Q1:

a) Define the joint probability P(A,B) and the conditional probability P (A given B). You may want to use a diagram/sketch. Give the formula that relates them. [3 marks]

b) Give the law of total probabilities, that is express p(A) using a set of events B1, B2, …, BN and the corresponding conditional (or joint) probabilities. What are the conditions that need to hold? [4 Marks]

c) Some emails received by users are spams, attempting to gain illicit access to employee’s accounts. You are building a classifier to detect such illicit emails. You start by using the feature of whether or not an email asks for the recipients password, as this an important datum indicating if the email is in fact a hacking attempt. In particular 95% of hacking emails contain requests for a user’s password, and 95% of legitimate emails do not contain requests for a user’s password. 1% of emails overall are hacking attempts.

i. If your classifier scans an email with a password request, what is the probability that the email is in fact a hacking attempt?

ii. Comment on this value that you calculate.

iii. What is the probability that your classifier makes an error? [12 marks]

d) Explain the difference between Maximum Likelihood (ML) and Maximum a Posterior (MAP) methods of parameter estimation. [6 marks]

Q2:

a) Compare and contrast the goals in Linear Regression and Logistic Regression. [4 marks]

b) The form of a linear regression model is y=w^T\*x. Assuming the mean squared error cost function, derive gradient descent updates for the weights w. [9 marks]

c) What is the limitation of the networks without hidden layers, that was overcome by Multilayer Networks? Is it is essential that the activation function is non-linear? [6 marks]

d) Practical pitfalls with training neural networks include:

(i) getting stuck in local optima,

(ii) underfitting or overfitting,

(iii) bad learning rate. Explain what each of these means. [6 marks]Q3:

(a) Describe the difference between supervised and unsupervised learning. Give an example of a real world problem that requires a supervised learning algorithm and an example of a real world problem that can be solved with an unsupervised learning algorithm. In both cases define the inputs and the outputs. [8 marks]

(b) Describe in detail the steps of the K-means algorithm. Make sure that you define the input to the algorithm, the output, and the dimensionality of all the variables that you use. [8 marks]

(c) Identify the two sets of variables that are estimated by the K-means algorithm. Explain what coordinate descent (or coordinate optimisation) is. Using a sketch, show that this general optimisation method is warranted to converge. [4 marks]

(d) The K-Means algorithm converges to a local minimum. Describe a practical method to deal with this problem. Can this method be used to determine the optimal value of K? [5 marks]Q4:

(a) With the help of a diagram explain the main principles of the first-order Markov Model. In your answer explain any notation that you use. Explain what is meant by the term ‘’first-order”. [6 marks]

b) What are the differences between a Markov-Model and a hidden Markov model (HMM)? What are the advantages of HMMs in comparison to Markov Models? Give an example of an application (a toy example will suffice) where an HMM can be used but a Markov Model cannot. In your answer, define the states w\_i , the symbols v\_k , and the matrices A = [a\_ij] and B = [b\_jk] . [6 marks]

(c) Given a Hidden Markov Model, its states, the observation symbols and the transition and emission probabilities,

i. Give the formula for the joint probability P(w^{1:T}, V^{1:T}) of a sequence of observation symbols V^{1:T} = (v(1), v(2), …, v(t), v(T)) and a sequence of states w^{1:T} = (w(1), w(2), …, w(t), w(T)).

ii. Give the formula for the probability p(w^{1:T}) of a sequence of states w^{1:T} = (w(1), w(2), …, w(t), w(T)). [6 marks]

(d) What is the evaluation problem? Using the results of (c) present a naïve algorithm that solves the evaluation problem. What is the computational complexity of that algorithm? Can this algorithm be used in practice? [7 marks]