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:	XXXXXXXXXXXXXXXXXXXXXX
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 TRUE

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

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

d) Clustering is an example of unsupervised learning TRUE

(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)

Least squares: $E = ½ \sum_i (t_i - o_i)^2$, where t is the target value and o is the output of the model for each example i.

A machine learning task is to find a set of parameters w such that E is minimized. For example, under the assumption of a linear model, the task may be to find the values of w_0 and w_1 in the equation $o = w_0+w_1x$, where x is the set of examples, such that E is minimized.

To minimize E, we make the partial derivatives of E with respect to w_0 and w_1 equal to zero. In most ML tasks, this cannot be done analytically so that the computation needs to be approximated by selecting values for w_0 and w_1 according to the direction of the gradient and evaluating such choices incrementally; this can be efficient but it does not guarantee global optimization.

Question 3

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

(20 marks)

Naïve Bayes is a family of classifiers which makes an assumption of independence between the features when applying Bayes theorem.

Applied to text data, this means that the probability of occurrence of a word in a text is assumed not to depend on the occurrence of other words. This is clearly not the case though, so it’s not a valid assumption.

Naïve Bayes is a family of classifiers because it may apply Bayes theorem under different distributions when calculating p(x|C), that is the probability of the feature vector x given the document classes C.

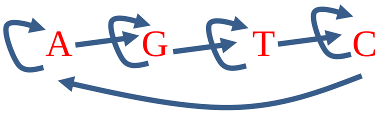
Following a Bernoulli distribution, features are independent boolean variables. In this case, instead of using e.g. word counts in the text to obtain x, each feature is in {0,1} depending on whether the word is present or not in the text.

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.

P(A)= 6/24

P(G)=6/24

P(T)=6/24

P(C)=6/24

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

P(A|A)=1/2 P(G|A)=1/2

P(G|G)=1/2 P(T|G)=1/2

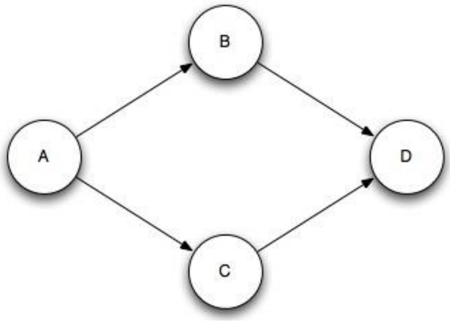
P(T|T)=1/2 P(C|T)=1/2

P(C|C)=3/5 P(A|C)=2/5

(20 marks)

Question 5

Consider the following directed graphical model:



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

P(A,B,C,D) = P(D|B,C) x P(B|A) x P(C|A) x P(A)

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

A Gaussian mixture model assumes that all the data points are generated from a combination of Gaussians. Typically:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta})$$

where θ are model parameters and π, the mixture weights, stands for *p*(*z*=*k*).

One way that the parameters of a mixture model can be learned under the assumption of maximum likelihood is to apply the Expectation-Maximisation algorithm. These can be regarded as types of unsupervised learning or clustering procedures.

For example, as a special case, GMM can be initialized using k-means clustering. A typical application therefore is image segmentation, that is, to partition (cluster) an image into segments (sets of pixels) for ease of analysis, e.g. to locate object boundaries in the image.

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

The expectation step:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

It calculates the expected value of the log likelihood function, with respect to the above conditional distribution under the current estimate of the parameters.

(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:	XXXXXXXXXXXXXXXXXXXXX
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

Answer: (b)

[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.

Answer: (a) and (c)

[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

Answer: (b) and (c)

[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.

Answer: (c)

[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)

See lecture notes 7, slide 10

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)

See lecture notes 9, slides 18 to 22

Part of speech tags are hidden; word sequences are observed, or…

Acoustic events are observed and you want to infer the presence of hidden words that have caused the acoustics.

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)

Decrease; c.f. coin example from lecture 1 and relate it with bagging in RFs.

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)

Define distortion w.r.t. mutual information and relate it with maximizing information gain. Notice it assumes a fixed number of values for k is given.

Question 9

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

(15 marks)

See lecture notes 1, slide 35, and relate with lecture notes 6, slides 12 and 13.

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)

Naïve Bayes since it is a probabilistic model; decision trees are deterministic.