UNIVERSIDADE DE AVEIRO
DEPARTAMENTO DE ELECTRÓNICA TELECOMUNICAÇÕES E INFORMÀTICA

ML - Final exam model

Q1. Which of these is a reasonable definition of machine learning?

- Machine Learning (ML) learns from labeled data.
- ML is the field of study that gives computers the ability to learn without being explicitly programmed.
- ML is the field of allowing robots to act intelligently.
- ML is the science of programming computers.

Q2
The amount of rain that falls in a day is usually measured in either millimeters (mm) or inches. Suppose you use a learning algorithm to predict how much rain will fall tomorrow. Would you treat this as a classification or a regression problem?

- Regression
- Classification

Q3
Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will win a patent infringement lawsuit (by training on data of companies that had to defend against similar lawsuits). Would you treat this as a classification or a regression problem?

- Regression
- Classification

Q4
Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. To which of the following you would apply supervised learning? (Select all that apply.) Determine if this is a classification or regression supervised learning. ~

- Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are.
- Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.
- Examine a web page, and classify whether the content on the web page should be considered "child friendly" (e.g., non-pornographic, etc.) or "adult."
- In farming, given data on crop yields over the last 50 years, learn to predict next year's crop yields.

Q5

Suppose that you have trained a logistic regression classifier, and it outputs on a new example x a prediction $h_\vartheta(x)$ = 0.2. This means (check all that apply):

- ○ Our estimate for P(y=1|x;θ) is 0.2.
- ○ Our estimate for P(y=0|x;θ) is 0.8.
- ○ Our estimate for P(y=1|x;θ) is 0.8.
- ○ Our estimate for P(y=0|x;θ) is 0.2.

Q6

Table 1 describes simple example with two classes. Represent the data set in the space. Is this a linearly or nonlinearly separable problem?

attribute 1	attribute 2	class
−1	1	1
1	−1	1
−1	−1	−1
1	1	−1

Table 1

Q7

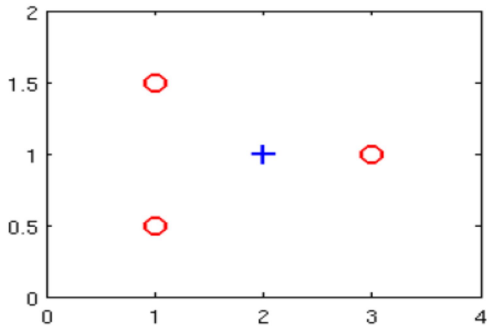
Which of the following statements are true? Check all that apply.

- ☐ In unsupervised learning, the training set is of the form $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$ without labels $y^{(i)}$.
- ☐ Clustering is an example of unsupervised learning.
- ☐ In unsupervised learning, you are given an unlabeled dataset and are asked to find "structure" in the data.
- ☐ Clustering is the only unsupervised learning algorithm.

Q8

Suppose you have the following training set, and fit a logistic regression classifier $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$.

x_1	x_2	y
1	0.5	0
1	1.5	0
2	1	1
3	1	0



Which of the following are true? Check all that apply.

☐ $J(\theta)$ will be a convex function, so gradient descent should converge to the global minimum.

☐ Adding polynomial features (e.g., instead using $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5 x_2^2)$) could increase how well we can fit the training data.

☐ The positive and negative examples cannot be separated using straight line. So, gradient descent will fail to converge.

☐ 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.

Q9

For logistic regression, the gradient is given by $\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$. Which of these is a correct gradient descent update for logistic regression with a learning rate of α? Check all that apply.

- [ ] $\theta := \theta - \alpha \frac{1}{m}\sum_{i=1}^{m} \frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)}\ x^{(i)}$.
- [ ] $\theta := \theta - \alpha \frac{1}{m}\sum_{i=1}^{m} \theta^T x - y^{(i)}\ x^{(i)}$.
- [ ] $\theta := \theta - \alpha \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$.
- [ ] $\theta_j := \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m} \theta^T x - y^{(i)}\ x_j^{(i)}$ (simultaneously update for all j).

Q10

Which of the following statements are true? Check all that apply.

- [ ] For logistic regression, sometimes gradient descent will converge to a local minimum (and fail to find the global minimum). This is the reason we prefer more advanced optimization algorithms such as fminunc (conjugate gradient/BFGS /L-BFGS/etc).
- [ ] The cost function $J(\theta)$ for logistic regression trained with $m \geq 1$ examples is always greater than or equal to zero.
- [ ] Linear regression always works well for classification if you classify by using a threshold on the prediction made by linear regression.
- [ ] The sigmoid function $g(z) = \frac{1}{1+e^{-z}}$ is never greater than one (> 1).

Q11

Suppose you train a logistic classifier $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suppose $\theta_0 = -6, \theta_1 = 1, \theta_2 = 0$. Which of the following figures represents the decision boundary found by your classifier?

Figure:

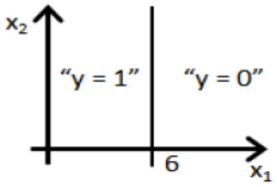


Figure:

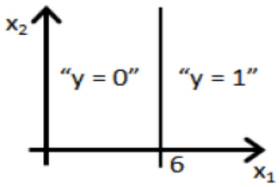
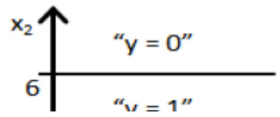


Figure:



Q12

You are training a classification model with logistic

regression. Which of the following statements are true? Check

all that apply.

- [ ] 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.
- [ ] 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.

Q13 (lambda- the regularization parameter)

Suppose you ran logistic regression twice, once with $\lambda = 0$, and once with $\lambda = 1$. One of the times, you got

parameters $\theta = \begin{bmatrix} 74.81 \\ 45.05 \end{bmatrix}$, and the other time you got

$\theta = \begin{bmatrix} 1.37 \\ 0.51 \end{bmatrix}$. However, you forgot which value of

λ corresponds to which value of θ. Which one do you

think corresponds to $\lambda = 1$?

- $\theta = \begin{bmatrix} 1.37 \\ 0.51 \end{bmatrix}$
- $\theta = \begin{bmatrix} 74.81 \\ 45.05 \end{bmatrix}$

Q14

Which of the following statements about regularization are

true? Check all that apply.

- [ ] Using too large a value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ.
- [ ] 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 $\lambda = 0$).
- [ ] Because logistic regression outputs values $0 \leq h_\theta(x) \leq 1$, its range of output values can only be "shrunk" slightly by regularization anyway, so regularization is generally not helpful for it.
- [ ] 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.

Q15

In which one of the following figures do you think the hypothesis has overfit the training set?

Figure:

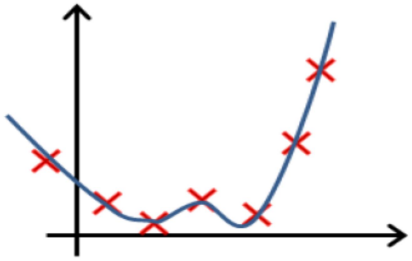
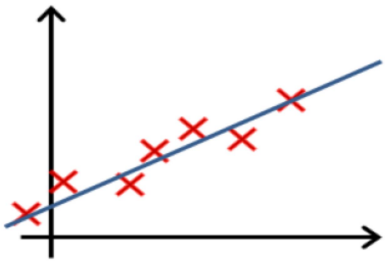


Figure:



Q16

In which one of the following figures do you think the hypothesis has underfit the training set?

Figure:

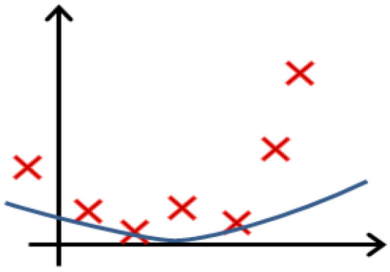
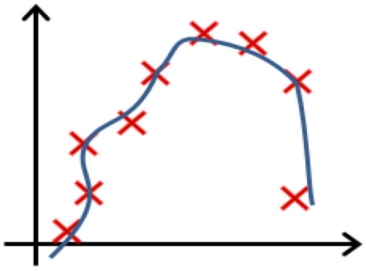


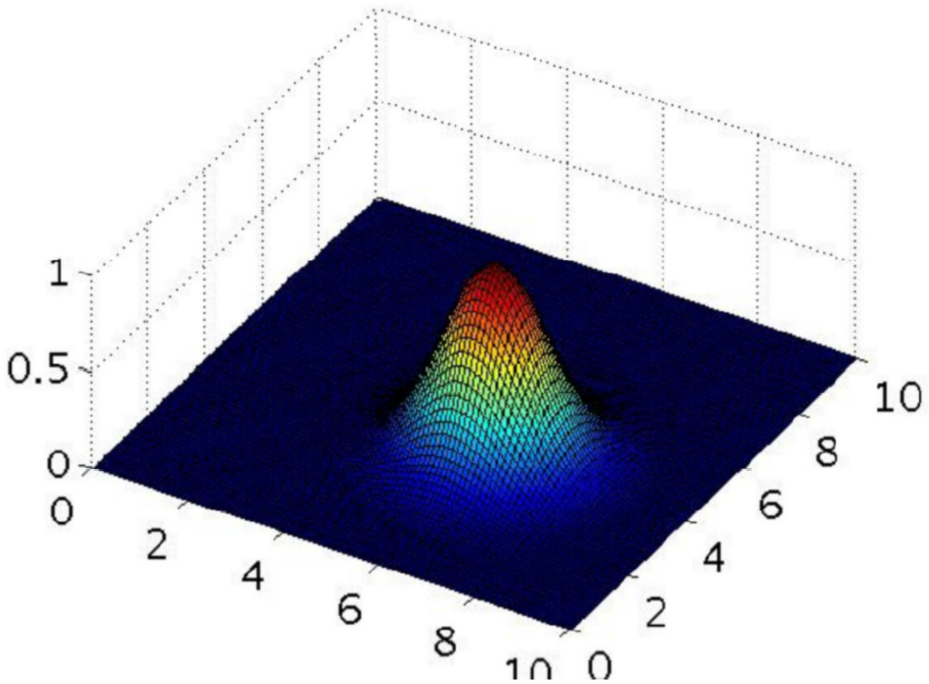
Figure:



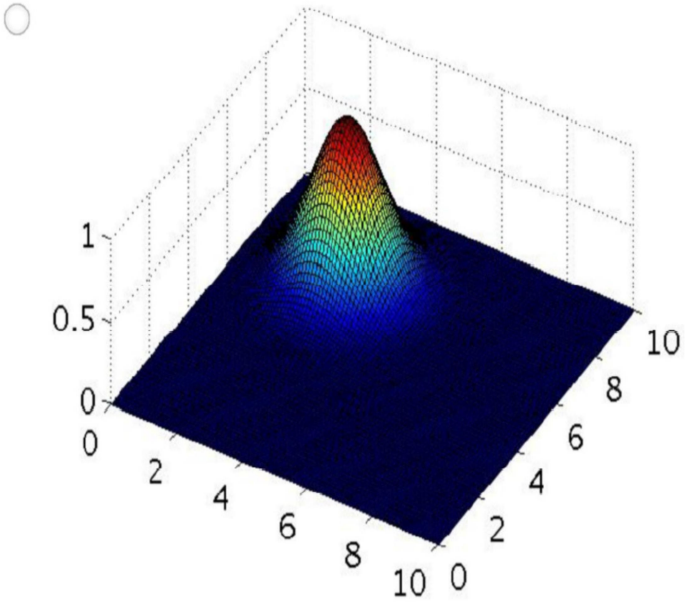
Q17

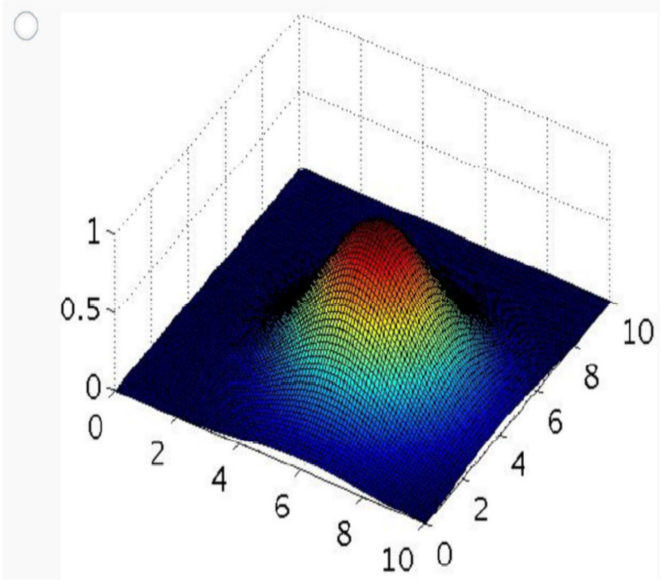
The formula for the Gaussian kernel is given by $\text{similarity}(x, l^{(1)}) = \exp(-\frac{||x - l^{(1)}||^2}{2\sigma^2})$.

The figure below shows a plot of $f_1 = \text{similarity}(x, l^{(1)})$ when $\sigma^2 = 1$.



Which of the following is a plot of f_1 when $\sigma^2 = 0.25$?





Q18

Suppose you are trying to decide among a few different choices of kernel and are also choosing parameters such as C, σ^2, etc. How should you make the choice?

- ○ Choose whatever performs best on the training data.
- ○ Choose whatever performs best on the cross-validation data.
- ○ Choose whatever performs best on the test data.
- ○ Choose whatever gives the largest SVM margin.

Q19

Suppose you run k-means using k = 3 and k = 5. You find that the cost function J is much higher for k = 5 than for k = 3. What can you conclude?

- ☐ This is mathematically impossible. There must be a bug in the code.
- ○ The correct number of clusters is k = 3.
- ○ In the run with k = 5, k-means got stuck in a bad local minimum. You should try re-running k-means with multiple random initializations.
- ○ In the run with k = 3, k-means got lucky. You should try re-running k-means with k = 3 and different random initializations until it performs no better than with k = 5.

Q20

Which of the following is the recommended way to initialize k-means?

- ☐ Pick a random integer i from $\{1, ..., k\}$. Set $\mu_1 = \mu_2 = \cdots = \mu_k = x^{(i)}$.
- ○ Pick k distinct random integers $i_1, ..., i_k$ from $\{1, ..., k\}$.

  Set $\mu_1 = x^{(i_1)}, \mu_2 = x^{(i_2)}, ..., \mu_k = x^{(i_k)}$.
- ○ Pick k distinct random integers $i_1, ..., i_k$ from $\{1, ..., m\}$.

  Set $\mu_1 = x^{(i_1)}, \mu_2 = x^{(i_2)}, ..., \mu_k = x^{(i_k)}$.
- ○ Set every element of $\mu_i \in \mathbb{R}^n$ to a random value between $-\epsilon$ and ϵ, for some small ϵ.

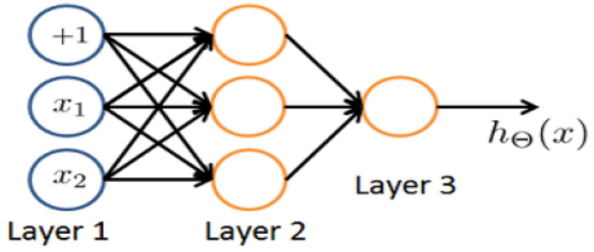
Q21

Which of the following are good / recommended applications of PCA? Select all that apply.

- ☐ To compress the data so it takes up less computer memory / disk space
- ☐ To reduce the dimension of the input data so as to speed up a learning algorithm
- ☐ Instead of using regularization, use PCA to reduce the number of features to reduce overfitting
- ☐ To visualize high-dimensional data (by choosing k = 2 or k = 3)

Q22

You have the following neural network:



You'd like to compute the activations of the hidden layer $a^{(2)} \in \mathbb{R}^3$. One way to do so is the following Octave code:

```
% Theta1 is Theta with superscript "(1)" from lecture
% ie, the matrix of parameters for the mapping from layer 1 (input) to layer 2
% Theta1 has size 3x3
% Assume 'sigmoid' is a built-in function to compute 1 / (1 + exp(-z))

a2 = zeros (3, 1);
for i = 1:3
  for j = 1:3
    a2(i) = a2(i) + x(j) * Theta1(i, j);
  end
  a2(i) = sigmoid (a2(i));
end
```

You want to have a vectorized implementation of this (i.e., one that does not use for loops). Which of the following implementations correctly compute $a^{(2)}$? Check all that apply.

- [ ] a2 = sigmoid (Theta1 * x);
- [ ] a2 = sigmoid (x * Theta1);
- [ ] a2 = sigmoid (Theta2 * x);
- [ ] z = sigmoid(x); a2 = Theta1 * z;

Q23

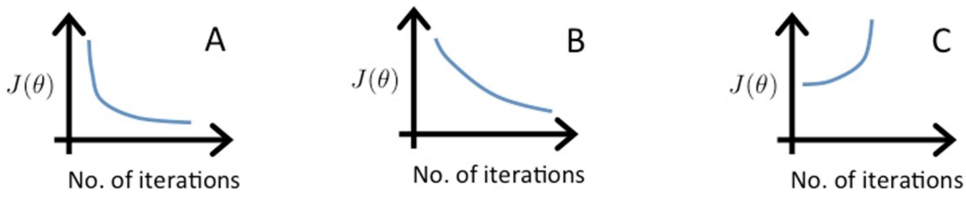
Let $J(\theta) = 2\theta^3 + 2$. Let $\theta = 1$, and $\epsilon = 0.01$. Use the formula $\frac{J(\theta+\epsilon)-J(\theta-\epsilon)}{2\epsilon}$ to numerically compute an approximation to the derivative at $\theta = 1$. What value do you get? (When $\theta = 1$, the true/exact derivati ve is $\frac{dJ(\theta)}{d\theta} = 6$.)

- ( ) 5.9998
- ( ) 6
- ( ) 6.0002
- ( ) 8

Q24 Is it true that the KNN classifier needs the training set during the test phase? Justify your answer.

Q25

Suppose you ran gradient descent three times, with different values for the parameter learning rate α=0.01, α=0.1, and α=1, and got the following three plots (labelled A, B, and C):



Which plots corresponds to which values of α?

- A is with α=0.01, B is with α=0.1, C is with α=1.
- A is with α=0.1, B is with α=0.01, C is with α=1.
- A is with α=1, B is with α=0.01, C is with α=0.1.
- A is with α=1, B is with α=0.1, C is with α=0.01.

Q26: Check all that apply regarding the typical characteristics of the back-propagation algorithm.

- It can be stacked into poor local minima.
- It does not require labeled data.
- It is very slow in networks with multiple hidden layers.
- It cannot be applied to multilayer perceptron.

Q27: Which of the following ML architectures are related with deep learning? Check all that apply.

- Sparse Stacked Autoencoder
- Multilayer perceptron.
- Convolution Neural Network (CNN)
- Reinforcement Learning

Q28: Which of the following statements regarding Softmax Regression (SR) are true? Check all that apply.

- o Softmax Regression is a supervised learning algorithm
- o Softmax Regression is a binary classifier.
- o Softmax Regression is more suitable than Logistic Regression for mutually exclusive classes.
- o The gradient descent cannot be applied for Softmax Regression.

Q29. Suppose you have *m*=14 examples with *n*=3 features. What are the dimensions of the data matrix X, the output y and the vector of parameters θ when you implement it .

- **A.** X is 14x4, y is 14x1, θ is 4x1
- **B.** X is 14x3, y is 14x1, θ is 3x3
- **C.** is 14x3, y is 14x1, θ is 3x1
- **D.** X is 14x4, y is 14x4, θ is 4x4

Q30. Propose a solution to treat the problem of class unbalanced data.