COMP529 : Jan 2017 Exam Answers + Solutions

1. a) You are setting up a Hadoop cluster. The cluster comprises 8 computers.

i) Which of the four Vs of Big Data does Hadoop primarily focus on?

Volume

ii) What operating system would be used and why?

Unix

Because it's set up to make it easy to configure the cluster

iii) Why would you sensibly run the DataNode and TaskTracker daemons on the same computer?

They will need to communicate a lot

Communication on the same computer is fast

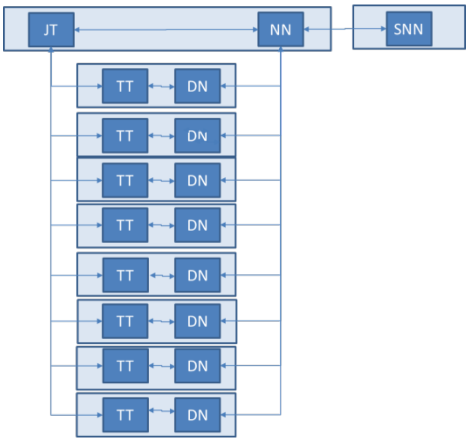
iv) Why would you sensibly avoid running the SecondaryNameNode and JobTracker Node daemons on the same computer?

jobtracker and Namenode both can run on the same computer

The SecondaryNameNode is a backup for the Namenode

Having the backup on the same computer as the Namenode would mean that both would stop running if that computer stopped working

v) Draw a diagram showing how you would allocate daemons to computers. Show communications paths between the daemons.



SecondaryNameNode on different computer to JobTracker

Job Tracker on same computer as Namenode

DataNodes each on the same computer as a TaskTracker

Communications shown between DataNodes and NameNode

Communications shown between JobTracker and TaskTracker

One NameNode. One JobTracker. One SecondaryNameNode

Multiple DataNodes. Multiple TaskTrackers

vi) What is the default number of replications of each block in a fully-distributed Hadoop cluster?

3

vii) HDFS is now running on the cluster and is configured to have 64MB blocks (with default settings for the number of replications of each block). How would a file that is 321MB in size be stored in the cluster?

6 blocks

Each replicated 3 times and distributed across the Namenodes

viii) Explain how the NameNode would know if one of the DataNodes was unplugged while processing one of these blocks.

The DataNode normally responds to the healthcheck signal sent by the Namenode.

If the DataNode is unplugged, the NameNode would get no response

ix) What is the name given to the configuration of Hadoop commonly used to debug MapReduce jobs?

Standalone mode

x) Why is java often used to describe MapReduce jobs?

It's a portable (so the code can go to the data)

xi) Provide a brief summary of a generic MapReduce job in terms of the inputs, the outputs and their interrelationship for each of the map and reduce tasks.

Input to map task is a set of values (accept key-value pairs)

Output of map task is a set of key-value pairs

The map task processes each input value to produce zero or more key-value pairs

Input to reduce task is a set of values for each key

Output of reduce task is a single value for each key

The reduce task processes each set of values to produce the single output value

b) i) In a Storm cluster, what name is given to the source of a stream of tuples?

A spout

ii) In a Storm cluster, what name is given to a component that takes (at least) one stream of tuples as input and outputs (at least) one stream of tuples?

A bolt

iii) In a Storm cluster, what do the Nimbus and Zookeeper each do?

The nimbus acts as the server.

The Zookeeper provides robustness

iv) In a Storm cluster, what is a topology?

A graph describing the processing of the streams (with spouts and bolts as nodes and streams as arcs).

2. You are building a system to analyse data related to crashes of autonomous cars for an insurance company.

a) Assuming that each individual car crash is an independent (instantaneous) random event and you wanted to use a conjugate prior, what distribution would you use for the likelihood (the number of car crashes that have happened) and for the posterior (your uncertainty related to the rate of occurrence of crashes)?

Likelihood is a Poisson distribution.

Posterior is Gamma distribution.

[Accept other correct answers (which the moderator implies to exist, but aren't discussed in the module or known to the setter)]

b) The number of crashes is N_1 and the rate of crashes is λ_1. Write down an expression for $p(N_1)$ in terms of the prior, $p(\lambda_1)$, and likelihood, $p(N_1|\lambda_1)$.

$$p(N_1) = \sum_{\lambda_1} p(\lambda_1)p(N_1|\lambda_1)$$

Sum (aka integral) correct

Summand correct

c) Using the answer from b) or otherwise, write down an expression for $p(\lambda_1|N_1)$ in terms of the prior, $p(\lambda_1)$, and likelihood, $p(N_1|\lambda_1)$ only.

$$p(\lambda_1|N_1) = \frac{p(\lambda_1)p(N_1|\lambda_1)}{\sum_{\lambda_1} p(\lambda_1)p(N_1|\lambda_1)}$$

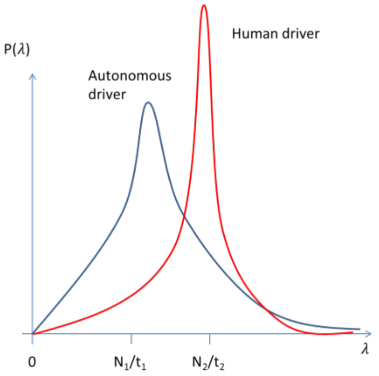
Denominator correct

Numerator correct

d) Autonomous cars have accumulated a total of t_1 driving hours. What is the average number of crashes by autonomous cars per hour?

N_1/t_1

e) In total, humans have had $N_2>>N_1$ crashes but have also accumulated a total of $t_2>>t_1$ driving hours. Draw a graph with two lines, one describing the uncertainty related to the rate of crashes of autonomous cars and the other describing the uncertainty related to the rate of crashes of cars driven by humans. Label the axes and lines. Ensure that you annotate the graph with the average number of crashes for the two kinds of car and that it is clear how the uncertainties associated with these two estimates compare. Assume that the uncertainty is described using the conjugate distribution you named in the answer to part c).



You should Labelled axes correctly.

Mean for humans is higher

Variance for humans is smaller

Labelled lines

Annotated averages

f) Explain why, even those autonomous cars might currently appear to be safer than cars driven by humans, there is a significant probability that they could eventually transpire to be less safe.

The uncertainty associated with the rate for autonomous cars is significant (relative to the difference between the two rates)

So, there is a much higher chance that the rate for autonomous cars is very different to the estimate

g) The posterior for *N* crashes during a time interval of *t*, can be approximated as being Normal with a mean of N/t and a variance of N/t^2. Data from the USA indicates that autonomous cars have had one crash in approximately 100 million hours of driving. Cars driven by humans have approximately 5 million crashes per year in approximately 5,000 billion hours of driving. Calculate the mean and standard deviation for both the posterior for autonomous cars and cars driven by vehicles. Comment on the ratio of the means relative to the ratio of the standard deviations

Autonomous cars:

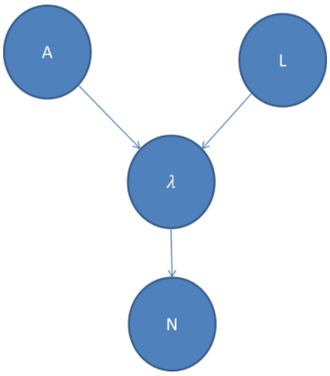
mean = $1/(100\text{x}10^{6})=10^{-8}$, s.d. = $\sqrt{}(1/(100\text{x}10^{6})^{2})=10^{-8}$

Human-driven cars:

mean = $5\text{x}10^{6}/(5{,}000\text{x}10^{9})=10^{6-12}=10^{-6}$, s.d. = $\sqrt{}(5\text{x}10^{6}/(5{,}000\text{x}10^{9})^{2})=4.5\text{x}10^{-10}$

Ratio of means is the same order as magnitude (2 orders of magnitude) as the ratio of standard deviations

h) Where a car drives affects the chance of a car crash. Draw a Bayesian Network that describes the relationship between the number of car crashes, *N*, the rate of car crashes, *λ*, the location of the cars, *L*, and whether the car is autonomous, *A*. Assume that the location of the car is not directly dependent on whether the car is autonomous.



Which is a Directed Acyclic Graph

4 labelled nodes

N connected to λ only

λ connected to A and L

A not connected to L

i) To perform inference in a very much larger version of this graph involving many contributory factors relating to the risk of a car crash, it is proposed to use Gibbs sampling, Belief Propagation or Mean Field. What would be the relative advantages of each technique in terms of their ability to be parallelised, the number of iterations required and any restrictions on the graph necessary to use the techniques? A tabular answer is acceptable.

	Ability to be parallelised	Number of iterations	Restrictions on graph
Gibbs	Y	Large	None
Belief Propagation	N	2 (accept small)	Must be a tree
Mean Field	Y	Small	None

One mark per correct entry

3. a) The stream of images comprising a CCTV video of a person is being processed using a Hidden Markov Model (HMM). The HMM's states represent whether, at the time of the current image, the person is standing still, walking left, walking right, running left or running right. The video is recorded at 25Hz such that the time between images is 4 milliseconds.

i) The current state is x_t and the history of states up to time t is $x_{1:t}$. The model assumed for the dynamics of the state is a Markov model. What does this mean in terms of $p(x_t|x_{1:t-1})$ and $p(x_t|x_{t-1})$?

They are the same

ii) Populate a matrix with plausible values for the transition matrix. Assume that people must walk for at least 4 milliseconds before they run, must stand still for at least 4 milliseconds before changing direction and that people in every state are most likely to be in the same state 4 milliseconds later.

	X_{t-1} =Run left	X_{t-1} =walk left	X_{t-1}=Still	X_{t-1}=Walk right	X_{t-1}=Run right
X_t=Run left	0.8	0.1	0	0	0
X_t=Walk left	0.2	0.8	0.1	0	0
X_t=Still	0	0.1	0.8	0.1	0
X_t=Run right	0	0	0.1	0.8	0.2
X_t=Walk right	0	0	0	0.1	0.8

Matrix size correct

Zeros correct

Diagonal is larger than sum of rest of column

Columns sum to unity

iii) The measurements, y_t, extracted from the image at time t, provide a likelihood, $p(y_t|x_t)$ which is conditionally independent of the previous states. Does knowing x_{t-1} as well as x_t affect the calculated likelihood?

No

iv) The output from the previous time-step is $p(x_{t-1}|y_{1:t-1})$. What is the minimum number of floating point numbers needed to be stored to describe this output?

4 (=5 states -1 (since you know they add up to 1)

v) Express $p(x_t|y_{1:t-1})$ in terms of $p(x_{t-1}|y_{1:t-1})$ and $p(x_t|x_{t-1})$.

$$p(x_t|y_{1:t-1}) = \sum_{x_{t-1}} p(x_{t-1}|y_{1:t-1})p(x_t|x_{t-1})$$

Correct terms used

Sum correct

Summand correct

vi) What matrix operation can be used to implement this expression?

Matrix multiply

vii) Express $p(x_t|y_{1:t})$ in terms of $p(x_t|y_{1:t-1})$ and $p(y_t|x_t)$.

$$p(x_t|y_{1:t}) = \frac{p(x_t|y_{1:t-1})p(y_t|x_t)}{p(y_t|y_{1:t-1})}$$

Correct terms used

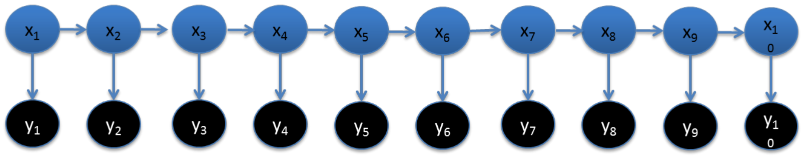
Numerator correct

Denominator correct

viii) What name is often given to this equation relating $p(x_t|y_{1:t})$ and $p(y_t|x_t)$?

Bayes rule

xi) Draw the Bayesian Network for $p(x_{1:10}|y_{1:10})$. Do not use plates. Be careful to indicate which nodes are observed and unobserved.



All nodes must be labelled as shown

10 observed nodes

10 unobserved nodes

Markov model represented

Conditional independence of likelihoods represented

b) i) Name three algorithms for sequential Bayesian estimation of a continuous state using nonlinear non-Gaussian models. For each algorithm, describe the approximation used.

Extended Kalman filter
Which uses a local linearization based on a Taylor series
Unscented Kalman filter
Which uses a local linearization based on finite difference
Particle filter
Which approximates the distribution with samples

ii) Write an equation describing the likelihood model used by a Kalman filter when processing *M*-dimensional data to make inferences about an *N*-dimensional state. Define the size of any matrices used in the models in terms of *M* and *N*.

$$y = Hx_t + w_t$$

Equations is linear

Equation is correct

$$H \text{ is } M \text{ x } N$$