CITY
UNIVERSITY OF LONDON
EST 1894

School of Mathematics, Computer Science and Engineering

MSc Data Science

INM431 Machine Learning
PG Examination

[day][month] 2017 **XXXX – XXXX**

Answer **ALL** questions
Division of marks: Marks vary per question. Marks are indicated below.
Give your answer in the booklet provided, not in the exam paper

BEGIN EACH QUESTION ON A FRESH PAGE

Number of answer books to be provided: 1
Calculators permitted: Casio FX-83/85 MS/ES/GT+ ONLY
Examination duration: 120 minutes
Dictionaries permitted: None
Additional materials: None
Can question paper be removed from the examination room: No

External Examiner: XXXXXXXXXXXXXXXXXXX
Internal Examiner: Artur Garcez

Question 1

Answer with “True” or “False”

a) As the number of training examples goes to infinity your model trained on that data will have lower variance

b) The result of adding two Gaussians is always a Gaussian

c) The curse of dimensionality does not apply to Big Data

d) Clustering is an example of unsupervised learning

(20 marks)

Question 2

Describe how you would go about using least squares as objective function (aka cost function) to learn a set of parameters w for a machine learning task

(20 marks)

Question 3

Define the process of using a Naive Bayes classifier with a Bernoulli distribution for text data

(20 marks)

Question 4

Consider the DNA sequence:

AAGGTTCCAAGGTTCCAAGGTTCC

a. Draw the diagram of a Markov model showing the transitions between the DNA symbols.

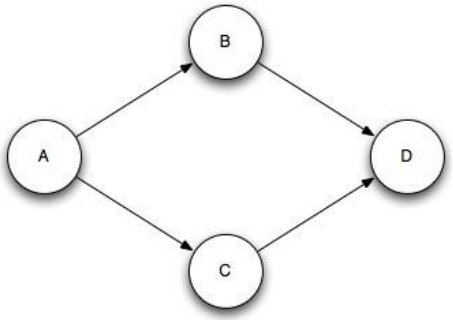
b. Calculate the prior probability for the Markov model.

c. Calculate the transition probability table based on the above sequence.

(20 marks)

Question 5

Consider the following directed graphical model:



a. Calculate the joint probability distribution for all random variables.

b. Give an example of how a Gaussian Mixture Model (GMM) could be applied to a task of your choice.

c. What quantity is computed during the E-step of training a GMM? What does that quantity indicate?

(20 marks)

CITY
UNIVERSITY OF LONDON
EST 1894

School of Mathematics, Computer Science and Engineering

MSc Data Science

INM431 Machine Learning
PG Examination

[day][month] 2016 **XXXX – XXXX**

Answer **ALL** questions
Division of marks: Marks vary per question. Marks are indicated below.
Give your answer in the booklet provided, not in the exam paper

BEGIN EACH QUESTION ON A FRESH PAGE

Number of answer books to be provided: 1
Calculators permitted: Casio FX-83/85 MS/ES/GT+ ONLY
Examination duration: 120 minutes
Dictionaries permitted: None
Additional materials: None
Can question paper be removed from the examination room: No

External Examiner: XXXXXXXXXXXXXXXXXXX
Internal Examiner: Artur Garcez

Question 1

When 5-fold cross-validation is applied on a dataset with 500 examples, how many models are trained with how many examples in each training set:

(a) 400 models and 5 examples

(b) 5 models and 400 examples

(c) 4 models and 500 examples

(d) 500 models and 4 examples

(e) None of the above

[5 Marks]

Question 2

Which **two sentences** best describe the curse of dimensionality:

(a) When the dimensionality increases, the volume of the space increases exponentially so that the available data may become sparse.

(b) When the dimensionality increases, PCA can be used to choose the most relevant dimensions.

(c) Sparsity is problematic in that it may lead to poor predictions without smoothness.

(d) Sparsity is no longer a serious problem because of Big Data.

[5 Marks]

Question 3

Choose **two tasks** to which it is most appropriate to apply a Hidden Markov Model:

(a) Film database used as part of a recommendation system

(b) Gene sequence data classification

(c) Stock market time series prediction

(d) Music database used as part of a recommendation system

(e) Stock market classification to buy or sell shares

[5 Marks]

Question 4

Which sentence best describes the main difference between regression and classification:

(a) Classification maps inputs to continuous values; regression maps inputs to discrete values.

(b) Classification maps inputs to {0,1}; regression maps inputs to an interval [0,1].

(c) Classification maps inputs to discrete values; regression maps inputs to continuous values.

(d) Classification maps inputs to an interval; regression maps inputs to {0,1,2,...}.

(e) Classification and regression are two sides of the same coin.

[5 Marks]

Question 5

For the K-nearest neighbour algorithm, is it true that larger K values lead to higher probability of overfitting? Briefly explain your answer.

(10 marks)

Question 6

Briefly describe how a HMM can be used for speech recognition. What are the observations and hidden states in this context? Why use an HMM for this task?

(10 marks)

Question 7

With the number of trees in a random forest increasing, would you expect variance to increase, decrease or stay roughly the same? Briefly explain why.

(15 marks)

Question 8

Consider the following k-means strategy for choosing k, the number of clusters: use different k values and choose the one that minimises the distortion criteria. Is this a good strategy? Briefly explain why.

(15 marks)

Question 9

Give an example of a decision tree. Describe how a collection of trees might be used for classification and regression.

(15 marks)

Question 10

Some patient features are expensive to collect (e.g. brain scans) whereas others are not (e.g. temperature). Therefore, you have decided to first ask your classification algorithm to predict whether a patient has a disease, and if the classifier is 80% confident that the patient has a disease, then you will ask for additional examinations to collect additional patient features In this case, which classification method do you recommend: decision tree or naive Bayes? Justify your answer.

(15 marks)