

**Module: M3. Machine learning for computer vision****Final exam**Date: February 17th, 2020

- Books, lecture notes, calculators, phones, etc. are not allowed.
- All sheets of paper should have your name.

Question 1: 0.8

You are trying to fit a polynomial model to your data. Which of the following can help in preventing overfitting? Check all that apply.

	TRUE	FALSE
Fit small degree polynomials to the data.		
Constrain the coefficients of the polynomial to have small values.		
Get more data points.		
Fit the data with a polynomial of degree greater than the number of training points.		

Question 2: 0.4

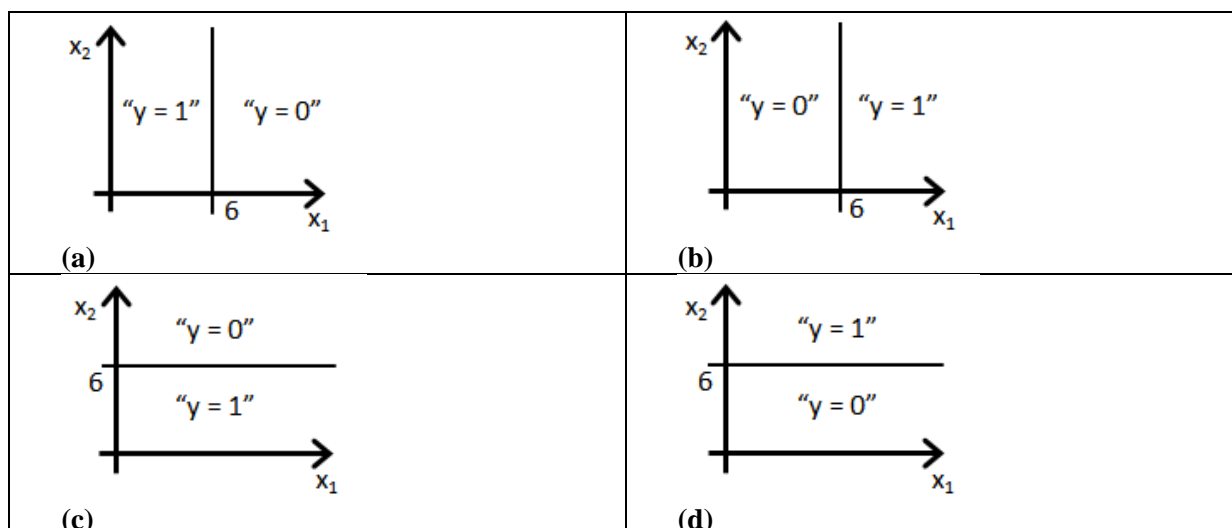
Suppose you trained two logistic regression models, one using regularization parameter $\lambda=0$ and the other using $\lambda=1$. You have the results from both runs, but you have forgotten which one corresponds to which experiment. Can you tell which set of parameters corresponds to the run with $\lambda=1$? Circle your answer

(a) $\theta = \begin{bmatrix} 15.52 \\ 78.34 \end{bmatrix}$

(b) $\theta = \begin{bmatrix} 0.96 \\ 10.52 \end{bmatrix}$

Question 3: 0.8

Suppose you train a logistic classifier $h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suppose also that after a while you find the solution: $\theta_0 = 6, \theta_1 = 0, \theta_2 = -1$. Which of the following figures represents the decision boundary found by your classifier? (Circle your answer)



Name_____

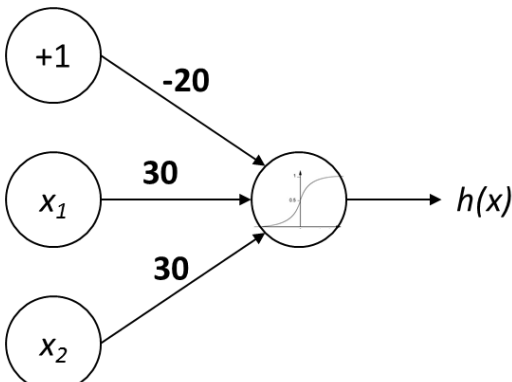
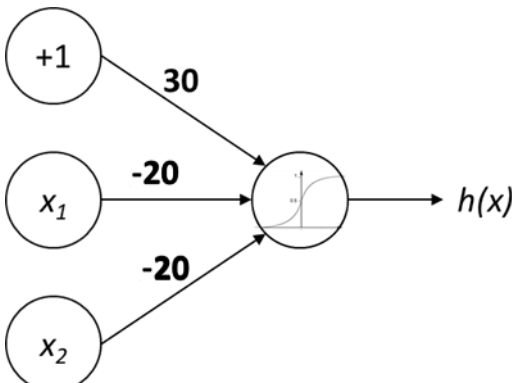
Question 4: 1.0

You are training a classification model with logistic regression. Which of the following statements are true?

	TRUE	FALSE
Adding many new features to the model helps prevent overfitting on the training set		
Introducing regularization to the model always results in equal or better performance on the training set		
Adding a new feature to the model always results in equal or better performance on the training set		
Introducing regularization to the model always results in equal or better performance on examples not in the training set.		
Logistic regression's weights w should be initialized randomly rather than to all zeros, because if you initialize to all zeros, then logistic regression will fail to learn a useful decision boundary because it will fail to "break symmetry".		

Question 5: 1.0

Consider the following neural networks which take two binary-valued inputs $x_1, x_2 \in \{0, 1\}$ output $h(x)$ through a sigmoid output unit. Which of the following logical functions does each network (approximately) compute? Circle the correct answer.

	<ol style="list-style-type: none"> 1. AND 2. NAND (NOT AND) 3. OR 4. XOR (exclusive OR)
	<ol style="list-style-type: none"> 1. AND 2. NAND (NOT AND) 3. OR 4. XOR (exclusive OR)

Question 6: 1.0

A ROC curve has the following behaviour:

- a) Cannot have a slope higher than 45 degrees, which corresponds to the 50% of classification accuracy.
- b) It is monotonically growing and the maximum value of the area under the curve is 1.
- c) The axis of the "Precision" cannot be negative.
- d) None of the above.

Name_____

Question 7: 1.0

A ROC curve has the following behaviour:

- a) Cannot have a slope higher than 45 degrees, which corresponds to the 50% of classification accuracy.
- b) It is monotonically growing and the maximum value of the area under the curve is 1.
- c) The axis of the "Precision" cannot be negative.
- d) None of the above.

Question 8: 1.0

For a given classification problem in a given feature space, the Bayesian error:

- a) Is always the same independently of the type of features used.
- b) Provides us with an asymptotical value for the error.
- c) Can be minimised by means of regularisation strategies.
- d) None of the above.

Question 9: 1.0

A hypothesis test is run to compare the average results from two different classifiers, and the p-value obtained is $p=0.03$. What can we say about both classifiers?

- a) Nothing, the test did not provide enough information.
- b) Both classifiers are indistinguishable in terms of performance.
- c) One classifier gets better results than the other one.
- d) None of the above.

Question 10: 1.0

The kernel trick in a SVM allows for:

- a) Explicitly calculate the features in a higher dimensional space, where the problem is linearly separable.
- b) Allowing some of the samples to violate the maximal margin condition by introducing slack variables.
- c) Classify non-linear problems without introducing any new parameter at all.
- d) None of the above.

Question 11: 1.0

One of the advantages of the boosting algorithms is that:

- a) They allow bias reduction.
- b) They allow variance reduction.
- c) They allow both bias and variance reduction.
- d) None of the above is a feature of boosting algorithms.

Question 12: 1.0

The XOR problem is well known due to the fact that:

- a) It can be solved with Fisher linear discriminant analysis.
- b) It can be solved with a unique perceptron with a sigmoid activation function.
- c) It can be solved with a one linear support vector machine with no kernel.
- d) None of the above.

Name_____

Question 13: 1.0

An overfitting during the training phase of a neural network can bring resulting learning curves which evolve through training epochs in the following way:

- a) The training data converges to zero loss while the validation data converges to zero loss too.
- b) The training data converges to zero loss while the validation data does not converge to zero loss.
- c) Neither the training data converges to zero loss nor the validation data converges to zero loss.
- d) None of the above.

Question 14: 1.0

Select one case in which the use of softmax activation with cross-entropy loss functions is suitable.

- a) A value estimation for non-linear regression.
- b) A performance estimate for cross-validation analysis.
- c) A class-probability estimation for mutually exclusive classes.
- d) None of the above cases will be suitable to use softmax activation with cross-entropy loss function.

Question 15: 1.0

Which of the following operations, frequently used in CNNs, preserves spatial dimensions while reducing the depth?

- a) A one by one (1x1) convolution.
- b) A drop-out.
- c) A max pool.
- d) Non of the above.

Question 16: 1.0

The max pool operation...

- a) Disables a number of units of the CNN.
- b) Introduces redundancy in the CNN.
- c) Reduces the number of parameters of a CNN.
- d) Non of the above.

Question 17: 1.0

Explain the basic intuition of the method of momentum (Polyak, 1964) and how it modifies the standard algorithm of Stochastic Gradient Descent (SGD).

Question 18: 1.0

Explain why it is not a good idea to initialize all weights of a neural network to zero. In the context of weights initialization, what does the concept of "breaking symmetry" mean? How can we implement it?

Name_____

Question 19: 0.5

What are the similarities between Visualization of a network through activation maps or using top-scoring images?

- a) Both methods need input images to visualize the network
- b) Both Methods focus on visualizing single neurons
- c) Both Methods use activation Map information
- d) All the above statements are true

Question 20: 0.5

Which of the following methods allows to visualize individual neurons in a neural network.

- a) Visualizing the network through modification of the activations
- b) Visualizing the network through weight comparison
- c) Visualizing the network through ablation with the lottery ticket hypothesis
- d) Visualizing the network through image generation.

Question 21: 0.5

Why the method based on directly visualizing weights in the CNN is not very used to understand neural networks.

- a) It is the best method for representing the first layer of a model, but it is impossible to get an exact reconstruction into image space of the weights due to irreversible operations
- b) It is useless for all the layers, since weights are giving information about intermediate representations which are not lying in the image space, and consequently they are not directly giving any understandable clue of the kind of image features that are activating that neuron
- c) It needs a huge dataset to visualize each neuron and thus, it is not an optimal method
- d) None of the above statements is true.

Question 22: 0.5

In the frame of Convolutional Neural Networks, briefly define the following two concepts:

Receptive field of a neuron:

Top-scoring images of a neuron:

Question 23: 0.5

Answer true or false

	TRUE	FALSE
Removing a layer in VGG reduces performance by less than 2.25%		
DARTS requires more computation than AMOEBA		
AMOEBA uses genetic algorithms		
Self-supervised learning requires labels during pre-training		

Name_____

Question 24: 0.5

Tell two differences between GoogleNet and previous architectures.

Question 25: 0.5

What is the main difference between a residual network and a highway network?.

Question 26: 0.5

Are stochastic depth networks faster than a traditional convnet during training time? Reason why.

Question 27: 0.5

From the ones explained in class, which techniques can reduce the size of the model?

Question 28: 0.5

What is the difference between magnitude pruning and Optimal Brain Damage?

Question 29: 0.5

What is the difference between xnor networks and binary connect?

Question 30: 0.5

Answer true or false

	TRUE	FALSE
Adaptive ResNet uses a cascade of models to classify an image		
Distillation transfers "dark knowledge" between models since it uses the relationships between classes		
Efficientnet was found by architecture search		
Trained quantization clusters weights by value and assigns binary codes to them. Huffman coding could be used to assign short codes to frequent weights, reducing the model size		