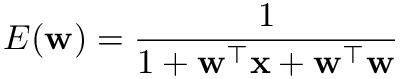
**Deep Learning Weekly MCQ**

**Handout 01**

Q1. If vector {bf w} is of dimension 3 times 1 and matrix {bf A} of dimension 5 times 3, then what is the dimension of left({bf w}^{top}{bf A}^{top}{bf A}{bf w}right)^{top}?

1x1

Q2. Consider the following loss function: 

What is the expression of its gradient frac{partial E}{partial {bf  w}}?



Q3. Which of the following models with input x_1,x_2, parameters w_1,w_2 and noise , are linear in the parameters and can be used for Least Squares (mark all models that are linear):

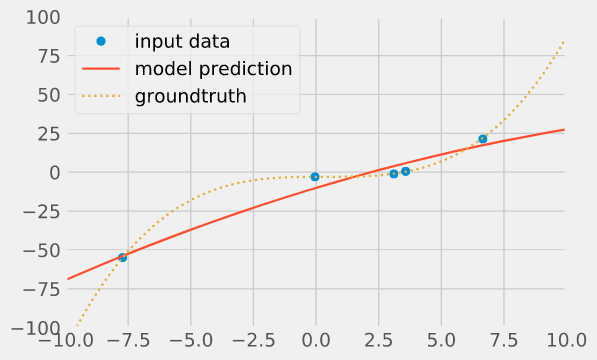


epsilonsim mathcal{N}(0,sigma^2)

Q5. Given a linear model {bf y} = {bf X}{bf w} + boldsymbol{epsilon}, we also want sum_{j=0}^p w_j b_j to be small. What is the expression that we need to minimise to obtain the optimal parameters {bf hat{w}}?



Q6. Consider the Least Square model prediction in the Figure below.



Which of the following statements are correct? (mark all correct)

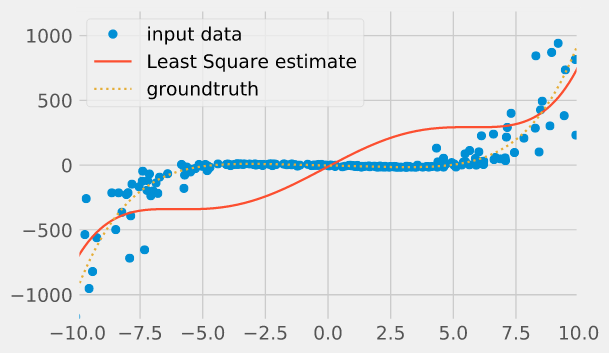
> this is a case of underfitting

Q7. Which of the following facts are true (mark all statements that are true):

> Regularisation biases the parameters towards the value 0

> Regularisation is usually used to reduce overfitting

Q8. We have a dataset and we want to fit a polynomial model of order 7 using Least Squares.



What are (is) the possible reason(s) why the estimated prediction model seems so poor? (mark all correct answers)

> The error is not independent and Gaussian

**Handout 02**

Q1. What is Logistic Regression used for?

> classification

Q2. Given weights w_0=0.1, w_1=1, w_2=2. What is the probabilty p of that observation with feature values x_1=0.3, x_2=0.4, belongs to class 1?



Q3. What is the loss function of the Logistic Regression?

> none of these

Q4. What is the range of possible values for the logit?

> between -infty and +infty

Q5. What do large values of the negative log-likelihood statistic indicate? (select all correct answers)

> That the statistical model is a poor fit of the data.

Q6. What does Logistic Regression assume? (select all correct answers)

> Linear relationship between feature variables and the logit.

Q7. The logistic model is estimated by way of?

> Maximum Likelihood Estimation

Q8. What is the probability corresponding to a logit of 0?

> 0.5

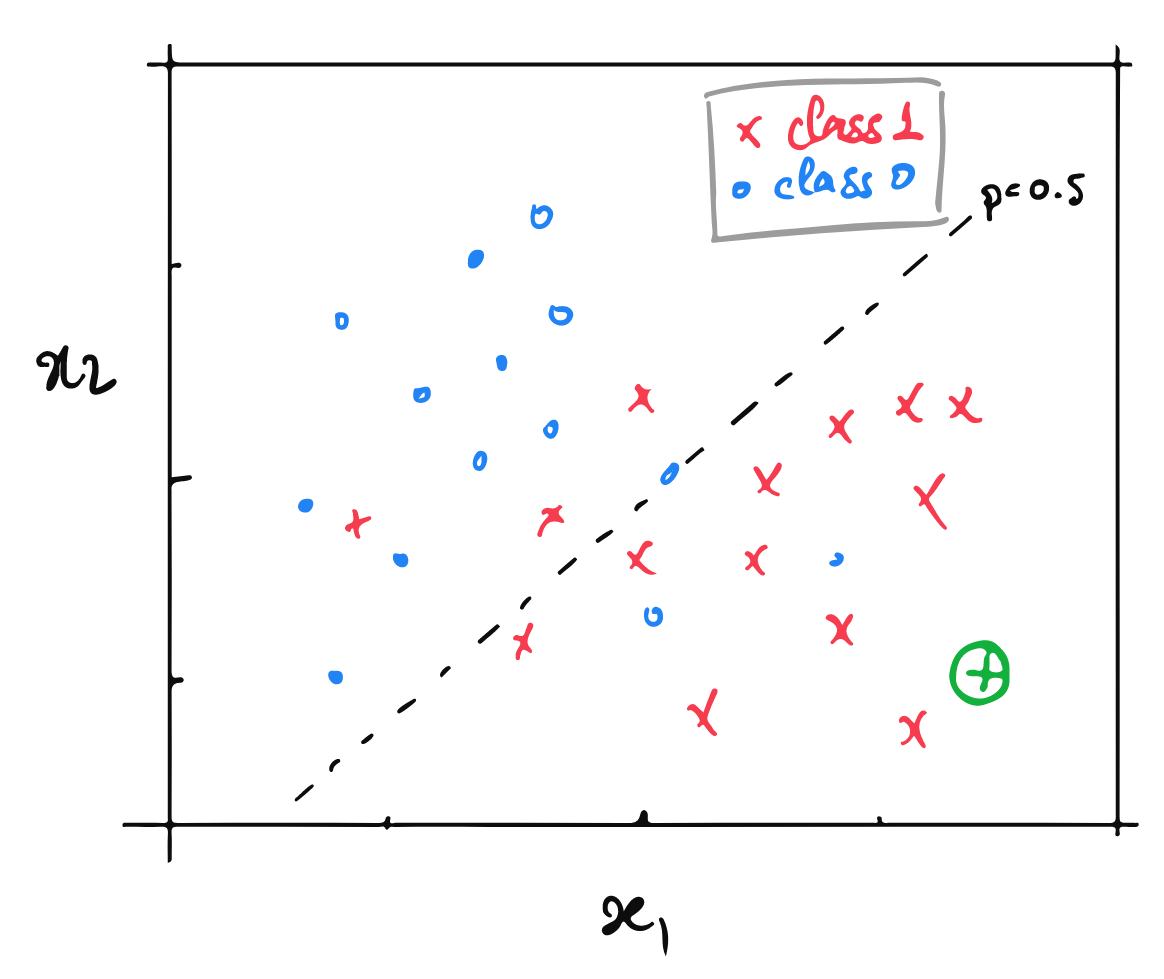
Q9. Which variant of Logistic Regression is recommended when you have a categorical dependent variable with more than two values?

> Multinomial Logistic Regression

Q10. Which of the following option is true? (select all correct answers)

> Linear Regression errors values have to be normally distributed, but, in case of Logistic Regression it is not the case.

Q11. Consider the scatter plot below for a 2-class problem (red cross: class 1, blue circle: class 0).



What is the expected logit value for the new observation (circled green cross)?

> High

Q12. Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic Regression will take very long time to train. Regardless of the final performance, what strategy will speed up the gradient descent optimisation?

> **Increase** the learning rate and **decrease** the number of iteration

**Handout 03**

Q1. What datasets are best suited for kernel methods?p

> datasets with small n and large

Q2. When are Machine Learning methods usually less effective? (select all correct answers)

> When the classes do mix (ie. dataset contains overlapping points).

Q3. Suppose you are using SVM with linear kernel of polynomial degree 2. You find that the corresponding training and testing accuracy is 100% (ie. all the observations in the dataset are correctly classified). You now increase the polynomial degree d of this kernel.

What would you think will happen? (select all correct answers)

> increasing d may overfit the data

Q4. Which of these following statements about kernel functions are true? (select all correct answers)

> The kernel function maps low dimensional data into a high dimensional space.

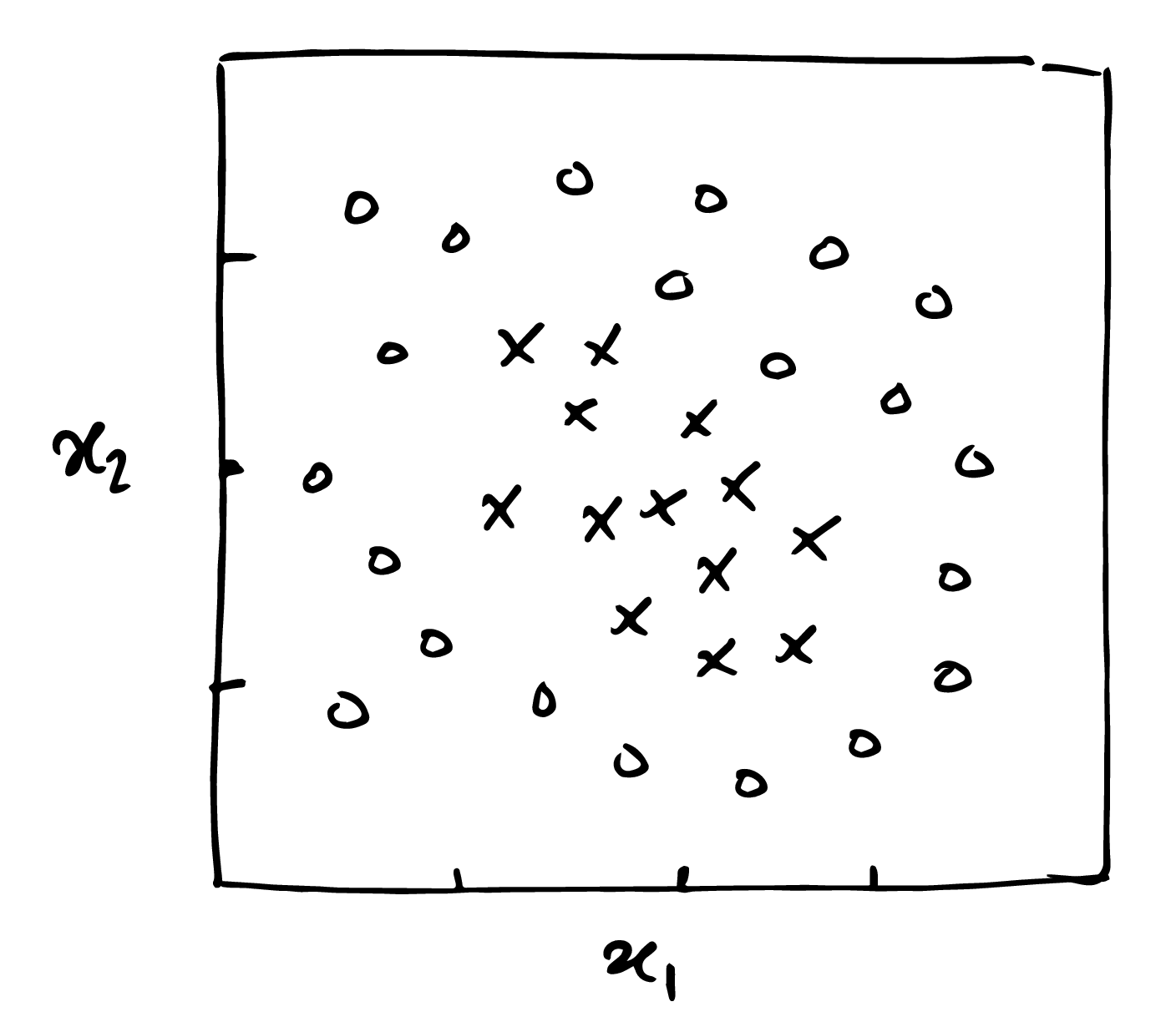
> The kernel function is a similarity function

> The computational complexity of the kernel function is independent of the corresponding mapped feature space

Q5. Which of the following facts about SVM and the kernel trick are true (multiple choices possible):

> The kernel trick can be used for both SVM and Logistic Regression.

Q6. Consider the binary class dataset below (with 2 features (x_1, x_2) and 2 classes (cross and circle). Which of the following classification techniques will be suitable for this dataset? (choose all that can give good results, using the model straight as, without any prior feature recombination).

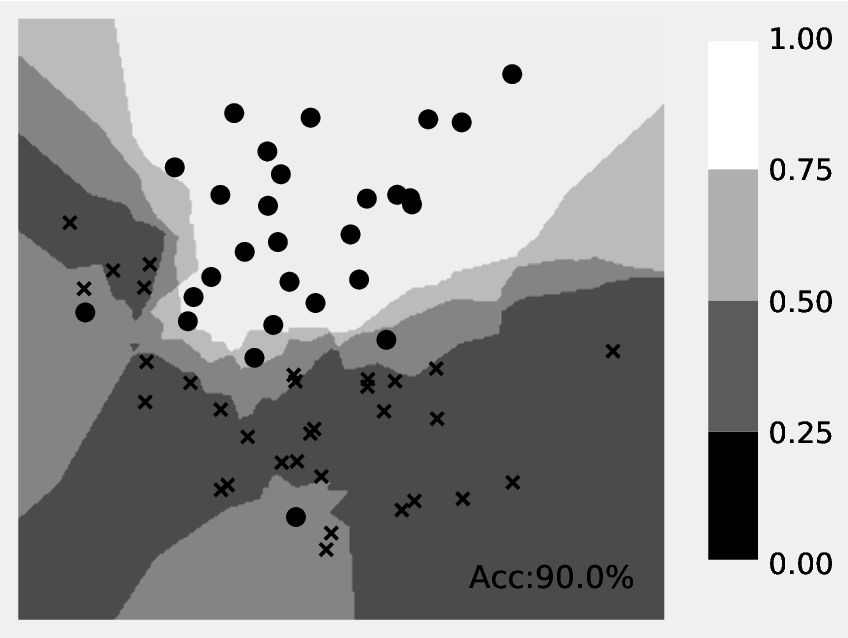


> Decision tree

> Random Forest

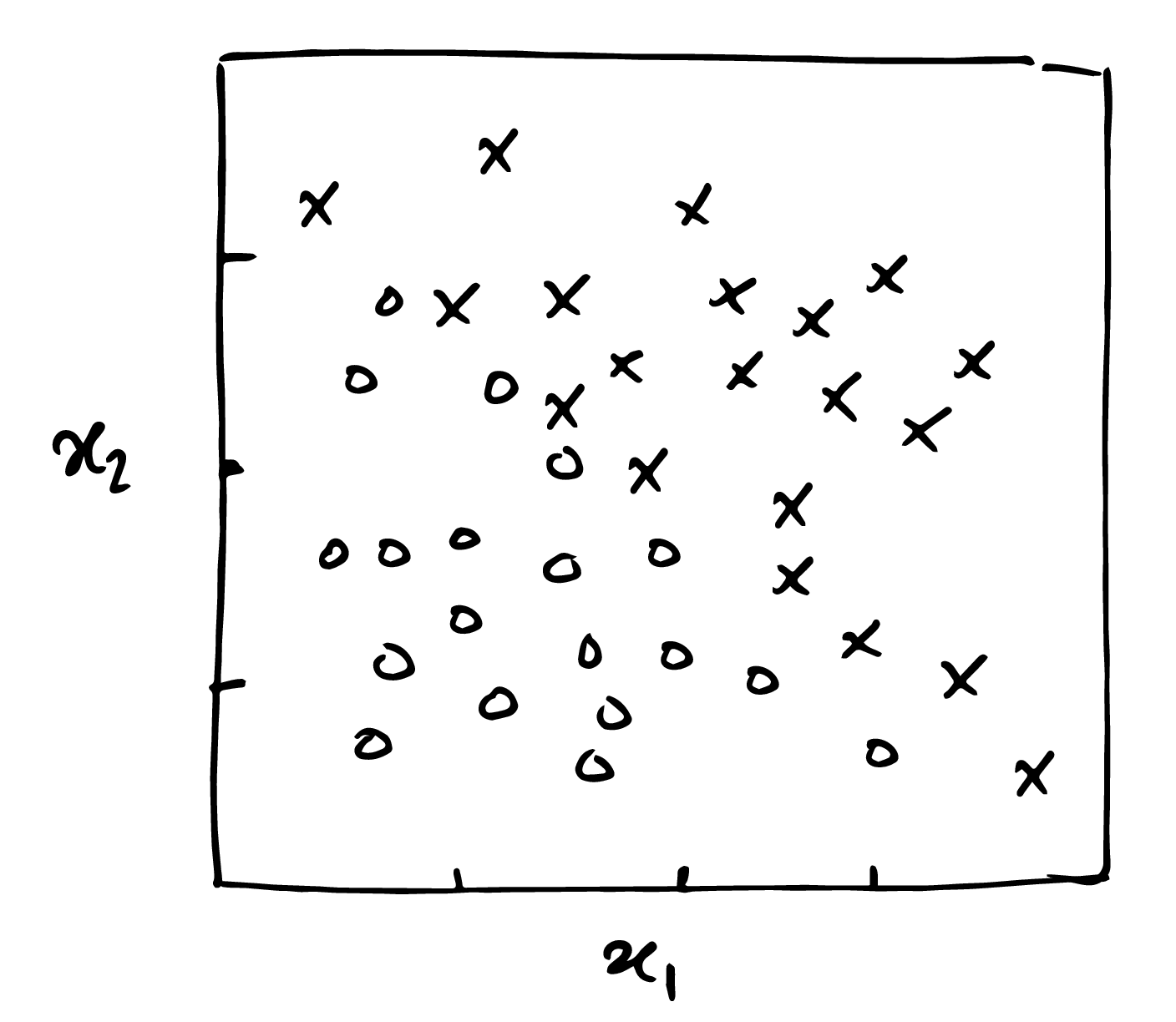
> SVM and RBF kernel

Q7. Which of the following binary classifiers produced the following decision boundary map?



> 3-nearest neighbor

Q8.Consider the binary class dataset below (with 2 features (x_1, x_2) and 2 classes (cross and circle). Which classifier is **least likely to overfit**? (mark one answer)

 > Logistic Regression

Q9. What are the effects of increasing k in k-Nearest Neighbours? (select all answers that are true *in general*)

> increase the area in the boundary decision where the prediction is 0<p<1

> increase the computational time when evaluating the prediction

Q10. Which of these statements about Decision Trees are true? (select all answers that are correct)

> Decisions Trees are computationally more efficient than kernel SVMs.

**Handout 04**

Q1. Consider a binary classifier with the following confusion matrix:

| **Confusion Matrix** | **actual:0** | **actual:1** |
| --- | --- | --- |
| predicted: 0 | TN=52 | FN=30 |
| predicted: 1 | FP=2 | TP=16 |

What is the Accuracy? > 68%

Q2. Consider a binary classifier with the following confusion matrix:

| **Confusion Matrix** | **actual:0** | **actual:1** |
| --- | --- | --- |
| predicted: 0 | TN=34 | FN=9 |
| predicted: 1 | FP=20 | TP=37 |

What is the False Positive Rate? > 37%

Q3. Consider a binary classifier with the following confusion matrix:

| **Confusion Matrix** | **actual:0** | **actual:1** |
| --- | --- | --- |
| predicted: 0 | TN=15 | FN=23 |
| predicted: 1 | FP=6 | TP=56 |

What is the Precision? > 90.32%

Q4. Consider a binary classifier with the following confusion matrix:

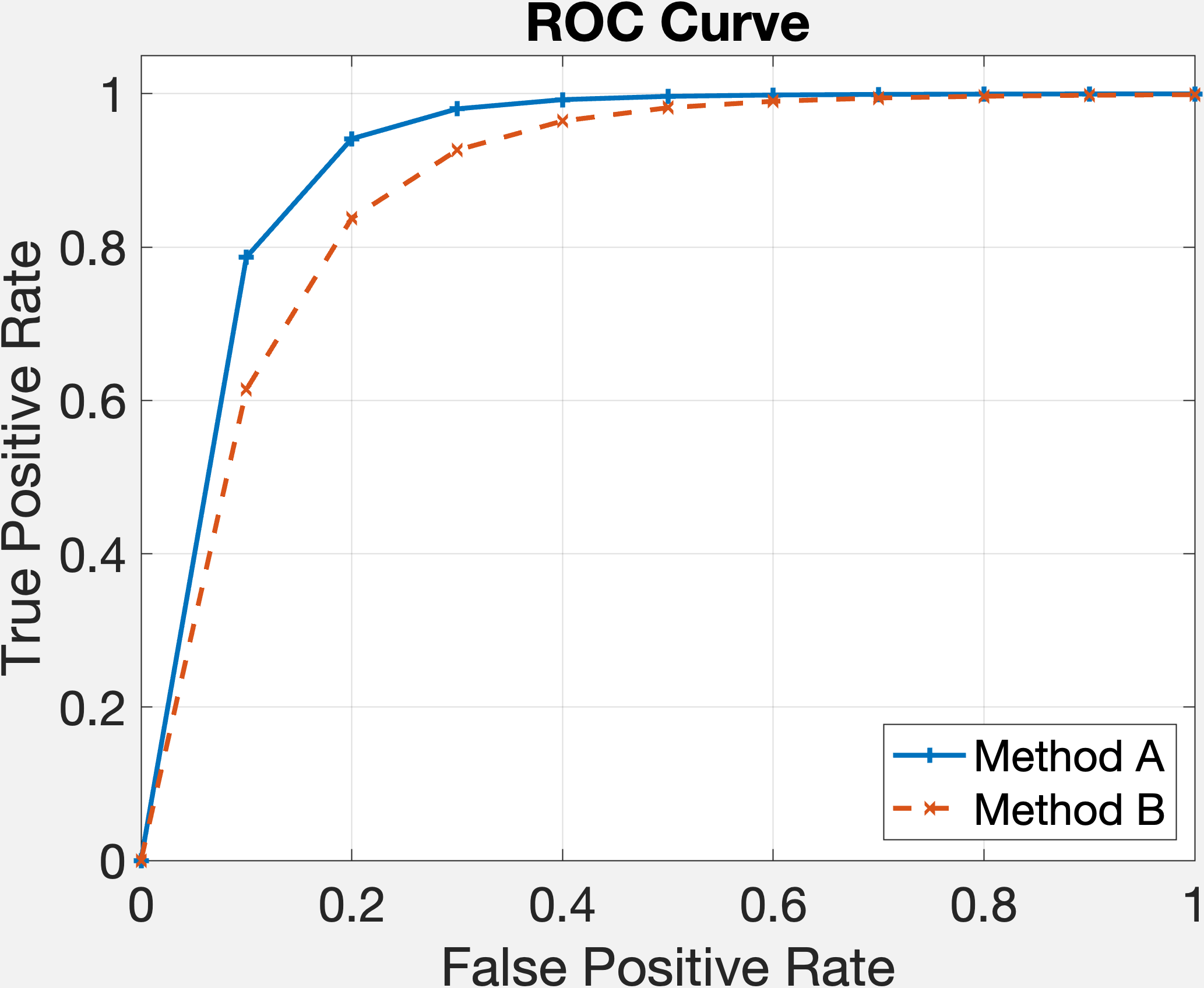
| **Confusion Matrix** | **actual:0** | **actual:1** |
| --- | --- | --- |
| predicted: 0 | TN=37 | FN=12 |
| predicted: 1 | FP=14 | TP=37 |

What is the F1 score? > 74%

Q5. A clinical laboratory reports a 99% accuracy on their latest screening test. What can we say about the the false positive rate?

> not enough information available

Q6. Consider the ROC curve below.



Is Method B more performant than Method A?

> False

Q7. When training and comparing machine learning algorithms, how many types of datasets should you consider?

> 3

Q8. A binary classifier gets 90% accuracy. The test set is well balanced with a 50/50 split between positive and negative samples. What is the worst possible Recall?

> 80%

**Handout 05**

Q1. Consider the network of operations:

<br>e=(a-1)^2+(b-0)^2,\;  b=\cos(d),\; a=\cos(c)<br>

Which of the following statements is true?

> ​​BackPropagation can compute frac{partial e}{partial a} and frac{partial e}{partial b} in any order.

Q2. Consider the network of operations:

<br>a=b\cos(c),\;  b=c+2,\; e=a+b+c<br>

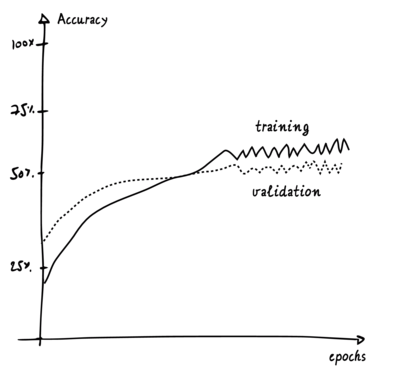
Which of the following statements is true?

> BackPropagation has to compute frac{partial e}{partial a} beforefrac{partial e}{partial b}

Q3. Assuming that the mini-batch fits in the GPU memory, it is faster to process an epoch with a batch size of 1024 than a batch size of 12.

> True

Q4. While training a neural net, we obtain an accuracy graph similar to the one below.



Which of the following statements are reasonable assessment of the situation. (multiple choices possible)

>There some overfitting.

>Dropout was applied.

>The learning rate is too high

Q5.Universal Approximation Theorem states that

> any continuous function can be arbitrarily well approximated by a single hidden layer neural network with a linear neuron output, given enough neurons.

Q6. For a DNN with n units, what is the worst case complexity of BackPropagation?

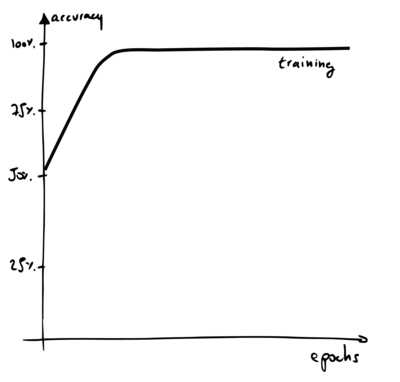
>

Q7. In a Neural Network, what is a unit? (only one answer)

> a differentiable function

Q8. Increasing the number of hidden layers in a Deep Neural Network will always improve the classification accuracy on your test data. > False

Q9.You obtain the following graph for the accuracy on the training set (over the first 50 epochs).



Which of the following statements can be reasonably made from this graph. (multiple choices possible)

> None of the above

Q10. Which of the following methods could be used to simulate more data? (multiple choices possible)

> adding a Dropout layer

> adding a Gaussian Noise layer

**Handout 06**

Q1. A convolutional neural network has 3 consecutive 5 times 5 convolutional layers with stride 1 and no pooling. What is the size of the area of the input image which is used in the computation for a filter in the 3rd convolutional layer.

> 

Q2. A convolutional neural network has 3 consecutive 3 times 3 convolutional layers with stride 1 and no pooling. What is the size of the area of the input image which is used in the computation for a filter in the 3rd convolutional layer.

>

Q3. Is a network made of convolutional neural layers a feed forward neural network?

>True

Q4. Based on the following code, what is the shape of the output?

inputs = keras.layers.Input(shape=(39, 39, 3))

out = Conv2D(2,kernel\_size=(5,5),strides=(2, 2),padding='valid')(inputs)

model = keras.models.Model(inputs, out)

> 18 x 18 x 2

Q5. For which purpose Convolutional Neural Network is used? (select all possible answers)

> Mainly to process and analyse digital images, with some success cases involving processing voice and natural language

Q6. Which answer explains better *flattening*?

> It transforms a tensor into a vector.

Q7. \_\_\_\_ computes the average of the elements in blocks of feature map covered by the filter.

>

Q8. Can softmax be used in Convolutional Neural Network?

> True

Q9. Where does weight sharing occurs?

> in convolution layers but not in fully connected layers

Q10. In what year was LeNet-5, the network that pioneered the use of convolutional layers in neural nets, published? (one answer is correct)

> 1998

**Mid-Term**

Q1.Which of the following models with input x_1,x_2, parameters w_1,w_2 and noise epsilonsim mathcal{N}(0,sigma^2), are linear in the parameters and can be used for Least Squares straight as? (mark all models that are linear)

>y = {w_1}x_1 + {w_2}x_2 + epsilon

>y = left({w_1}x_1 + {w_2}right)x_1^2 + 3{w_1} + epsilon

>y = log(x_1x_2) w_1 + w_2 + epsilon

Q2. To reduce overfitting, which of the following strategies are reasonable?

> use a larger training set.

> increase regularisation.

> increase dropout.

Q3. What does a high value of logit mean in logistic regression (when applied to a binary classification)?

>The observation is far from the decision boundary

>The likelihood for the observation to belong to class 1 is high.

Q4. Which of the following facts are true?

> In *Logistic* Regression, the predicted value h_{bf w}({bf x}) is the likelihood that the data point should be classified as positive given the observed input feature vector.

> In *Probit* Regression, we assume that the error follows a normal distribution

> The *Logistic* Regression loss can be used in Deep Neural Networks.

Q5. Y​​ou have developped machine learning system to spot defects in a manufacturing process. The True Positive Rate of your method is 0.51 and the False Positive Rate 0.48.

Which of the following statements are reasonable assessements of the situation?

> The method is performing badly.

Q6. Which of these is the best measure to decide which threshold to use on a Logistic Regression binary classifier

> ROC Curve

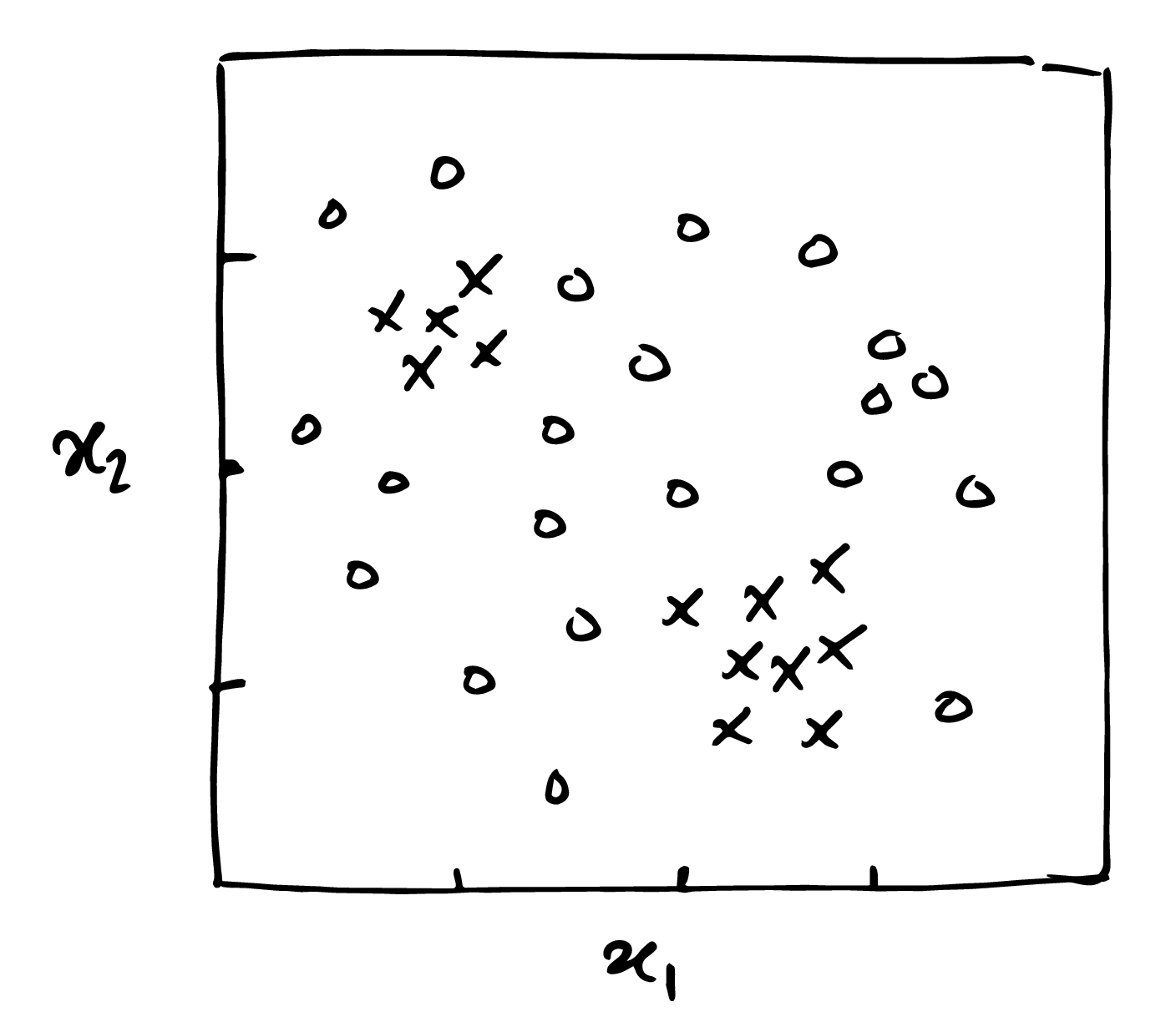
Q7. Consider a binary classifier with the following confusion matrix:

| **Confusion Matrix** | **actual:0** | **actual:1** |
| --- | --- | --- |
| predicted: 0 | TN=28 | FN=9 |
| predicted: 1 | FP=17 | TP=46 |

What is the False Positive Rate?

> 37.78%

Q8. ​​Consider the binary class dataset below, with 2 features (x_1, x_2) and 2 classes (cross and circle). Which classification techniques are suitable, straight as, for this dataset? (choose all that could give good results without further feature re-combination)

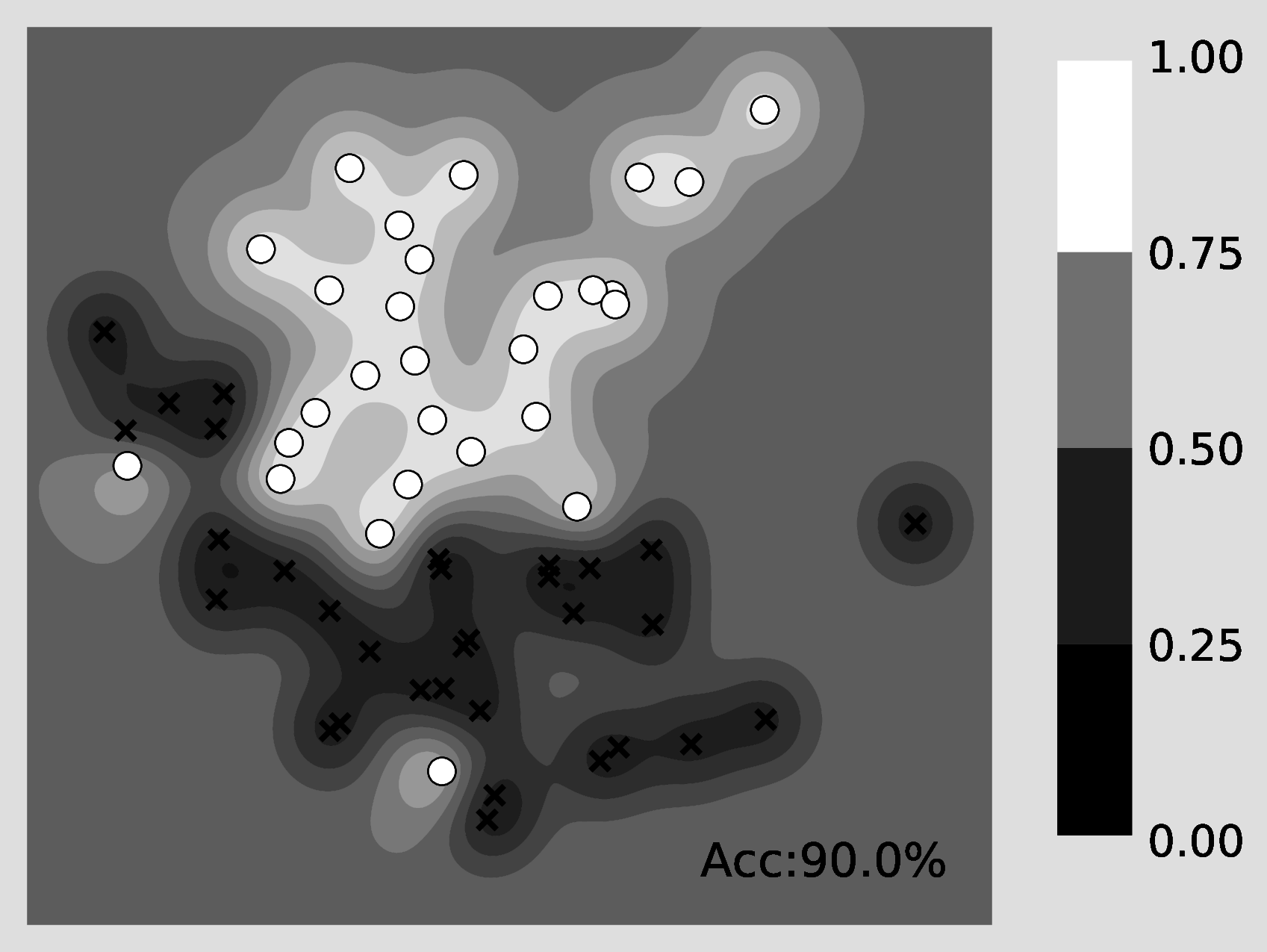


>SVM with RBF kernel

>Random Forest

>K-Nearest Neighbours

Q9. Which of the following binary classifiers produced the following decision boundary map?



> SVM with RBF kernel

Q10. Which of the following statements are correct?

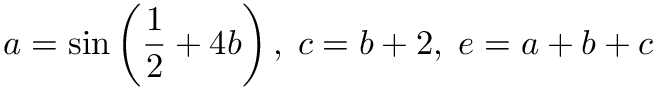
> An artificial neuron is a linear combination of inputs followed by an activation function

> A neural network is called **deep** if it is made of at least 2 layers with non-linear activations.

Q11. What does the backpropagation algorithm compute?

>the gradient of the loss function with respect to its nodes

Q12. Consider the network of operations:



Which of the following statements is true?

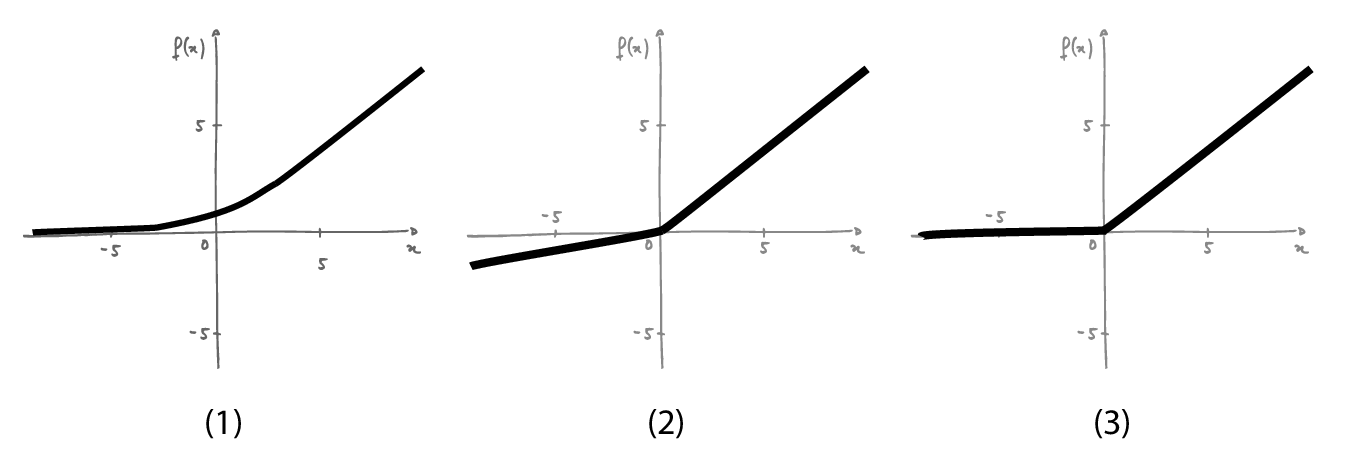
>BackPropagation can compute frac{partial e}{partial a} and frac{partial e}{partial c} in any order.

Q13. Which of the following statements are correct?

> Backpropagation was published in 1986

> AlexNet, by Alex Krizhevsky et al., was published in 2012

Q14.Irrespective of their impact on the network performance, rank the following activation functions by how likely they are to induce vanishing gradient in training (eg. 1 < 2 < 3 means that 1 is the least likely to cause vanishing gradient whilst 3 is the most likely).



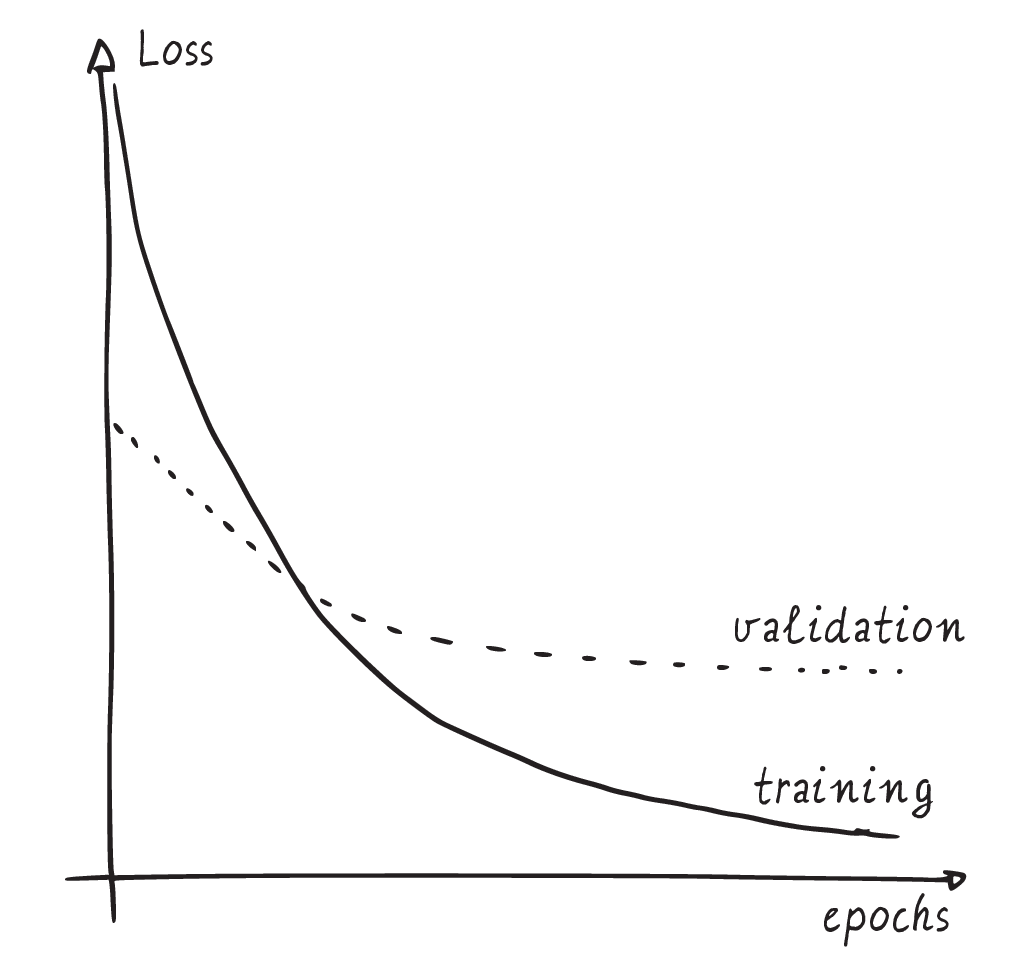
> 2 < 1 < 3

Q15.Which of the following statements are true?

> In a mini-Batch gradient descent, the gradient used for updating the parameters is the average of all the gradients computed in the mini-batch.

>If a dataset contains 1000 observations and the mini-batch size is 10, an epoch will be traversed after 100 iterations

Q16.While training a neural net, we obtain a loss graph similar to the one below.



Which of the following statements are reasonable assessment of the situation.

> There is overfitting

> Dropout applied

Q17. What is the best suited architecture for image recognition

> a convolutional neural network followed by a sequence of fully connected layers

Q18. A convolutional neural network has 4 consecutive 3 times 3 convolutional layers with stride 1 and no pooling. What is the size of the area of the input image which is used in the computation a pixel for a filter of the 4th convolutional layer.

> 9x9

Q19.based on the following code, what is the shape of the output?

inputs = keras.layers.Input(shape=(17, 17, 3))

out = Conv2D(2,kernel\_size=(3,3),

strides=(3, 3), padding='valid')(inputs)

model = keras.models.Model(inputs, out)

> 5x5x2

Q20. What can 1 times 1 convolutions be used for?

>reducing the number of channels

**Handout 7+ 8**

Q1 Given the credit history of a person, a bank tries to predict if the client will fail to repay their mortgage next month. Which of the following neural network architectures would be **the most suitable** to complete this task?

> Recurrent Neural Networks

Q2 Which of the following techniques can help with transfer learning?

> using Batch Normalisation layers.

Q3 We want to classify whether pedestrians are on roller-blades or not. We want to start from a ResNet-50 network pre-trained on ImageNet. We have access to 250 images of pedestrian on roller-blades. Which of the following strategies is the most appropriate? - wrong

> pre-compute the ResNet visual features and use a shallow NN or a SVM with RBF kernel classiffier.

Q4 Taking the conventions from Feed Forward networks, what is the depth of the following network? - wrong

nfeatures = 5

model = keras.Sequential(

[

keras.Input(shape=(12, nfeatures)),

layers.SimpleRNN(128),

layers.Dense(1, activation="sigmoid"),

]

)

>129

Q5 Is the problem of exploding gradients specific to Recurrent Neural Architectures?

> True

Q6 Which answer can describe the following Keras code? (select one answer)

currLayer = Dense(32, (3, 3), trainable=False)(prevLayer)

> Frozen Fully Connected layer.

Q7 In which of the following architectures, does weight sharing occur? (multiple choices possible)

> convolutional neural Network

> recurrent neural network.

> ResNet

Q8 The technical difference between backpropagation and backpropagation through time (BPTT) is that, in BPTT, the recurrent network is first unrolled at runtime into a deep feedforward network.

>True

Q9 Inception Modules are used in Convolutional Neural Networks to allow for more efficient computation and deeper Networks through a dimensionality reduction.

> Ture

Q10 In GANs, the nature of the input noise does matter if we want to ensure that the generated samples follow the target dataset distribution.

>False

Q11 We want to automatically analyse the customersâ product reviews on online shop and predict, based on the text alone, their rating (out of 5 stars). Which of the following statements are true. (multiple possible choices)

>The architecture could use some ResNet modules

>We could use the cross-entropy loss.

>We could use the L1 loss.

>The architecture could be based on a LSTM.