

**Module: M3. Machine learning for computer vision****Final exam**Date: February 18th, 2019

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

Question 1: 0.75

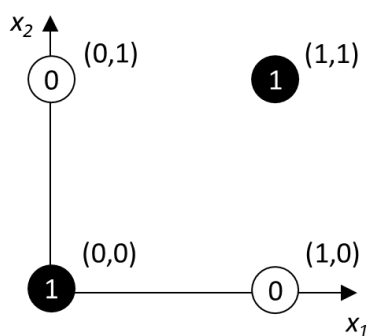
Suppose you have a dataset with $n = 10$ features and $m = 5000$ examples. After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets.

Which of the following might be promising steps to take?

	TRUE	FALSE
Use an SVM with a linear kernel, without introducing new features		X
Use an SVM with a Gaussian Kernel	X	
Create / add new polynomial features	X	
Increase the regularization parameter λ .		X
Reduce the number of examples in the training set		X

Question 2: 0.6

Suppose we are given four training cases:



x1	x2	y
1	1	1
1	0	0
0	1	0
0	0	1

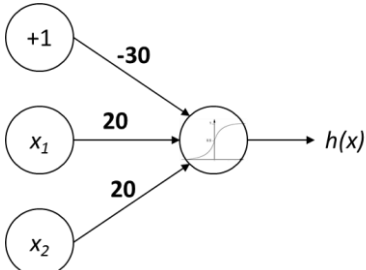
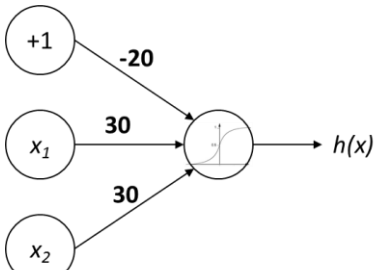
It is impossible for a linear classifier to produce the desired target outputs for all four cases. Now suppose that we add an extra feature x_3 so that each of the four input vectors consists of three numbers instead of two.

Which of the following ways of setting the value of the extra feature will create a set of four input vectors that is linearly separable (i.e. that a linear classifier with appropriate weights and bias can give the target values).

	TRUE	FALSE
Make the extra feature x_3 be the same as the target value y for that input vector	X	
Make the extra feature x_3 be the opposite of the first feature x_1 (i.e. use 1 if the first feature is 0 and 0 if the first feature is 1)		X
Make the extra feature x_3 be 1 for one of the four input vectors and 0 for the other three	X	
Make the extra feature x_3 be the same as the first feature		X

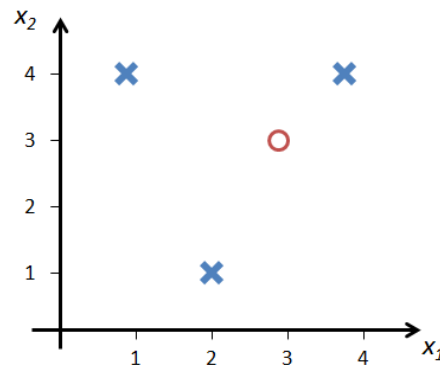
Question 3: 0.4

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 4: 0.5

Suppose you are fitting a logistic regression classifier:
 $h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$ on the dataset on the right. Which of the following statements are true?



x1	x2	y
1	4	1
2	1	1
3	3	0
4	4	1

	TRUE	FALSE
At the optimal value of θ the value of the cost function will be $J(\theta) \geq 0$	X	
If we train gradient descent for enough iterations, for some examples $x^{(i)}$ in the training set it is possible to obtain $h_{\theta}(x^{(i)}) < 0$		X
The cost function $J(\theta)$ will be a convex function, so gradient descent might converge to a local minimum		X
The positive and negative examples cannot be separated using a straight line. So, gradient descent will fail to converge		X
Because the positive and negative examples cannot be separated using a straight line, linear regression will perform as well as logistic regression on this data		X

Name _____

Question 5: 0.75

Which of the following statements about regularization are true?

	TRUE	FALSE
Using too small a value of the regularization parameter λ can cause your hypothesis to underfit the data		X
Because logistic regression outputs values $0 \leq h\theta(x) \leq 1$, it's range of output values can only be "shrunk" slightly by regularization anyway, so regularization is generally not helpful for it		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, i.e. when $\lambda=0$)	X	
Using a very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems		X
Because regularization causes $J(\theta)$ to no longer be convex, gradient descent may not always converge to the global minimum (when $\lambda > 0$, and when using an appropriate learning rate α)		X

Question 6: 1

Which of the following statements about regularization are true?

	TRUE	FALSE
For detection of local features, SIFT and SURF are both based on approximations of the Laplacian of Gaussian.		X
Dense sampling allows reducing the number of local keypoints in comparison to standard feature detectors, such as SIFT.		X
SIFT descriptor is based on computing histograms of gradient orientations in a regular grid over the patch.	X	
One of the differences between SIFT and HOG for feature description is block normalization, which is applied in HOG, but not in SIFT.	X	
Dimensionality reduction helps to reduce the number of training samples required to learn the classification model.	X	

Name_____

Question 7: 2

Which of the following statements about regularization are true?

	TRUE	FALSE
Minibatch based updates in Stochastic Gradient Descent usually offer a good approximation of the true gradient of the loss function over the entire dataset.	X	
Using very small batches can make our model easily overfit the training data.		X
Stochastic Gradient Descent is a scale-free algorithm and thus we do not need any data normalization when training deep neural networks with SGD.		X
The best practice to initialize the weights of a deep neural network is by setting all the initial weights to a constant value near zero.		X
The momentum algorithm accumulates an exponentially decaying average of past gradients and continues to move in their direction.	X	
The method of momentum makes training slower but more stable.		X
The AdaGrad algorithm individually adapts the learning rates of all model parameters by scaling them inversely proportional to the square root of the sum of all the historical squared values of the gradient.	X	
Adaptive learning rate optimization algorithms like Adagrad, RMSProp, Adam, or AdaDelta, always perform better than standard SGD and SGD with momentum.		X
Randomly dropping units (along with their connections) from the neural network during training is a common technique to reduce overfitting.	X	
The early stopping criterion is a technique designed to halt SGD whenever overfitting begins to occur and is usually based on the true underlying loss function, such as the 0-1 loss (for classification problems), measured on a validation set.	X	

Name_____

Question 8: 1

The area under the ROC curve can be used...

1. To assess the general rankings for classifier performance.
2. To assess the particular behavior of a classifier for a specific working point.
3. To assess the precision of a classification system along all the working points.
4. None of the above.

Question : 1

When we are tackling a classification problem.

1. The ratio between the dimensionality and the number of samples is limiting our performance error (curse of dimensionality).
2. The Bayesian error is limiting our performance error.
3. The computational cost is limiting our performance error.
4. None of the above.

Question 10: 1

In a hypothesis test,

1. we create a null-hypothesis which has to be validated with a p-value > 0.05 .
2. we create a null-hypothesis which has to be validated with a p-value < 0.05 .
3. we create a null-hypothesis which has to be rejected with a p-value > 0.05 .
4. we create a null-hypothesis which has to be rejected with a p-value < 0.05 .

Question 11: 1

In order to allow some vectors to violate the maximum margin condition, the SVMs introduce...

- 1 The kernel trick.
- 2 The Lagrange multipliers.
- 3 The slack variables.
- 4 None of the above.

Question 12: 1

With respect to the mathematical solution of the SVM, the support vectors correspond with:

1. The incorrectly classified samples.
2. The vectors with associated non-zero Lagrange multipliers.
3. The vectors with associated slack variables greater than zero.
4. The vectors with associated non-zero regularization factor.

Question 13: 1

The types of classification problems, such as the XOR problem:

1. Can be solved with a linear support vector machine with no kernel.
2. Can be solved with a perceptron with a sigmoid activation function.
3. Can be solved with Fisher linear discriminant analysis.
4. None of the above.

Question 14: 1

The dropout technique in NN is fundamentally used:

1. To reduce the computational cost of the validation process.
2. To allow for a greater generalization power of the trained NN.
3. To increase the learning rate.
4. None of the above.

Name_____

Question 15: 1

Loss functions, such as mean square error, can provide no sufficient slope for gradient descent in specific cases. In order to overcome this for the particular case of classification of mutually exclusive classes, we can suggest:

1. To perform maximum dropout and adaptive learning rate.
2. To use sigmoid activation functions exclusively.
3. To use softmax activation with cross-entropy loss functions.
4. None of the above strategies will be suitable to solve the mentioned problem.

Question 16: 1

A one by one (1x1) convolution is frequently used in CNNs to...

1. Preserve depth will reducing spatial dimension.
2. Reduce both depth and spatial dimensions.
3. Preserve spatial dimensions while reducing the depth.
4. Non of the above.

Question 17: 1

The max pool operation.

1. Decreases the number of parameters of a CNN.
2. Introduces redundancy in the CNN.
3. Disables a number of units of the CNN.
4. Non of the above.

	8	9	10	11	12	13	14	15	16	17
1)	X									X
2)		X			X		X			
3)				X				X	X	
4)			X			X				

Name_____

Question 18: 1

What is the problem of traditional Convolutional Neural Networks? What is the solution proposed in Capsule Networks?

Question 19: 1

What is the main benefit of residual connections in DenseNet, ResNet or Highway networks?

Name_____

Name_____

Question 20: 1

According to what we can conclude from the approaches that try to understand CNNs, answer if the following sentences are TRUE or FALSE:

	TRUE	FALSE
The feature map obtained by a single neuron has only one channel.	X	
Neurons of a CNN are not related on the intermediate representation of an image through such CNN.		X
Top-scoring images of a neuron are the set of images that mostly activates a specific neuron.	X	
Neurons of shallower layers of a hierarchical CNN are devoted to simple image features.	X	
The receptive field of a Fully-connected layer is smaller than the input original image		X
A high pixel value of a feature map means of a specific neuron means that the codified feature by this neuron is found in the image.	X	
Details of the images are encoded in deeper layers while a global visualization of the image is represented in shallower layers.		X
The main problem of describing the neuron activity by generating a new image that maximizes the activation of a specific neuron is that the result may be a non-realistic image. These approaches require good regularizations.	X	
Visualizing directly neuron weights is an easy and good approach to understand what has been learned by the CNN.		X
Image feature complexity decreases from shallow to deep layers.		X

Question 21: 1

Understanding what has been learned on a trained CNN can be done in different ways. One of them is to project feature maps into the image space in order to visualize what is represented in a specific layer. This technique can be used for describing either a single neuron (typically from the top scoring images) or a entire layer representation. Can you describe the main ideas of these approaches?

Name_____

Name_____

From the next 3 questions, choose 2 of them

Question 22: 1

It is frequent to batch several input images for doing a joint inference using a single CNN. Describe the effect of the size of the mini-batch (or batch) on the inference time.

Question 23: 1

Describe the two main options for using two or more GPUs for training a single CNN faster, and the advantages and disadvantages of each one.

Name_____

Question 24: 1

Name and describe very briefly two special H/W designs for fast inference and explain why are them more suitable for NN inference than general-purpose CPUs.

Name _____

8 Sol: 1. The AUC is used to have a global overall view of the classification system, since it integrates information for all the possible working points regarding sensitivity vs fp-rate.

9 Sol: 2. The Bayesian error is asymptotically bounding the performance error of any given classifier, since it incorporates the intrinsic indistinguishability of the data.

10 Sol: 4. The null-hypothesis is the equivalent of “nothing changes”, so it has to be rejected. The p-value indicates how possible is that this rejection can be a matter of chance, and the scientific community agrees as acceptable below a 5% of possibilities of chance.

11 Sol: 3. Slack variables allow for some vectors to violate the maximum margin condition, allowing in this way to use the SVM for non separable classification problems.

12 Sol: 2. The SVM formulation implies the resolution of the quadratic optimization problem, solved using the Lagrange multipliers and a dual problem. The solution is a linear combination of the vectors with non-zero Lagrange multipliers (i.e., the support vectors).

13 Sol: 4. The non-linear nature of the XOR-like problems makes it not feasible for the techniques mentioned above to provide a solution for the classification problem.

14 Sol: 2. By disabling a number of units, during the forward or backwards phases of the training, the network is forced to readapt and ignore potential local minima, increasing in this way its generalization power and avoiding potential overfitting.

15 Sol: 3. Softmax activation with cross-entropy loss functions provides a view of the class-probability, which is particularly suited for mutually exclusive classes.

16 Sol: 3. 1x1 Convolutions preserve the spatial dimensions of the outputs, but reduces the final depth through a combination of the feature maps.

17 Sol: 1. The max pool operation fundamentally reduces the number of parameters within the model, while provides a higher generalization power for the convolutional filters.

18 Answer: Traditional CNNs do not preserve the **relative position** between image features, specially when using **max pooling**. Capsule networks **encode feature pose information in vectors** instead of just encoding the presence of a feature with a scalar.

19 Answer: They allow to train **deeper** models by adding a **shortcut paths** to the gradient.

20 Answer: The idea is to invert the effects of any layer of a specific trained CNN. First, let us be focused on a specific neuron. The idea is, then, to understand what this neuron is selecting from an image. For this visualization, the feature map obtained by this specific neuron when an input image is analyzed in a forward way, is taken to be projected into the image space by inverting each previous layer. Zeiler proposed to invert convolutional layers using the deconvolution operation, keeping the ReLU effects on the projection and using switches for inverting pooling layers. Later, Symonian improved that method by combining the ReLU operation with the switches to invert the RELU layers. Second, for inverting the layer representation, two main ideas have been proposed: generation an image whose feature map best matches a given feature map or train a deconvolutional network to invert such feature maps.

22 Answer: Increasing the batch size may increase inference latency (seconds) slightly but improves the inference throughput (images/second). The reason is an improvement of the arithmetic intensity (higher ratio of computation versus accesses to slow memory). Weights are read from slow memory into fast memory and can be reused for each of the elements in the batch, affording expensive memory accesses.

Comment: the effect of using batches on training is more complex (affects learning) and is not the question.

Keys: Differentiate between training and inference. Distinguish between the latency of the inference (seconds) and the throughput (images inferred per second). Identify the memory hierarchy and the arithmetic intensity as key performance factors.

22: Answer: Data parallel training distributes the data among GPUs, and each GPU must hold/access all the weights of the complete NN. Coordinated weight update is required (can be synchronous or asynchronous). The gradients must be communicated. The GPUs contain redundant data (weights) and very large models may not fit into memory. An asynchronous update affects the learning task.

Model parallel training distributes the model (weights) among the GPUs. The activations must be communicated among GPUs, but the amount of memory required per GPU is reduced, which allows using models larger than the memory available on a single GPU.

Keys: Partitioning weight or input data. Effect on GPU memory requirements and communication requirements.

23 Answer: Tensor cores on GPUs accelerate matrix multiplication with reduced precision real numbers (16-bit multiplication and 32-bit addition)

Google TPUs include similar matrix-multiplication H/W with 8-bit numbers and special H/W for activation functions. NVIDIA Deep Learning Accelerator includes H/W for Convolution operations.

The most important advantage is lower energy consumption due to a reduction of data movement from/to general memory, and a reduction on the size of the operands.

Keys: Identify Matrix-Multiplication and convolution as the main tasks that need acceleration.

Recognizing that energy consumption is the hard performance limiter.