Questions

1. General Questions

(a) (1 point) Which of the following problems are more suitable for classification? There might be more than one answer, and you must indicate all correct answers to obtain 1 point.

a. Predicting if a job advert is an advert for a real job or a scam.

b. Predicting the blood alcohol content of a person based on data including the persons BMI, how many and what types of drink they have had, and how long they have been drinking.

c. Predicting which film a user will watch next based on their preferences and their past history.

(b) (1 point) Consider the following data-sets and problems. Which data sets would require un-supervised learning. There might be more than one answer, and you must indicate all correct answers to obtain 1 point.

a. An un-labelled set of pictures of animals. Your job is to classify the pictures into different classes of animals.

b. A labelled set of pictures of cats, dogs and automobiles.

c. A data set containing the closing price of every house sold in Uppsala since 2000. You are to predict the final house price of a house based on its location and other data.

(c) (1 point) Which of the following features are categorical. There might be more than one answer, and you must indicate all correct answers to obtain 1 point.

a. The gender of a person.

b. The age of a person.

c. Which country the person lives in.

d. The weight of a person.

(d) (1 point) Consider using gradient descent to learn a hypothesis $h_\theta = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$ for regression. During gradient descent an error/loss function J is minimised. Does the gradient descent for linear regression always converge to a global minimum for any training set?

a. True

b. False

(e) (1 point) Again, consider using gradient descent to learn a hypothesis $h_\theta = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$ for regression. During gradient descent an error/loss function J is minimised. At the end of gradient descent for linear regression does the function J always equal 0 for *any* training set?

a. True

b. False

(f) (1 point) Why should the training set always be a different set from the validation set. Please indicate the correct answer.

a. To avoid over-fitting.

b. Learning algorithms become inefficient if the training set is too large.

(g) (1 point) You are developing a classifier to detect cancer. The algorithm should report true if the patient has cancer. You want true cancer patients not to be diagnosed as non-cancer patients. Given confusion matrix which of the following situations is better.

a. A high false positive rate.

b. A high true positive rate.

(h) (2 points) You have the following data concerning the occurrence of words in spam email.

Spam (Y/N)	'Home' Occurs	'BitCoin' Occurs
Y	Y	Y
N	Y	N
N	N	N
Y	Y	N
Y	N	N

You are given an email that contains the word "BitCoin" which of the following is the correct value of $P(\text{Spam}|\text{BitCon})$

a. $\frac{2}{3} \times \frac{3}{5}$

b. $\frac{2}{3} \times \frac{3}{5}$

c. $\frac{2}{3} \times \frac{2}{5}$

d. $\frac{1}{3} \times \frac{3}{5}$

(i) (1 point) Consider using logistic regression with a linear hypothesis to classify 2-dimensional data points separated into two classes. Which of the following statements is true, note that there might be more than one correct answer and you must indicate all correct answers to get full marks:

a. Logistic regression is able to classify any data set.

b. Logistic regression can only classify a data set if and only if there is no overlap in the classes.

a. Logistic regression requires the two classes to be linearly separable.

b. Logistic regression can only classify the two classes if and only if it is possible to use linear regression to separate two classes.

(j) (1 point) Consider using logistic regression with a linear hypothesis to classify 2-dimensional data points separated into two classes. Which of the following statements is true, note that there might be more than one correct answer and you must indicate all correct answers to get full marks:

(a) Logistic regression is able to classify any data set.

(b) Logistic regression can only classify a data set if and only if there is no overlap in the classes.

(C) Logistic regression requires the two classes to be linearly separable.

(D) Logistic regression can only classify the two classes if and only if it is possible to use linear regression to separate two classes.

(k) (1 point) Given two probabilities p_1 and p_2. If you are told that the entropy $H(p_1, p_2) = -(p_1 \log_2 p_1 + p_2 \log_2 p_2)$ is 0, then what are the possible values for p_1 and p_2. There may be more than one answer. You must indicate all correct answers to get full marks.

(a) $p_1 = (1 - p_2)$

(b) $p_1 = \frac{1}{2} = p_2$

(c) $p_1 = 0, p_2 = 1$

(d) $p_1 = 1, p_2 = 0$

(l) (1 point) The ID3 algorithm for constructing decision trees always constructs the smallest decision tree possible for *any* training set.

(a) True

(b) False

(m) (2 points) You are using the ID3 algorithm to construct a classifier to work out if somebody can play golf or not. You are given the following data:

Humidity	Sunny	Windy	Play
L	N	Y	True
L	N	Y	True
H	Y	N	True
L	Y	Y	True
H	Y	Y	False
L	Y	N	False
H	N	N	False

Consider the probability of playing golf and the probability of not playing golf. Which of the correct value of entropy of our dataset

(a) $-\left(\frac{4}{7} \log_2 \frac{4}{7} + \frac{3}{7} \log_2 \frac{3}{7}\right)$

(b) $-\left(\frac{3}{7}\log_2\frac{4}{7}+\frac{3}{7}\log_2\frac{3}{7}\right)$

(c) $-\left(\frac{1}{4}\log_2\frac{4}{7}+\frac{1}{3}\log_2\frac{3}{7}\right)$

(n) (2 points) Again using the ID3 algorithm and the same data-set as in the previous question part. Which of the following is the correct expression for the information gain for splitting on the attribute "Windy"? Note that $H(P)$ is the entropy of the whole data set calculated in the previous question part. The exam had a misprint:

(a) $H(P) - (H(P|\text{Windy} = \text{Yes}) + H(P|\text{Windy} = \text{No}))$

(b) $H(P) - \left(\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{4}{7}H(P|\text{Windy} = \text{No})\right)$

(c) $H(P) + \left(\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{1}{4}H(P|\text{Windy} = \text{No})\right)$

(d) $H(P) + \left(\frac{3}{7}H(P|\text{Windy} = \text{Yes}) + \frac{4}{7}H(P|\text{Windy} = \text{No})\right)$

(o) (1 point) As part of the principle component algorithm the eigen-vectors and eigen-values are calculated of the co-variance matrix of the training data. What does the largest eigen-value tell you? There is only one correct answer.

(a) The number of dimensions your reduced data set will have.

(b) The direction to project in order to maximise the variance in that dimension.

(c) A normalising coefficient for the first feature that you need to avoid over-fitting.

(p) (1 point) For logistic regression it is possible to use a regularisation parameter λ as follows:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\Big[-y^{(i)}\log(\sigma(h_\theta(x^{(i)}))) - (1-y^{(i)})\log(1-\sigma(h_\theta(x^{(i)})))\Big] + \lambda\sum_{i=1}^{n}\theta_i^2$$

Why is regularisation used?

(a) It normalises the input data so that each feature has mean zero.

(b) Increasing λ reduces the considered dimension of the training data.

(c) It avoids over-fitting on the training data by forcing gradient descent to learn small weights θ_i.