Question 1: Short Questions

(a) We use a support vector machine (SVM) with a soft-margin SVM to perform a classification task. Which of the following types of samples will have zero slack variables ξ_i?

(i) All support vectors

(ii) All correctly classified samples

(iii) All misclassified samples

(iv) All samples lying within the margin

(i) is correct. Correctly classified samples might still lie within the margin, therefore the slack variable will be non-zero. Misclassified samples have slack variables greater than zero, since they contribute to the soft error.

(b) Explain in which of the following cases the risk of overfitting a network decreases.

(i) Regularizing the weights

(ii) Increasing the number of the hidden layers

(iii) Using dropout to train a deep neural network

(iv) Getting additional training data that are very similar to the training data that have been seen before

(i) and (iii) are correct. Increasing the number of hidden layers will result in more complex model and more weights that need to be learned, therefore the risk of overfitting increases. If we use additional training data that are very similar to the ones already observed, we do not get any additional information, therefore the risk of overfitting remains the same.

(c) You have a neural network with two inputs $x_1 = 2$, $x_2 = 2$, connected to the output with two weights $w_1 = 0.5$ and $w_2 = -0.2$. The bias term of the input is $b_1 = 0.1$. You use three different activation functions and get the following output for each function $\{0.7, 0.67, 1\}$. Which type of activation function was used for each of the three outputs?

(i) (linear, indicator/step, sigmoid)

(ii) (ReLU, sigmoid, indicator/step)

(iii) (ReLU, indicator/step, sigmoid)

(iv) (indicator/step, sigmoid, linear)

(ii) is the correct answer. The output node before the activation function is $2\times0.5-2\times0.2+0.1 = 0.7$. When passing a linear or ReLU function to this, the result remains 0.7. When passing an indicator or step function, the result becomes 1. When passing a sigmoid function the result becomes 0.67.

(d) Please select the correct answer(s).

(i) It is possible to successfully train a network by initializing all the weights to zero

(ii) It is not possible to successfully train a network by initializing all the weights to zero

(iii) It is possible to successfully train a network by initializing all the biases to zero

(ii) is the correct answer. If all weights are zero, there is not chance that the neural network learns something, since all resulting transformations will be zero. For (iii), if all biases are zero, then there is still chance that the neural network learns something (e.g., if the data are centered around zero).

(e) The number of nodes in the input layer is 10 and the hidden layer is 5. The maximum number of connections from the input layer to the hidden layer is:

(i) 50

(ii) 10

(iii) 5

(iv) 55

(iv) is the correct answer. We have $5 \times 10 = 50$ weights and 5 bias terms.

(f) We perform a convolution operation to an input image of size 28×28 using a kernel/filter of size 7×7 with a stride of 1. What will be the size of the resulting matrix if we assume that there is not zero-padding at the boundaries of the image?

(i) 22×22

(ii) 28×28

(iii) 21×21

(iv) 7×7

(i) is the correct answer. We perform the convolutional kernel with the first pixel starting from row/column 1 until row/column 22, increasing the step size by 1. After row/column 22 the 7×7 kernel will not have full overlap with the image, therefore we cannot proceed.

(g) Which of following activation function is the most suitable at output layer of a neural network to classify an image in a binary classification task?

(i) Sigmoid

(ii) ReLU

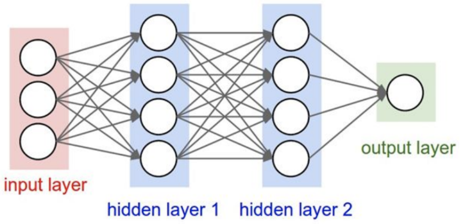
(iii) Linear

(iv) None of the above

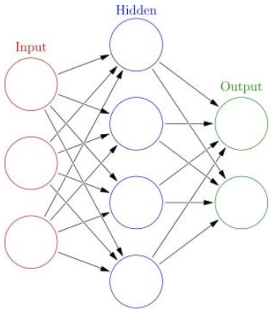
(i) is the correct answer. The sigmoid activation function is the only one of the above which gives an output between 0 and 1, which can be used as a likelihood for the classification task.

(h) For which type of activation function in the nodes of the hidden layers would Architecture 1 be equivalent to Architecture 2?

1.



2.



(i) Sigmoid

(ii) Hyperbolic tangent

(iii) Linear

(iv) ReLU

(iii) is correct. If the activation of the hidden layers is linear, then the transformation implemented by the two layers is equivalent to the transformation resulting when these two are combined.

Question 2: Maximum likelihood estimation

The voters in a given town arrive at the place of voting according to a Poisson process of rate λ voters per hour, where $\lambda = 1 - (\theta t - 1)^2$ and $t = 1, \ldots, 12$. Using the Poisson distribution, we can express the probability of x voters coming to the poll within each hour using the following equation $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$. We further assume N samples $\mathcal{X} = \{(t_1, x_1), \ldots, (t_N, x_N)$ that represented the number of voters x_n that came to the poll at time t_n.

(a) Compute the likelihood of sample (t_n, x_n).
$p(t_n, x_n) = \frac{e^{-1+(\theta t_n - 1)^2}[1-(\theta t_n - 1)^2]^{x_n}}{x_n!}$

(b) Compute the likelihood of all samples $l(\mathcal{X})$.

$$l(\mathcal{X}) = \prod_{n=1}^{N} p(t_n, x_n) = \prod_{n=1}^{N} \frac{e^{-1+(\theta t_n - 1)^2}[1-(\theta t_n - 1)^2]^{x_n}}{x_n!}$$

(c) Compute the log-likelihood of all samples $logl(\mathcal{X})$.

$$\begin{aligned} logl(\theta) &= log \prod_{n=1}^{N} \frac{e^{-1+(\theta t_n - 1)^2}[1-(\theta t_n - 1)^2]^{x_n}}{x_n!} \\ &= \sum_{n=1}^{N} \left[-1 + (\theta t_n - 1)^2 + x_n log\left(1 - (\theta t_n - 1)^2\right) - log\left(x_n!\right)\right] \\ &= -N + \sum_{n=1}^{N} (\theta t_n - 1)^2 + \sum_{n=1}^{N} x_n log\left(1 - (\theta t_n - 1)^2\right) - \sum_{n=1}^{N} log\left(x_n!\right) \end{aligned}$$

(d) Describe how you would find the maximum likelihood estimate of θ given the samples $\mathcal{X}$.
We would compute the derivative of the log-likelihood with respect to θ.

$$\frac{\vartheta logl(\theta)}{\vartheta\theta} = \; = \sum_{n=1}^{N} 2t_n(\theta t_n - 1) + \sum_{n=1}^{N} x_n \frac{-2(\theta t_n - 1)t_n}{1 - (\theta t_n - 1)^2}$$

We can find the MLE of θ, θ^{MLE}, with approximate methods, such as gradient descent. We notice that the inverse of *theta*, $\frac{1}{\theta^{MLE}}$ represents the time in which the voter arrival rate is maximal.