Q	Grade
21	0.6
22	0.6
23	0.6
24	0.6
25	0.6
26	0.6
27	0.6
28	0.6
29	0.6
30	1.3

UNIVERSIDADE DE AVEIRO
DEPARTAMENTO DE ELECTRÓNICA TELECOMUNICAÇÕES E INFORMÀTICA

Machine Learning final exam - 9/June 2020 PART 3 (40 min)

Nº: 85129 Name: Gabriel Augusto Santos Silva

Instructions: You have 40 min. to write down your answers of the questions below. During this time, please, **keep switched on the camera of your PC**. Save and name the file with your answers as
"ML_P3_XXXXX" and substitute XXXXX with your academic (mechanographic) number.
Send a **PDF** version of the file with your answers and a PDF file of the digitalized pages, you may have produced while solving the problems, to petia@ua.pt with **Subject: ML_P3 + your academic number**

Q21. Suppose we have three cluster centroids $\mu_1 = [1, 2]$, $\mu_2=[-3, 0]$, $\mu_3= [4, 2]$ and a training example $x^{(i)} = [3, 1]$. After a cluster assignment step, what will $c^{(i)}$ be (i.e. to which cluster will be assigned $x^{(i)}$) ? **Justify why ?**

- **A.** $c^{(i)}$ is not assigned
- **B.** $c^{(i)} =3$
- **C.** $c^{(i)} = 2$
- **D.** $c^{(i)} = 1$

Answer: B

Q22. Suppose you have an unlabeled dataset $\{x^{(1)},\ldots,x^{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clustering of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

A. Use the elbow method.

B. Compute the distortion function $J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_k)$, and pick the one that minimizes this.

C. Compute the mean of the cluster centroids obtained after each clustering and use them as the final clustering.

D. Visually examine the clusterings, and pick the best one.

Answer: B

Q23. Which of the following statements are true? Check all that apply.

A. Principal Component Analysis (PCA) is a recommended approach to deal with over fitting.

B. PCA is only used to reduce data dimensionality by 1 (from 3D to 2D, from 2D to 1D).

C. Given an input $x \in R^n$, PCA compresses it to a lower-dimensional vector $z \in R^k$

D. If the input features are on very different scales, it is a good idea to perform feature scaling before applying PCA.

Answer: C, D

Q24. Which of the following are recommended applications of PCA? Select all that apply.

A. To get more features to feed into a learning algorithm.

B. Data visualization: Reduce data to 2D (or 3D) so that it can be plotted.

C. Data compression: Reduce the data dimension, so that it takes less memory / disk space.

D. Preventing overfitting: Reduce the number of features, so that there are fewer parameters to learn.

Answer: B,C

Q25. Recall that "np.dot(a,b)" performs a matrix multiplication on a and b, whereas "a*b" performs an element-wise multiplication. Consider the two following random arrays "a" and "b":

a = np.random.randn(4, 3) # a.shape = (4, 3)
b = np.random.randn(3, 2) # b.shape = (3, 2)
*c = a*b*
d = np.dot(a,b)

What will be the shape of "c"?

Answer: c will fail to be computed. For element wise multiplication both matrices need to have the same shape, or one of them has to be a vector with the same amount of rows or columns as the other matrix.

What will be the shape of "d"?
Answer: d will be 4x2

Q26. In Decision Tree algorithm, the Entropy, the Information gain and the Gain ratio are measures to choose the best feature for splitting at that node. Which of the following statements are true? Check all that apply.

A. Choose the feature that maximizes the entropy.
B. Choose the feature that maximizes the information gain.
C. Choose the feature that minimizes the node impurity.
D. Choose the feature that minimizes the gain ratio.

Answer: B,C

Q27. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

A. Given a set of news articles from many different news websites, find out what are the main topics covered.

B. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

C. From the user profiles on a website, figure out what different groups of users exist.

D. Given many emails, you want to determine if they are Spam or Non-Spam emails.

Answer: A,C

Q28. You have two coins: one fair with probability P1 (heads)=0.5 and one loaded with probability P2 (heads)=1.
You now pick a coin at random with 0.5 chance. You flip this coin, and see “heads”.
What is the probability this is the loaded coin? What is the probability this is the fair coin ?

Answer:

First define our events: L -> choosing the loaded coin
n_L -> choosing the normal coin
H -> seeing heads

From the problem we get that : P(H | L) = 1

P(H | N_L) = 0.5

P(L) = P(n_L) = 0.5

The probability of it being the loaded coin is P(L|H). P(L|H) = (P(H|L)*P(L) / P(H))
P(H|L) = 1.; P(L) = 0.5; The P(H) = (P(H|L)*P(L) + P(H|n_L)*P(n_L)) = 1*0.5 + 0.5*0.5 = 0.75;
So The probability of it being a loaded coin is P(L|H) = 1*0.5/(0.75) = 66.(6) %
The probability of being the fair coin is P(n_L|H) = (P(H|n_L)*P(n_L)/P(H)).
P(H|n_L) = 0.5; P(n_L)=0.5; P(H) = 0.75 (from before)
So the probability of it being the fair coin is P(n_L|H) = (0.5*0.5)/0.75 = 33.(3)%

Q29. Which of the following are true? Check all that apply.

A. If you develop an anomaly detection system, you do not use labelled data to improve the system.

B. In a typical anomaly detection setting, we have a large number of anomalous examples, and a relatively small number of normal/non-anomalous examples.

C. When developing an anomaly detection system, it is often useful to select an appropriate numerical performance metric (threshold *epsilon*) to evaluate the effectiveness of the learning algorithm.

D. In anomaly detection, we fit a model p(x) to a set of negative (y=0) examples, without using any positive examples we may have collected of previously observed anomalies.

Answer: C,D

Q30. The Health System collected data (weight and height, shown in Fig.Q30) from 20 teenagers (10 girls and 10 boys) and concluded that both features (weight and height) follow different Gaussian distributions.

The mean (μ) and the standard deviation (σ) of each feature and class are given in Table Q30.

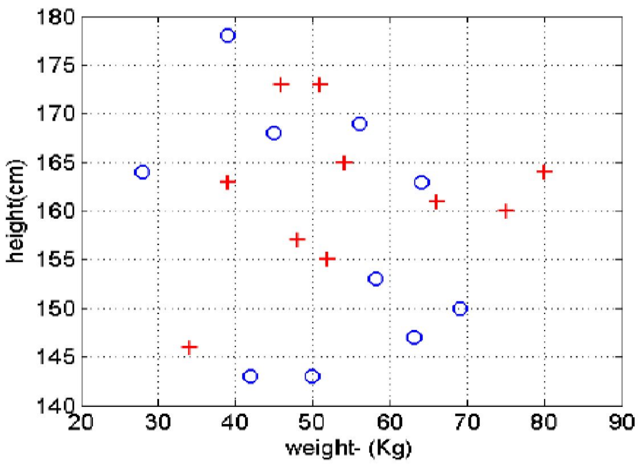

Fig. Q30 The teenager dataset : boy(+); girl(o)

Table Q30 Mean and standard deviation

class	Weight ($\mu;\sigma$)	Height ($\mu;\sigma$)
boy	54; 15	162; 8
girl	52; 13	158; 12

Apply Naive Bayes Classifier to decide if a new example *x* =[50 kg 150 cm] is a boy or a girl.

Answer: The similarity is measured by exp(-|x - mu|/2sigma^2);

For the weight we have a 50 kg example. 50 kg is closer to 52 than 54, meaning the numerator in that expression is smaller when using 52, making the similarity higher (because of the minus sign, a smaller distance means we're closer to max similarity). So for the weight feature we have a girl, most likely. For the height, 150 cm is closer to 158 than 162, so, again, we're more likely to have a girl. So the classifier will classify this new example as a girl. Probability of being boy is the similarity of 50 to 54 + similarity of 150 to 162. The probability of being girl is similarity of 50 to 52 + similarity of 150 to 168. The girl probability will be higher.