Name:____ UniversityID:____________________
*

Master in Computer Vision

Barcelona

UAB UOC UPC upf

Module: M3. Machine learning for computer vision **Final exam**

Date: February 27th, 2023

- Books, lecture notes, calculators, phones, etc. are not allowed.
- Write down your name and UAB in all sheets.

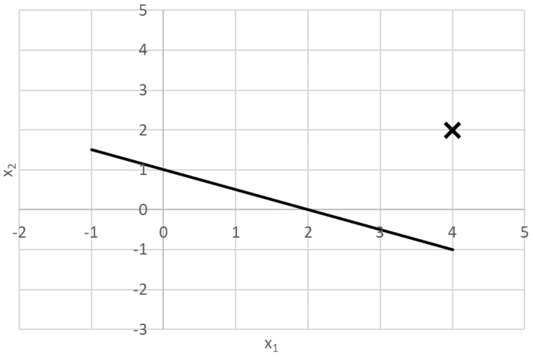
Question 1: **1.0 pt**

Suppose you train a logistic classifier $f_{\mathbf{w}}(\mathbf{x}) = g(w_0 + w_1x_1 + w_2x_2)$, where $g(\quad)$ is the logistic function. After some time, you find the solution: $w_0 = 2, w_1 = -1, w_2 = -2$.

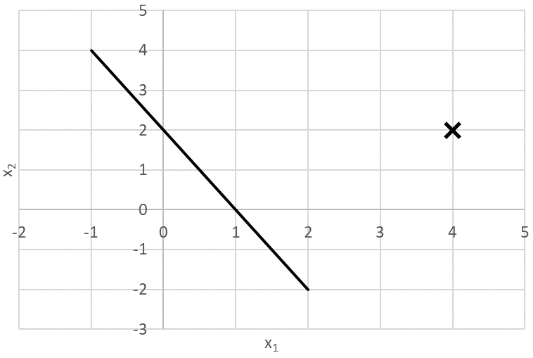
(a) Which of the following figures represents the decision boundary found by your classifier?

(b) What is the class of the point with features (4, 2), marked with an × in the figure?

Explain your answer.



(1)



(2)

RESPONSE:

(a) The correct decision boundary is depicted in (1)
The decision boundary is defined by the line $w_0 + w_1x_1 + w_2x_2 = 0$. Substituting the values of our solution this should be the line $2 - x_1 - 2x_2 = 0$. This corresponds to the line on the left.

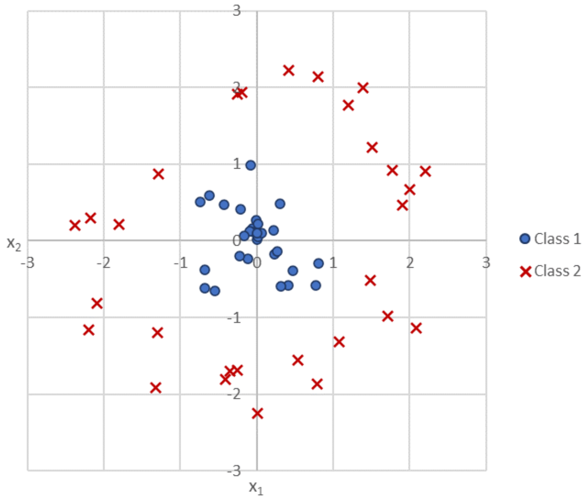
(b) The marked point is of class “0”
We substitute the features of the point to the linear part and we get a negative number: $2 - 4 - 2 * 2 = -6$. This corresponds to class 0, as passing it through the sigmoid would give us a value < 0.5.

Name:____ UniversityID:____________________

Question 2: **1.0 pt**

Suppose you are fitting the following non-linear logistic regression classifier on the dataset on the right.

$f_{\mathbf{w}}(\mathbf{x}) \ = \ g(w_0 + w_1 x_1^2 + w_2 x_2^2)$



Which of the following statements are true? Each wrong answer discounts 0.1 pts

	TRUE	FALSE
At the optimal value of $\mathbf{w}$ the value of the cost function will be $J(\mathbf{w}) \ \geq \ 0$	X	
The trained classifier will reach an accuracy of 100% on the training set	X	
The trained classifier will give $f_{\mathbf{w}}(\mathbf{x}^{(i)}) \leq \ 1$ for the points of class 1 (dots) and $f_{\mathbf{w}}(\mathbf{x}^{(i)}) \geq \ 1$ for the points of class 2 (crosses) of the training set.		X
The features we are using are not enough; we need to create more features to reach a good result in this scenario.		X
The positive and negative samples cannot be separated using a straight line. So, applying logistic regression as described here will not be able to find a good solution		X

Question 3: **1.0 pt**

Which of the following statements about regularization are true? Each wrong answer discounts 0.1 pts

	TRUE	FALSE
Given that logistic regression already bounds the output values in $[0,1]$, regularization in general is not necessary when we do logistic regression		X
Adding an L2 regularization term to the “Mean Squared Error” cost function, makes it a non-convex function		X
The “Lasso” (L1) regularization can act as a model selector, as it can force certain parameters of the model to become exactly equal to zero	X	
Fixing the regularization coefficient $\lambda \ = \ 0$, is the same as not applying any regularization at all	X	
Consider a classification problem. Adding regularization may cause your classifier to incorrectly classify some training examples (which it had correctly classified when not using regularization)	X	

Name:____ UniversityID:____________________

Question 4: 4 pts

Each wrong answer discounts 1/4 of the value of the question.

1 The test of hypothesis allows …

a) To reject the Null hypothesis.
b) To validate the proposed hypothesis.
c) To compare different types of hypothesis
d) Non of above.

2 The linear SVM can be viewed as a ....

a) Linear optimisation problem subject to linear constraints.
b) Linear optimisation problem subject to non-linear constraints.
c) Quadratic optimisation problem subject to non-linear constraints.
d) Quadratic optimisation problem subject to linear constraints.

3 Why do we use the slack variables in a SVM?

a) To allow margin violations.
b) To allow multiple classes.
c) To allow a non-linear definition of the separating hyperplane.
d) To project the original feature space into a higher dimensional space.

4 Which interpretation for the support vectors in the SVM is correct?

a) The vectors supporting all the correct classification.
b) The reference points for the minimal margin solution.
c) The vectors supporting all the misclassifications.
d) The reference points for the maximal margin solution.

5 Why is the kernel trick useful?

a) To avoid the explicit calculation of the slack variables.
b) To avoid the explicit calculation of the higher dimension functions.
c) To avoid the explicit calculation of the support vectors.
d) To avoid the explicit calculation of the error function.

6 In boosting algorithms the learners one would train …

a) Should be real time performing.
b) Should be as accurate as possible, in order to get the asymptotical convergence to the Bayesian error.
c) Should be random classifiers, in order to increase the diversity.
d) Should be at least better than random.

7 The local receptive field (LRF) of a CNN

a) Allows to have a generalisation for every pixel of the image.
b) Allows to treat individual pixels.
c) Allows to connect only a small region of the image.
d) None of above

8 The max poling technique provides

a) A linear processing of the images.
b) A nonlinear processing of the images.
c) An exponential processing of the images.
d) None of above

Name:____ UniversityID:____________________

Question 5: 0.8 pt

You are training a deep MLP, with 10 hidden layers and sigmoid (logistic) activations. You note that the weights of the first layers hardly move during the training process. Which of the following statements are True? Each wrong answer discounts 0.1 pts

	TRUE	FALSE
We are experiencing a problem due to saturated units		X
We are experiencing a vanishing gradient problem	X	
Introducing an intermediate loss would improve the situation	X	
Changing the activations to ReLU would solve the problem	X	

Question 6: 1.2 pt

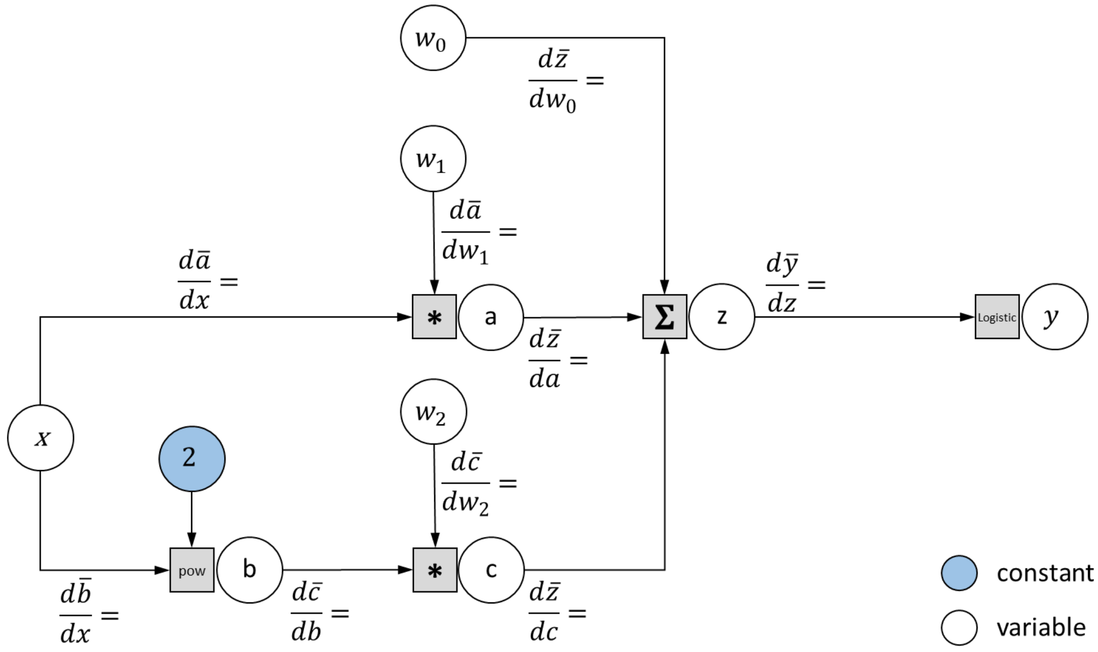
Neural networks are universal approximators. What does this mean exactly? Each wrong answer discounts 0.1 pts

	TRUE	FALSE
A feedforward, fully-connected neural network with a single hidden layer can represent any continuous function to an arbitrary precision	X	
In addition to the previous statement, the derivatives of a feedforward, fully-connected neural network with a single hidden layer can represent the derivatives of any continuous function to an arbitrary precision	X	
A feedforward, fully-connected neural network with a single hidden layer can represent any Boolean function exactly	X	
The fact that neural networks are universal approximators has high practical significance, as it gives us a guarantee that we can find a good solution for any problem		X
Adding more neurons at the hidden layer can cause out model to overfit	X	
To add expressive power to our fully-connected feed-forward network we can either add more neurons to the hidden layer, or add depth by adding more hidden layers.	X	

Marking instructions: a saturated unit would also kill the gradient, although we should have a whole layer (at least many) of saturated units to see a generalized effect as the one described here. Here the focus is put on the depth, saying that there are 10 hidden layers.

Question 7: 2.0 pt

Given the following computation graph for the expression $y = \sigma(w_0 + w_1 x + \ w_2 x^2)$, where $\sigma()$ is the logistic function:



(a) Fill in the gaps with the expressions for the local derivatives (you can do that directly on the figure if you want) [0.9]

(b) Do the forward pass and calculate y for the following values: $x = 2.0$, $w_0 = w_1 = w_2 = 0.0$ [0.1]

(c) Apply back propagation and calculate the derivatives $\frac{dy}{dw_0}, \frac{dy}{dw_1}, \frac{dy}{dw_2}$ and $\frac{dy}{dx}$ for that point [1.0]

Remember that the derivative of the logistic function is given by $\sigma'(z) = \sigma(z)(1 - \sigma(z))$

Response:

(a) **Marking instructions**: 0.1 for each correct local derivative, total of 0.9

$$\frac{d\bar{y}}{dz} = y(1-y) \quad \frac{d\bar{z}}{dw_0} = 1 \quad \frac{d\bar{z}}{da} = 1 \quad \frac{d\bar{z}}{dc} = 1 \quad \frac{d\bar{a}}{dw_1} = x \quad \frac{d\bar{a}}{dx} = w_1 \quad \frac{d\bar{c}}{dw_2} = b \quad \frac{d\bar{c}}{db} = w_2 \quad \frac{d\bar{b}}{dx} = 2x$$

(b) $y = \sigma(0) = 0.5$ **Marking Instructions**: 0.1 for correct answer.

(c) **Marking instructions**: 0.25 for each correct local derivative, total of 1.0

$$\frac{dy}{dw_0} = \frac{d\bar{y}}{dz}\frac{d\bar{z}}{dw_0} = y(1-y) * 1 = 0.25 * 1 = 0.25$$

$$\frac{dy}{dw_1} = \frac{d\bar{y}}{dz}\frac{d\bar{z}}{da}\frac{d\bar{a}}{dw_1} = y(1-y) * 1 * x = 0.25 * 1 * 2 = 0.5$$

$$\frac{dy}{dw_2} = \frac{d\bar{y}}{dz}\frac{d\bar{z}}{dc}\frac{d\bar{c}}{dw_2} = y(1-y) * 1 * b = 0.25 * 1 * 4 = 1.0$$

$$\begin{aligned}\frac{dy}{dx} &= \frac{d\bar{y}}{dz}\frac{d\bar{z}}{da}\frac{d\bar{a}}{dx} + \frac{d\bar{y}}{dz}\frac{d\bar{z}}{dc}\frac{d\bar{c}}{db}\frac{d\bar{b}}{dx} \\ &= y(1-y) * 1 * w_1 + y(1-y) * 1 * w_2 * 2x \\ &= 0.25 * 1 * 0 + 0.25 * 1 * 0 * 4 \\ &= 0\end{aligned}$$

Question 8: 1 pt

Can we shuffle two contiguous layers of a resnet (switch one with the other)? Why? Why not?

Question 9: 1 pt

About classification architecture, which of the following statements are True? Each wrong answer discounts 0.1 pts

	TRUE	FALSE
Stochastic depth networks train faster than resnets.	X	
Resnets and highway networks introduce architectural modifications to improve information and gradient flow.	X	
Resnets act as ensembles of relatively shallow networks because of the relu activations		X
Capsule networks address the problem of gradient vanishing		X
Moco and SimCLR are contrastive learning methods	X	

Question 10: 1 pt

Briefly explain what is the rationale behind randomly initializing the parameters in convolutional neural networks (CNNs). Explain also the difference between the simple random initialization and Xavier (Glorot) and He (Kaiming) initializations.

Question 11: 1 pt

Briefly explain what are adaptive learning rate optimization algorithms, and how do they differ from traditional learning rate schedules in deep learning? How can one determine which adaptive learning rate algorithm to use for a particular task or dataset?

Question 12: 1 pt

About efficient NN, which of the following statements are True? Each wrong answer discounts 0.1 pts

	TRUE	FALSE
Reducing the number of parameters of a neural network always implies a reduction of its computational cost.		X
Weight pruning can remove 90% of the network’s weight while maintaining the original performance.	X	
The Dense Sparse Dense technique can improve the performance of a neural network while keeping the same number of parameters.	X	
Structured weight pruning cannot speed-up a convolutional neural network.		X
The lottery ticket hypothesis says that increasing the number of parameters of a network always leads to better performance.		X

Question 13: 0.5 pt

Is it better to perform low-rank approximation by reconstructing the network filters or feature maps? Why?

Feature maps. Approximating the feature maps gives better performance because it reconstructs the results of the convolution instead of trying to reconstruct the filters and the final aim is a good approximation of the convolution at each layer to not change the classification results. However, approximating filters does not require data and it's therefore easier and faster.

Question 14: 0.5 pt

What is conditional computation?

It’s adapting the computation depending on the given input. Less computation for simple examples and more for difficult

Question 15: 1 pt

The following articles describe explainability techniques to better understand neural networks. For each article, provide the name of the explainability category that they belong to, and whether they are considered Post-hoc analysis or Ad-hoc modelling.

Inverting Visual Representations with Convolutional Networks
We propose a new approach to study image representations by inverting them with an up-convolutional neural network. We apply the method to shallow representations (HOG, SIFT, LBP), as well as to deep networks.

Neuron Analysis, Post-hoc

Understanding Black-box Predictions via Influence Functions
In this paper, we use influence functions to trace a model's prediction through the learning algorithm and back to its training data, thereby identifying training points most responsible for a given prediction.

Data Inspection, Post-hoc

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization
Our approach - Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in the image for predicting the concept.

Saliency Based, Post-hoc

Neural network explanation using inversion
We then present “HYPINV” a new explanation algorithm which relies on network inversion; i.e. calculating the NN input which produces a desired output. HYPINV is a pedagogical algorithm, that extracts rules, in the form of hyperplanes.

Proxy model, Post-hoc

Deep Visual-Semantic Alignments for Generating Image Descriptions
We present a model that generates natural language descriptions of images and their regions. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data.

Modifications, Post-hoc

Human-in-the-Loop Interpretability Prior
In this work, we optimize for interpretability by directly including humans in the optimization loop. We develop an algorithm that minimizes the number of user studies to find models that are both predictive and interpretable and demonstrate our approach on several data sets.

Name:____ UniversityID:____________________

InterpretableRepresentation, Ad-hoc

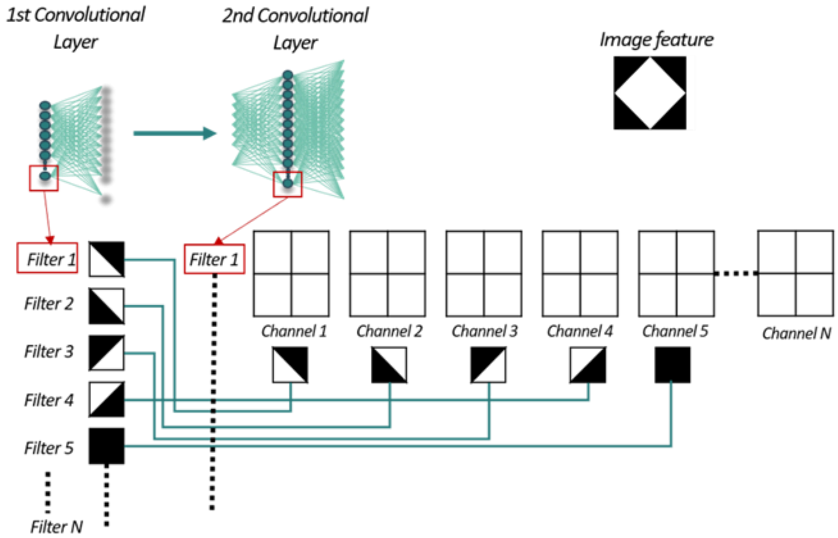
Gaining Free or Low-Cost Interpretability with Interpretable Partial Substitute

Our solution is to find an interpretable substitute on a subset of data where the black-box model is overkill or nearly overkill while leaving the rest to the black-box. This transparency is obtained at minimal cost or no cost of the predictive performance.

Model Renovation, Ad-hoc

Question 16: 1 pt

Regarding the hierarchical composition of convolutions for shape representation in CNNs, and given the learnt filter for the 1st convolutional layer, indicate which filter weights (0 or 1) should have a filter in the 2nd convolutional layer to be selective to the image feature given at the right in the next figure



Correct answer: (from left to right and top to bottom)
Channel 1: 0001 / Channel 2: 1000 / Channel 3: 0100 / Channel 4: 0010 / Channel 5: 0000