January 2018 Exam- Answers and Solutions

1) a)

i) Which V is Hadoop primarily designed to cater for?

Volume

ii) Where does a programmer articulate that Hadoop is running in standalone mode?

Xml Files

iii) What are the names of the daemons that act as master and slave in a cluster running Hadoop? Be clear which are which.

Master: TaskTracker

Slave: JobTracker

Master: NameNode

Slave: DataNode

iv) What is the name of the daemon responsible for maintaining a backup of the information describing where each file is being stored?

Secondary NameNode

v) Which feature of Hadoop makes it necessary to use a portable programming language such as java?

The code is sent to the data

vii) What does HDFS stand for?

Hadoop Distributed File System

viii) A HDFS system uses 64 bit addresses and each block is 64 Megabytes. What is the maximum volume size for this HDFS system?

64 MB

ix) In what ways does a FAT32 hard disk differ from such an HDFS system?

32 bit addresses

Sectors are smaller than 64 MB

x) What are the inputs to and outputs from a reducer within a MapReduce job?

Input: key, list of values

Input: keys are unique

Output: key, single value

Output: keys are unique

b) i) Which V is Storm primarily designed to cater for?

Velocity

ii) Where does a programmer articulate that Storm is being run in local mode?

In the java

iii) What are the names of the daemons that act as master and slave in a cluster running Storm? Be clear which is which.

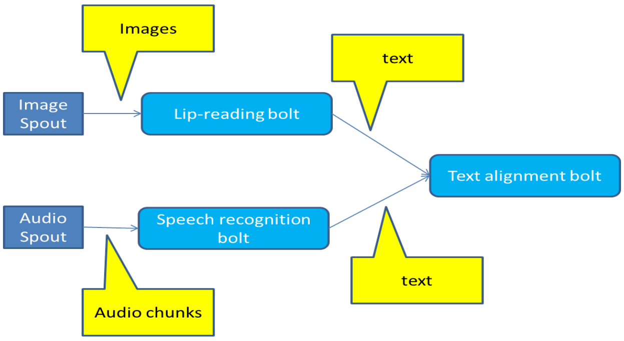
Master: Nimbus

Slave: supervisor

iv) What is the name of the daemon responsible for ensuring the stability of a Storm cluster?

Zookeeper

v) A topology comprises two spouts and three bolts. Assume one spout generates a stream of images and the other spout generates a stream of 30 millisecond audio chunks. Assume one bolt performs lip-reading, one performs speech recognition and the third bolt aligns two streams of text. Draw a diagram describing the topology. Label all spouts and bolts. Annotate all streams with the information being transmitted.



Spouts correct

Bolts correct

Arcs correct

Streams correctly labelled

vi) Name a middleware product that is not Storm and that is designed to support streaming analysis.

Flume

c) Aside from Volume and Velocity, state the other two Vs of Big Data and explain what they each mean.

Veracity

Which is trusting the output

Variety

Which is the unstructured nature of the input

2.

a) i) How many numbers would be needed to store P(T,S,U,D) in general?

16

ii) Write P(T|D) in terms of P(D|T) and P(D).

$$P(T|D) = \frac{P(D|T)\,P(T)}{P(D)}$$

iii) How many additions and multiplications are needed to calculate P(T|D) in this general case when D is true and T is true?

$$\mathrm{P(T|D)} = \frac{\sum_{S,U} P(T,S,U,D)}{\sum_{S,U,T} P(T,S,U,D)}$$

There are 4 elements in the sum in the numerator and 8 in sum in the denominator

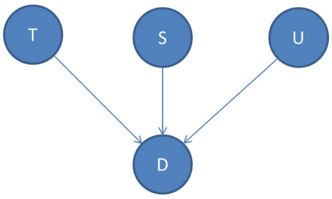
So, (4)+(8) = 12additions (and no multiplications)

iv) How many further mathematical operations are needed to calculate P(T|D) when D is true and T is false?

1

b)

i) Draw a Bayesian Network to describe this postulated model. Clearly label all nodes.



4 nodes shown

Nodes (including labels) correct

ii) How many numbers would be needed to store the probabilities parameterising the postulated model?

1 number for each of P(T), P(U) and P(S)

1 number for each of the 8 combinations of S,T and U for P(D|S,T,U)

iii) Write P(D|T) in sum-product form for this model.

$P(D|T) = \sum_{S,U} P(U)\, P(S)\, P(D|S,T,U)$

iv) Write P(T|D) as a ratio of two sum-products.

$$P\ (D) = \sum_{S,U,T} P\ (T)\ P\ (U) P\ (S) P\ (D|S,T,U)$$

$$\mathrm{P(T|D)} = \frac{\sum_{S,U,T}\ P(T)\, P(U)\, P(S)\, P(D|S,T,U)}{\sum_{S,U} P(U)\, P(S)\, P(D|S,T,U)}$$

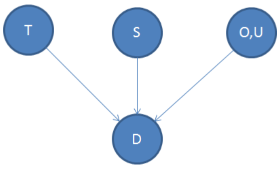
v) How many additions and multiplications are needed to calculate P(T|D) using this postulated model when D is true and T is true?

3 multiplications for each of 8 terms in the numerator and 2 multiplications for each of 4 terms in the denominator: 3x8+2x4 = 24

7 additions for the numerator and 3 additions for the denominator: 7+3=10

c) Another model assumes that P(T,O,S,U,D)=P(T)P(O,U)P(S)P(D|S,T,U).

i) Draw a Bayesian Network to describe this model. Clearly label all nodes.



ii) Write an expression for P(T,O,S,U,D) in terms of P(O|U).

P(O,U) = P(O|U)P(U)

So: P(T,O,S,U,D)=P(T)P(O|U)P(U)P(S)P(D|S,T,U).

iii) Describe (in words not equations) the operation of belief propagation in the context of this model.

Use a downward pass from the topmost node to the bottom-most node and then an upward pass from the bottom-most node to the upmost node

Consider each node in turn during each pass

Combine the outputs from the two passes for each node

iv) Describe (in words not equations) the operation of Pearl's algorithm in the context of this model.

For each node, send a downward message to its neighbours assuming the messages it previously received from the neighbours above were the result of a downward pass of belief propagation

Do the same thing for the upward messages

Repeat this process for some finite number of iterations

3. **a)** You are part of a team building a predictive text system for a MOOC that helps people to learn Swedish. The system has been configured to consider a dictionary of 600,000 unique words. A Hidden Markov Model (HMM) is to be used to process the incoming stream of text. The Levenshtein distance is to be used as a way of quantifying the (non-negative) number of changes (e.g., additions and deletions) to one string to transform it into another string. The likelihood is defined, using Levenshtein distance, such that the likelihood is high for words in the dictionary that are similar to the current text string and low for words that are dissimilar to the current text string. The strings that have been processed up to and including now are $y_{1:t}$ and the true current word is x_t.

i) How many numbers are used to parameterise are the transition matrix?

600,000 x 600,000

ii) What two reasons would motivate the approximate the transition matrix as sparse?

It'd be faster

It'd require less storage

iii) What would be a disadvantage of such an approximation?

It would degrade accuracy

iv) Write an equation that expresses the fact that the state is Markov.

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

v) Write an equation that expresses the fact that the current state is a sufficient statistic of the past in terms of predicting the measurement.

$p(y_t|x_{1:t},y_{1:t-1}) = p(y_t|x_t)$

vi) Write an equation for $p(x_t|y_{1:t})$ in terms of $p(x_{t-1}|y_{1:t-1})$, $p(x_t|x_{t-1})$ and $p(y_t|x_t)$. Clearly label the posterior, the prior, the likelihood and the dynamic model.

$[p(x_t|y_{1:t}) = \Sigma_{xt-1}p(x_{t-1}|y_{1:t-1})\ p(x_t|x_{t-1})\ p(y_t|x_t) / p(y_t|y_{1:t-1})]$

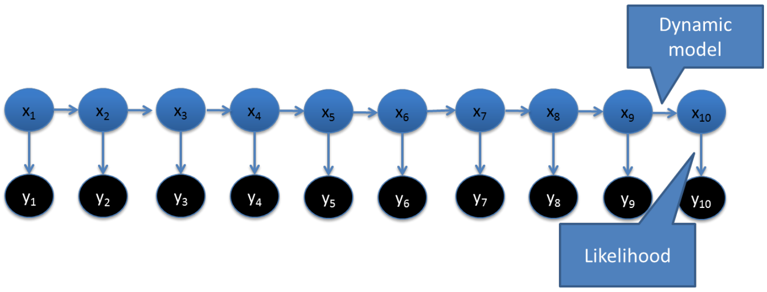
$p(x_t|y_{1:t})$ labelled as posterior

$p(x_{t-1}|y_{1:t-1})$ labelled as prior

$p(y_t|x_t)$ labelled as likelihood

$p(x_t|x_{t-1})$ labelled as dynamic model

vii) Draw a Bayesian Network for the joint distribution of 10 true words and 10 observed text strings. Clearly label one arc that represents a likelihood and one that represents an instance of the dynamic model.



viii) Why would the output from the current time-step not be the same as the most likely word given the most likely word from the previous time-step?

There is uncertainty over the previous word

This uncertainty could change the most likely word at the current time-step

b) i) What properties of the dynamics and likelihood are such that the Kalman filter exactly characterises the current uncertainty in the state given some historic data?

Linear and Gaussian

ii) What two parameters are passed between consecutive iterations of a Kalman filter?

Mean

Covariance

iii) What approximation is used by each of the Extended Kalman filter, Unscented Kalman filter and particle filter?

Extended Kalman filter: linearisation based on Taylor series

Unscented Kalman filter: linearisation based on sigma points

Particle filter: random samples