Questions

General Questions

(a) (1 point) On these algorithms is an unsupervised learning algorithm. Which one is is?

A. Linear Regression

B. K-means clustering

C. Support vector machines

D. Naive Bayes Classification

(b) (1 point) You have a data-set labelled into two classes “True” and “False”. Which of the following machine learning algorithms could you try without any modification. There is more than one answer, please circle or indicate all possible correct answers.

A. Linear Regression

B. Logistic Regression

C. Naive Bayes Classification

D. K-means clustering

(c) (1 point) Which of the following data types are categorical variables. There is more than one answer. You must circle or indicate all possible correct answers.

A. Gender (Male or Female)

B. Weight (in Kg)

C. Age

D. Member of Gotland Nation Yes or No.

(d) (1 point) Which of these classification problem is a regression problem. There is more than one answer, please circle or indicate all possible correct answers.

A. Predicting the probability that a message is spam or not.

B. Labelling a message as spam or not.

C. Predicting somebodies exam grade based on the number of lectures they attend and their score on assignments given during the course.

D. Deciding if a patient has cancer or not.

(e) (1 point) You are building a self driving car, and you are building a system to recognise pedestrians (fotgängare). If the car kills pedestrians then you will go to prison. The algorithm output positive if there is a pedestrian in the path of the car. Which of the following are you trying to minimise?

A. False Negative

B. True Positive

C. False Positive

(f) (1 point) Assume that you are using gradient descent to train your learning algorithm, and you are given an error or loss function J used during training of a machine learning algorithm. If $\theta_0, \dots, \theta_n$ are the weights or parameters of the algorithm and

$$\frac{\partial J(X, \theta)}{\partial \theta_i} = 0$$

for all i then which of the following statements are true (there is only one correct answer):

A. The values of $\theta_0, \theta_1, \dots, \theta_n$ are the values of the global minimum of the function J.

B. The values of $\theta_0, \theta_1, \dots, \theta_n$ are the values of a local minimum of the function J.

C. There is not enough training data and you must collect more before you can train the algorithm further.

(g) (2 points) Given the following data set:

x	y
1	2
2	3
3	4

You are using gradient descent to fit a linear regression model. $h_{\theta_0,\theta_1}(x) = \theta_0 + \theta_1 x$. Which of the following expressions is the correct value of the loss (or error) function J?

A. $J(\theta_0, \theta_1) = 2 - (\theta_0 + \theta_1) + 42 - (\theta_0 + 2\theta_1)$

B. $J(\theta_0, \theta_1) = \frac{1}{2}((2-(\theta_0+\theta_1))^2 + (42-(\theta_0+2\theta_1))^2 + (43-(\theta_0+3\theta_1))^2)$

C. $J(\theta_0, \theta_1) = \frac{1}{2x}((1-(\theta_0+2\theta_1))^2 + (2-(\theta_0+3\theta_1))^2 + (3-(\theta_0+4\theta_1))^2)$

D. $J(\theta_0, \theta_1) = \frac{1}{2x}((2 - (\theta_0 + \theta_1)) + (2 - (\theta_0 + 2\theta_1)) + (4 - (\theta_0 + 3\theta_1)))$

(h) (1 point) Logistic regression is a regression algorithm:

A. True

B. False

(i) (1 point) Logistic regression requires all variables to be categorical

A. True

B. False

(j) (1 point) Which of these statements best describe what is learnt when doing logistic regression:

A. A linear hypothesis $h_\theta = \theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i$, such that $h(x) = 1$ if the data point x belongs to the class and $h(x) = 0$ if x does not belong to the class.

B. A hyperplane defined by the set of values for which $\theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i = 0$ that separates the data into two regions with those points belonging to the class and those points not belonging the class.

C. The value $\theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i$ is the probability that the point x_i belongs to the class or not.

(k) (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	Y
N	N	Y
Y	Y	N
Y	N	N

(l) (1 point) 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. $\frac{2}{3} \times \frac{4}{5}$
C. $\frac{2}{3} \times \frac{2}{5}$
D. $\frac{1}{3} \times \frac{3}{5}$

(m) (1 point) Which of the following options can be used to get global minima in K-Means Algorithm?

(a) Try to run algorithm for different centroid initialisation
(b) Adjust number of iterations
(c) Find out the optimal number of clusters

Answers:

A. 1 and 3
B. 3
C. 2 and 1
D. 1, 2 and 3

(n) (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.

(o) (2 points) Given a d-dimensional data set with n points, $x_1, \dots, x_n$, where each x_i belongs to $\mathbb{R}^d$ after running principle component analysis (PCA) you pick the first P principle component directions. Your dimension reduced dataset is now n points $y_1, \dots, y_n$, where each y_i belonging to $\mathbb{R}^P$, is a P-dimensional point. Can you always reconstruct any data point x_i from y_i?

A. Yes, if $P < d$
B. Yes, if $P < n$,
C. Yes, if $P = d$,
D. It is never possible.

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

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

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.

D. Increasing λ will make the algorithm converge to a solution much more quickly.

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

A. True

B. False