată avea sens rezultatul. Singura funcție care se aplică dintre cele din listă este sigmoid (produce un număr în [0, 1]).

2. Care dintre urmatoarele asigura media 0 pentru fiecare din trasaturile din setul de date? 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. L1 Normalization (x / sum |x_i| pentru fiecare exemplu x)
B. Standard Normalization (X_i / std(X_i) - mean(X_i) pentru fiecare trasatura i)
C. Min-Max Scaling ((X_i - min(X_i)) / (max(X_i) - min(X_i)) pentru fiecare trasatura i)
D. L2 Normalization (x / sqrt(sum x_i^2) pentru fiecare exemplu x)

Normalizarea standard aduce datele la medie 0 și deviație standard 1.

7. Care dintre urmatoarele functii nu este functie nucleu?
A. K(x,y) = 5x - 2y
B. K(x,y) = sum(sqrt(x_i·y_i)) + sum(min{x_i,y_i})
C. K(x,y) = sum(min{x_i,y_i})
D. K(x,y) = (x·y + 10)**5

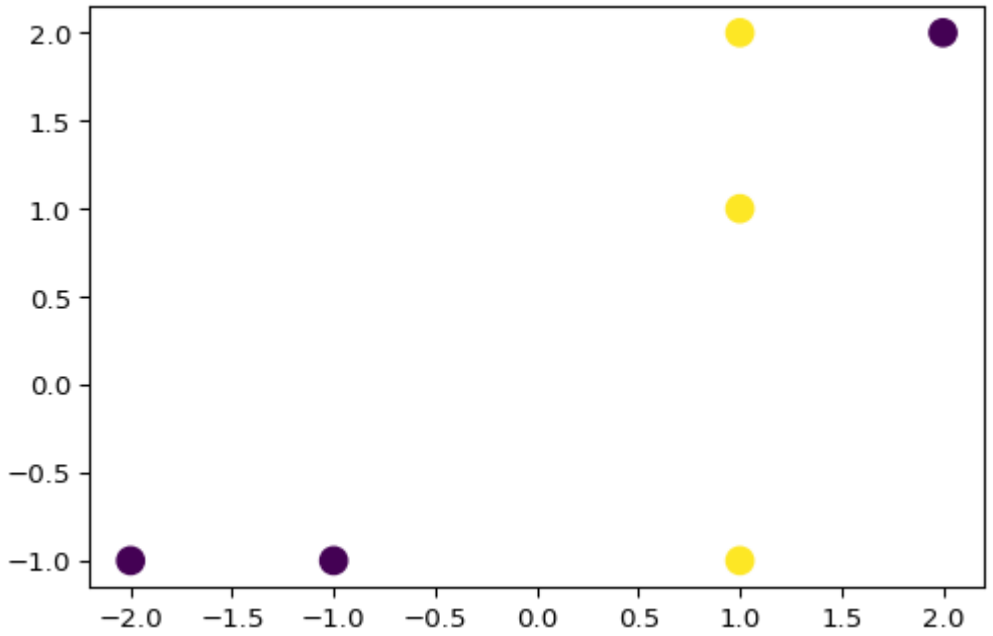
Funcția nucleu trebuie să respecte aceleași proprietăți ca produsul scalar:

- <x, y> = <y, x>
- <x, y> >= 0
- <x, y> = 0 ddacă x = 0 sau y = 0

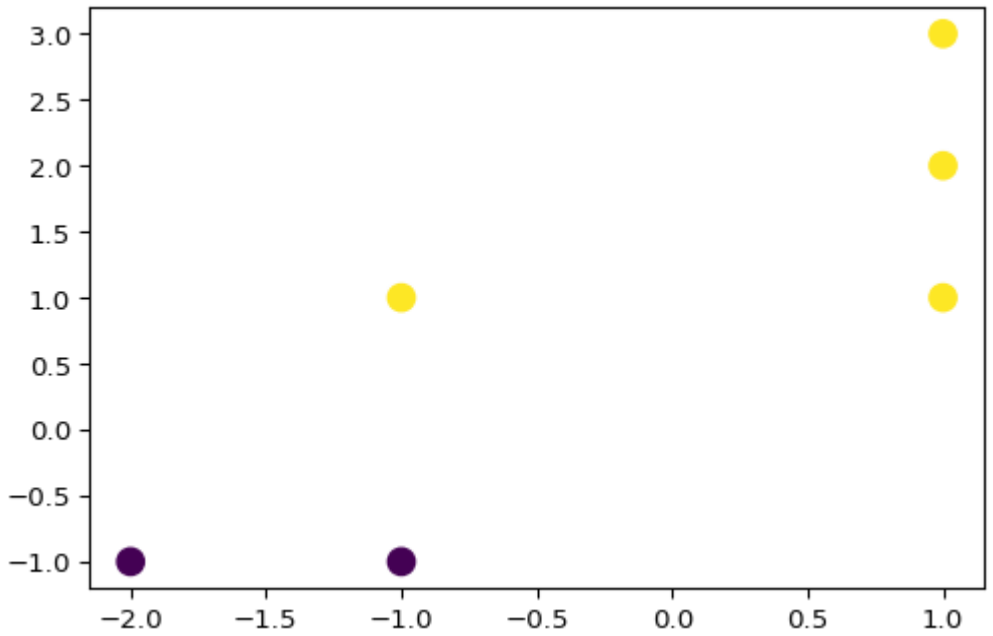
Funcția de la A nu respectă aceste proprietăți.

Care dintre punctele urmatoare, alaturi de etichetele corespunzatoare, pot fi discriminate corect de un perceptron?
A. X = ((1,1),(1,2),(2,2),(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),(-1,2),(1,3),(1,-1),(-1,-1),(-2,-1)) Y = (1,1,-1,1,-1,-1)
D. X = ((1,1),(1,3),(2,3),(2,1),(-1,2),(3,2)) Y = (1,1,1,-1,-1,-1)

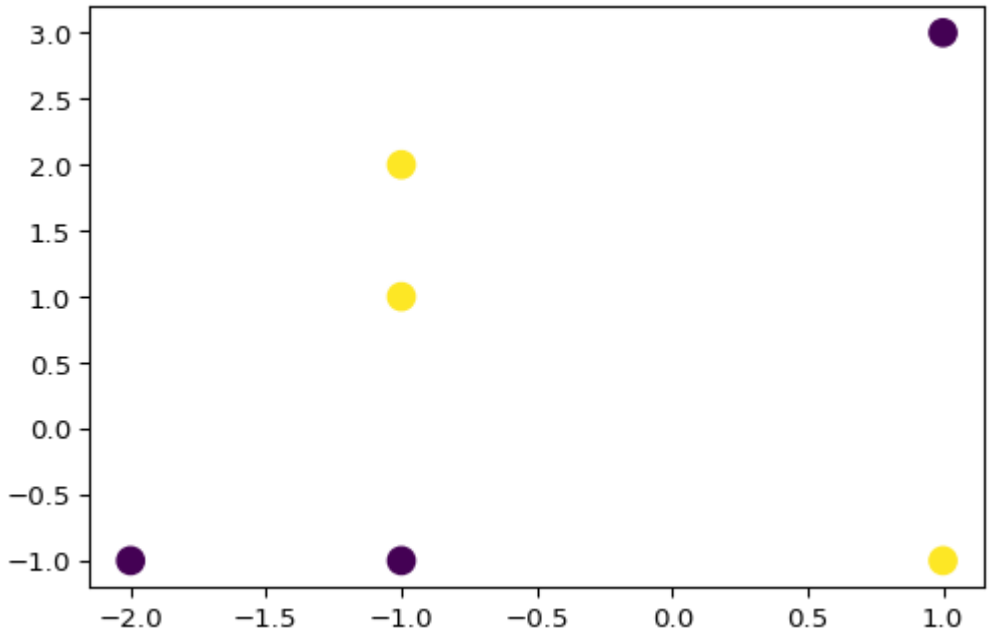
Trebuie să reprezentăm grafic datele:
A:


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

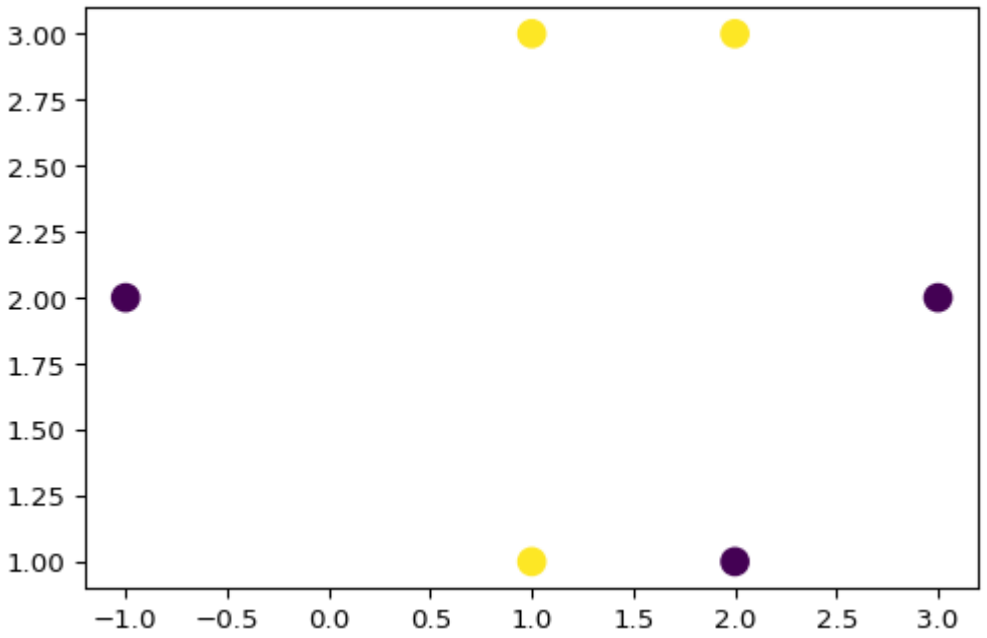
B:


3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0

C:



D:



Alegem varianta B, la care putem găsi o dreaptă liniară de separație (un singur perceptron poate clasifica doar date liniar separabile).

Cod:

```
import matplotlib.pyplot as plt

points = ((1,1),(1,2),(2,2),(1,-1),(-1,-1),(-2,-1))
labels = (1,1,-1,1,-1,-1)
```

```
x, y = zip(*points)
plt.figure(dpi=96)
plt.scatter(x, y, s=100, c=labels)
plt.plot()
```

9. Cand este mai eficient sa folosim reprezentarea duala a datelor?
A. Cand avem o problema de clasificare cu foarte multe clase (mai mult de doua)
B. Cand avem o problema de clasificare binara (cu doua clase)
C. Cand numarul de trasaturi este mai mic decat numarul de exemple
D. Cand numarul de trasaturi este mai mare decat numarul de exemple

De exemplu, din https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html:
dualbool, default=False
Dual or primal formulation. Dual formulation is only implemented for l2 penalty with liblinear solver. Prefer dual=False when n_samples > n_features.

Deci e mai bine dual=True când n_samples << n_features.

9. 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. Cosinus
C. L1
D. Niciunul dintre raspunsuri

Dacă atunci când calculăm acuratețea nu luăm și punctul în sine în datele de intrare (pentru că atunci ar fi 100% pentru orice metrică), observăm că punctele cu aceeași clasă se află pe aceeași dreaptă, deci ar putea fi clasificate corect.

7. Pot fi folosite modelele de regresie pentru a face clasificare?
A. Nu, pentru ca nu sunt facute pentru clasificare si nu pot si adaptate in niciun mod pentru astfel de probleme.
B. Da, pot fi folosite aproape intotdeauna cu mici ajustari.
C. Da, dar doar daca setul de date este balansat.
D. Nu, pentru ca seturile de date nu sunt suficient de mari in general.

Se pot antrena modele de clasificare folosind aceleași reguli matematice (de exemplu, SVM classifier și SVM regressor), dar nu poate fi făcută o simplă ajustare pentru clasificare.

3. De ce este necesar setul de test?
A. Pentru ca setul de antrenare nu este in general suficient de mare.

B. Pentru ca testarea pe validare este neriguroasa din moment ce folosim acest set pentru determinarea hiperparametrilor.
C. Pentru ca setul de test ajuta la obtinerea overfitting-ului.
D. Pentru ca datele ar trebui intotdeauna impartite 70%-15%-15%

Când modificăm hiperparametrii pe datele de validare, noi înșine facem un fel de optimizare, și suntem la fel de expuși la a produce overfitting.

Care este scopul ratei de invatare in contextul retelelor neurale?
A. Ajuta procesul de convergenta prin prevenirea pasilor mult prea mari care pot depasi minimul global
B. Specifica rata de selectie pentru batch-uri, astfel ajutand procesul de invatare
C. Este o foarte necesara initializare pentru bias-uri
D. Favorizeaza depasirea exagerata a punctului de minim si reguleaza viteza de convergenta

Explicația vine din înțelegerea algoritmului de gradient descent.

5. Fiind date etichitele
(1, 2, 1, 1, 2, 1, 1, 1)
si predictiile
(1, 2, 1, 1, 2, 2, 1, 2),
care sunt acuratetea si scorul f1?
A. acc = 0.82 f1 = 0.66
B. acc = 0.75 f1 = 0.75
C. acc = 0.75 f1 = 0.8
D. acc = 0.5 f1 = 0.66?

Acuratețe = nr_corecte / nr_total
F1 = 2 * ((precision * recall)/(precision + recall))

Cod:
```
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score

y_true = (1, 2, 1, 1, 2, 1, 1, 1)
y_pred = (1, 2, 1, 1, 2, 2, 1, 2)

print(accuracy_score(y_true, y_pred))
print(f1_score(y_true, y_pred))
```

8. Ce avantaj aduce folosirea bias-ului in contextului neuronului artificial?
A. Nu aduce niciun avantaj
B. Ajuta aducand constragerea de a trece prin origine a hiperplanelor determinate de neuroni

C. Ajuta prin evitarea situatiilor in care toate hiperplanele determinate de neuroni trebuie sa treaca prin origine
D. Ajuta semnificativ in contextul in care datele sunt debalansate inclinand (bias) decizia catre clasa defavorizata

Un neuron calculează funcția WX + b. Dacă nu am avea termenul b, atunci am avea un sistem de ecuații omogene, care determină un hiperplan care trece prin origine.

3. Care dintre urmatoarele nu constituie o functie de pierdere valida pentru o retea neurala antrenata prin metoda coborarii pe gradient?
A. L2 Loss
B. Cross Entropy
C. L(output, label) = 0 daca output - label < 0, altfel 1
D. Media partatelor erorilor

C nu este nici continuă, nici derivabilă

4. Care dintre urmatoarele functii nu este functie nucleu?
A. K(x,y) = sum(sqrt(x_i·y_i)) + sum(min{x_i,y_i})
B. K(x,y) = sum(min{x_i,y_i})
C. K(x,y) = 5x - 2y
D. K(x,y) = (x·y + 10)**5

Funcțiile nucleu trebuie să fie simetrice în x, y.

1. 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. 0.55
B. 0.65
C. 0.75
D. 0.45

Aplicăm teorema lui Bayes:
P(B|A) = P(A|B) * P(B) / P(A) = 0.33 * 0.5 / 0.3 = 0.55

2. Fiind data urmatoarea multime de antrenare cu etichetele corespunzatoare:
X_train = ((1,0),(1,1),(1,2),(2,1),(0,1),(-1,1),(-1,-1),(-1,2))
Y_train = (1,5,4,3,2,3,3,2), cati clasificatori sunt antrenati pentru un SVM folosind abordarea ONE vs ALL?
A. 5
B. 4
C. 6
D. 10

În one-vs-all, antrenăm câte un clasificator pentru fiecare clasă. Dacă ne uităm în mulțimea de antrenare, avem clasele {1, 2, 3, 4, 5}. Deci vom antrena 5 clasificatori.

3. Care dintre urmatoarele asigura media 0 pentru fiecare din trasaturile din setul de date? 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. L1 Normalization (x / sum |x_i| pentru fiecare exemplu x)
B. L2 Normalization (x / sqrt(sum x_i^2) pentru fiecare exemplu x)
C. Standard Normalization (X_i / std(X_i) - mean(X_i) pentru fiecare trasatura i)
D. Min-Max Scaling ((X_i - min(X_i)) / (max(X_i) - min(X_i)) pentru fiecare trasatura i)

Feature scaling

4. Cate ponderi (inclusiv bias) are o retea neuronala cu configuratia 2-5-8-1 (primul numar este dimensiunea datelor de intrare, urmatoarele numere reprezinta numarul de neuroni de pe fiecare strat)?
A. 58
B. 74
C. 72
D. 76

Pe primul strat: 2 * 5 (ponderi) + 5 (bias) = 15
Pe al doilea strat: 5 * 8 + 8 = 48
Pe al treilea strat: 8 * 1 + 1 = 9
Total: 72

5. Care sunt punctele in care dreapta de separare a percetronului w = [2, -4], b = [-1] intersecteaza axele?
A. (-1, 0) si (0, 0.5)
B. (0, -1) si (0.5, 0)
C. (0, 0.5) si (0.25, 0)
D. (0.5, 0) si (0, -0.25)

Dreapta de separare a perceptronului este
d: W * X + b = 2x - 4y - 1 = 0
Intersectăm d cu axele Ox și Oy:

- Intersecția cu Oy: 2 * 0 - 4y - 1 = 0 ⇔ 4y = -1 ⇔ y = -1/4 = -0.25
- Intersecția cu Ox: 2x - 4 * 0 - 1 = 0 ⇔ 2x = 1 ⇔ x = 1/2 = 0.5

6. Care este rezultatul modelului "Masini cu vectori suport" pentru datele de intrare X = [0.5, -2, -5, 0.9], daca ponderile sunt W = [-2, -1.2, -3, 1.2] si bias-ul b = -0.5?
A. 1
B. -1
C. 2
D. 0

7. Care este valoarea de iesire a perceptronului daca input=[2.4, 3.0], ponderi=[-0.5, 0.2], bias=0.0 (functia de activare - sign)?

A. -1
B. 0
C. 1
D. -0.6

Output = W X + b =
= 2.4 * (-0.5) + 3.0 * 0.2 + 0.0
= -0.6
Aplicăm funcția de activare:
sign(-0.6) = -1

8. Care este precizia unui clasificator daca etichetele corecte sunt
y = [1, 1, 1, 1, 0, 0, 0, 1] si cele prezise sunt
y_hat = [1, 0, 0, 1, 0, 1, 1, 1]?
A. 0.2
B. 0.4
C. 0.1
D. 0.6

True Positives = y este 1 și y_hat este 1 => 3 exemple
False Positives = y este 0 și y_hat este 1 => 2 exemple

Precision = True Positives / (True Positives + False Positives) = 3 / (3 + 2) = 3/5 = 0.6

9. Cum ajuta normalizarea in contextul unui clasificator KNN?
A. Normalizarea creste performanta cand setul de date este balansat
B. Normalizarea aduce trasaturi diferite pe aceesi scala astfel incat distantele L1/L2 dintre exemple sa nu fie impactate de variantele diversificate
C. Normalizarea ajuta la reducerea zgomotului si asigura o convergenta mai inceata
D. Normalizarea nu ajuta

Acest lucru este menționat și pe https://en.wikipedia.org/wiki/Feature_scaling#Motivation

10. Care dintre urmatoarele afirmatii este gresita?
A. Retele neuronale pot invata problema XOR
B. Setarea parametrului C la o valoare prea mare in timpul antrenarii unui SVM poate conduce la supra-invatare
C. Tangenta hiperbolica nu se satureaza
D. Al doilea strat dintr-o retea neuronala nu poate primi date de intrare

Tangenta hiperbolică ia valori în -1, 1 deci se saturează. O funcție de activare care nu se săturează ar trebui să poată lua valori oricât de mari.

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 14 clase?
A. 24

B. Depinde de problema si ar trebui determinat prin validare
C. 14
D. 3

Dacă vrem să prezicem una din cele 14 clase, stratul de ieșire trebuie să aibă 14 neuroni, pe care aplicăm softmax și obținem probabilitățile corespunzătoare.

2. Fiind date etichetele
y = [23, 14, 30, 45, 18, 31]
si predictiile aferente
p = [26, 20, 39, 38, 18, 33],
care este masura Kendall Tau?
A. 0.4
B. 0.6
C. 0.2
D. 1.0

Cod:

```
from scipy.stats import kendalltau

x = [23, 14, 30, 45, 18, 31]
y = [26, 20, 39, 38, 18, 33]

correlation, _ = kendalltau(x, y)
print(correlation)
```

3. Care dintre urmatoarele este o tehnica folosita pentru SVM in contextul a mai mult de doua clase?
A. N way split
B. One versus all
C. Split group classification
D. All versus all

La SVM cu mai multe clase putem folosi one-versus-all sau one-versus-one.

4. Setul de antrenare contine urmatoarele date [(3, PASS), (2, PASS), (2, PASS), (1, PASS), (0, FAIL), (1, FAIL), (3, FAIL), (1, FAIL)], Prima componenta reprezinta numarul de ore de studiu, iar a doua componenta indica daca studentul a trecut examenul. Care e probabilitatea trecerii examenului daca studentul a invatat o singura ora - P(PASS|1)?
A. 75%
B. 66%
C. 33%
D. 100%

Probabilitatea condiționată P(A|B) = P(A intersectat cu B)/P(B)

În cazul nostru: P(PASS | 1) = P(PASS și a studiat o oră)/P(a studiat o oră)

Numărul de exemple cu studenți care au studiat o singură oră: 3
Numărul de exemple cu studenți care au studiat o singură oră și au trecut: 1

P(PASS | 1) = 1/3 = 33%

5. Care dintre urmatoarele nu este o strategie de prevenire a overfittingului in retelele neuronale?
A. Normalizarea datelor
B. Ajustarea parameterilor pe datele de validare
C. Introducerea regularizarii
D. Reducerea dimensiunii retelei

Conform https://en.wikipedia.org/wiki/Feature_scaling#Motivation, normalizarea ajută la antrenare și la calculele din algoritmi, dar nu are vreo legătură cu overfitting-ul.

Dacă la B e vorba de modificarea hiperparametrilor, atunci ar fi putea fi un răspuns corect. Dacă prin B se înțelege antrenarea ponderilor pe datele de validare, atunci asta nu ar ajuta cu overfitting-ul.

6. Care este scufundarea asociata functie nucleu liniare?
A. k(x,y) = <x,y>, unde <,> denote produsul scalar
B. f(x) = x
C. f(x) = sqrt(x)
D. Nu exista

Scufundare (embedding) înseamnă o funcție care duce vectorii dintr-un spațiu cu puține dimensiuni într-unul cu mai multe dimensiuni.
În cazul nostru, la A avem funcția nucleu liniară, care doar calculează produse scalare. Funcția de scufundare asociată ei ar fi f(x) = x

7. Care este acuratetea unui clasificator daca etichetele corecte sunt
y = [8, 7, 4, 1, 0, 6, 3, 2] si cele prezise sunt
y_hat = [5, 6, 6, 1, 0, 2, 2, 1]?
A. 0.55
B. 0.2
C. 0.25
D. 0.52

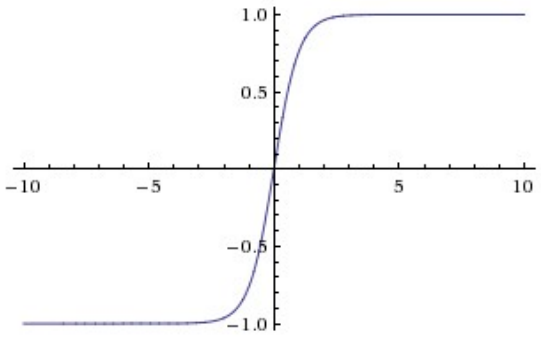
Acuratețea este nr_corecte / nr_total, în acest caz avem 2 răspunsuri corecte și 8 în total, deci acuratețea este 2/8 = ¼ = 0.25

8. Care dintre urmatoarele afirmatii este gresita?
A. Tangenta hiperbolica nu se satureaza
B. Al doilea strat dintr-o retea neuronala nu poate primi date de intrare

C. Retele neuronale pot invata problema XOR
D. Setarea parametrului C la o valoare prea mare in timpul antrenarii unui SVM poate conduce la supra-invatare

O funcție de activare nu se saturează dacă output-ul ei poate fi oricât de mare.
De exemplu, ReLU = max(0, x) poate fi oricât dacă x crește.
tanh este o funcție mărginită.

The tanh (hyperbolic tangent) activation function is saturating as it squashes real numbers to range between $[-1, 1]$:



https://stats.stackexchange.com/a/174438

9. Daca datele sunt impartite in 9 clase si folosim un SVM pentru antrenare, cate clasificatoare binare vor fi antrenate prin metoda one-vs-many?
A. 36
B. 81
C. 18
D. 9

La metoda one-vs-many, antrenăm un clasificator pentru fiecare clasă, deci vom avea 9 clasificatori.

10. Care dintre urmatoarele activari ale neuronilor este rezultatul functiei de activare softmax?
A. [0.9, 0.1, 0.0]
B. [0.9, 0.1, 0.01]
C. [0.9, 0.11, 0.0]
D. [0.8, 0.1, 0.11]

Rezultatul funcției softmax trebuie să fie un vector în care valorile adunate să dea 1.
Deci singura posibilitate este A.