# CPSC 540 Assignment 3 (due Friday March 12 at midnight)

1. Name(s):
2. Student ID(s):

## 1 Markov Chains

### 1.1 Inference with Discrete States

The function *example_markovChain.jl* loads the initial state probabilities and transition probabilities for a Markov chain model,

$$p(x_1, x_2, \dots, x_d) = p(x_1) \prod_{j=2}^{d} p(x_j \mid x_{j-1}),$$

corresponding to the "grad student Markov chain" from class.

1. Write a function, *sampleAncestral*, that uses ancestral sampling to sample a sequence $x$ from this Markov chain of length $d$. Hand in this code and report the univariate marginal probabilities for time 50 using a Monte Carlo estimate based on 10000 samples.
   Hint: you can use *sampleDiscrete* in *misc.jl* to sample from a discrete probability mass function using the inverse transform method.
2. Write a function, *marginalCK*, that uses the CK equations to compute the exact univariate marginals up to a given time $d$. Hand in this code, report all exact univariate marginals at time 50, and report how this differs from the marginals in the previous question.
3. What is the state $c$ with highest marginal probability, $p(x_j = c)$, for each time $j$?
4. Write a function, *viterbiDecode*, that uses the Viterbi decoding algorithm for Markov chains to find the optimal decoding up to a time $d$. Hand in this code and report the optimal decoding of the Markov chain up to time 50 and up to 100.

5. Report all the univariate conditional probabilities at time 50 if the student starts in grad school, $p(x_{50} = c \mid x_1 = 3)$ for all $c$. Hint: you should be able to do this by changing the input to the CK equations.

6. Report for all $c$ the univariate conditional probabilities $p(x_5 = c \mid x_{10} = 6)$ ("where you were likely to be 5 years after graduation if you ended up in academia after 10 years") obtained using a Monte Carlo estimate based on 10000 samples and rejection sampling. Also report the number of samples accepted among the 10000 samples.

7. Give code implementing a dynamic programming approach to exactly compute $p(x_5 = c \mid x_{10} = 6)$, and report the exact values for all $c$.

8. Why is $p(x_j = 7 \mid x_{10} = 6)$ equal to zero for all $j$ less than 10?

Hint: for some of the quesitons you may find it helpful to use a $k$ by $d$ matrix $M$ to represent a dynamic programming table

## 1.2 Inference with Gaussian States

Consider a continuous-state Markov chain where the initial distribution is given by

$$x_0 \sim \mathcal{N}(m_0, v_0^2),$$

and the transition distributions for $j > 1$ are given by

$$x_j \mid x_{j-1} \sim \mathcal{N}(w_j x_{j-1} + m_j, v_j^2).$$

This model could be used to model an object moving through $\mathbb{R}$.[1] Because of the Gaussian assumptions, this defines a joint Gaussian distribution over the variables while the marginal distributions are also Gaussian. For a generic $j > 1$, derive the form of the marginal distribution of $x_j$, expressing the marginal parameters $\mu_j$ and $\sigma_j$ recursively in terms of the parameters $\mu_{j-1}$ and $\sigma_{j-1}$ of the previous marginal, $p(x_{j-1}) \sim \mathcal{N}(\mu_{j-1}, \sigma_{j-1}^2)$.

Hint: You can use Theorem 4.4.1 of Murphy's book.

[1] In practical applications like object tracking, we typically have that the states $x_j$ are 2- or 3-dimensional if we are modeling an object moving through space, or even higher-dimensional if we are modeling things like stock prices.

### 1.3 Learning with Discrete States

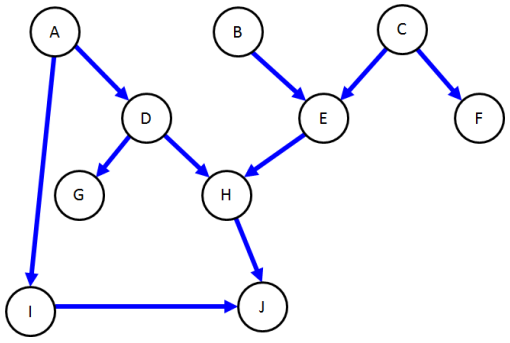
If you run *example_rain* it will: load the Vancouver rain data set, split it into a training and validation set, fit an indepdendent (and homogeneous) Bernoulli model to the training set, and then compute the negative log-likeilhood (NLL) of this model on the validation set (a lower validation NLL means a better fit). As discussed in class, we expect that a Markov chain could be a better model of this dataset.

1. Give code for finding the MLE for the initial probabilities and transition probabilities in a homogeneous Markov chain, and report the MLE values for the training set.
2. Report the NLL of the Markov chain model on the validation set.

# 2 Directed Acyclic Graphical Models

## 2.1 D-Separation

Consider a directed acyclic graphical (DAG) model with the following graph structure:

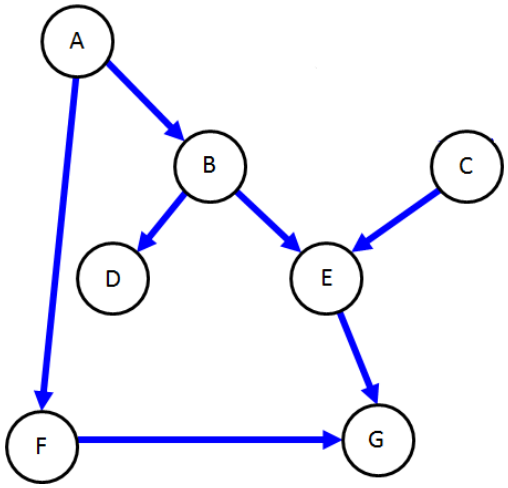



Assuming that the conditional independence properties are faithful to the graph, using d-separation briefly explain why the following are true or false:

1. $H \perp I$.
2. $H \perp I \mid A$.
3. $H \perp I \mid J$.
4. $H \perp I \mid A, J$.
5. $C \perp I$.
6. $C \perp I \mid J$.
7. $C \perp I \mid H$.
8. $C \perp I \mid A, H$.
9. $C \perp I \mid E, H, J$.

## 2.2 Exact Inference

Consider a directed acyclic graphical (DAG) model with the following graph structure:




Assume that all variables are binary and that we use the following parameterization of the network:

$$
\begin{aligned}
&p(A = 1) = 0.7 \\
&p(B = 1 \mid A = 0) = 0.8 \\
&p(B = 1 \mid A = 1) = 1.0 \\
&p(C = 1) = 0.8 \\
&p(D = 1 \mid B = 0) = 0.8 \\
&p(D = 1 \mid B = 1) = 0.6 \\
&p(E = 1 \mid B = 0, C = 0) = 0.3 \\
&p(E = 1 \mid B = 0, C = 1) = 0.7 \\
&p(E = 1 \mid B = 1, C = 0) = 0.4 \\
&p(E = 1 \mid B = 1, C = 1) = 0.5 \\
&p(F = 1 \mid A = 0) = 0.5 \\
&p(F = 1 \mid A = 1) = 0.9 \\
&p(G = 1 \mid E = 0, F = 0) = 0.5 \\
&p(G = 1 \mid E = 0, F = 1) = 0 \\
&p(G = 1 \mid E = 1, F = 0) = 0.1 \\
&p(G = 1 \mid E = 1, F = 1) = 0.1
\end{aligned}
$$

Compute the following quantities:

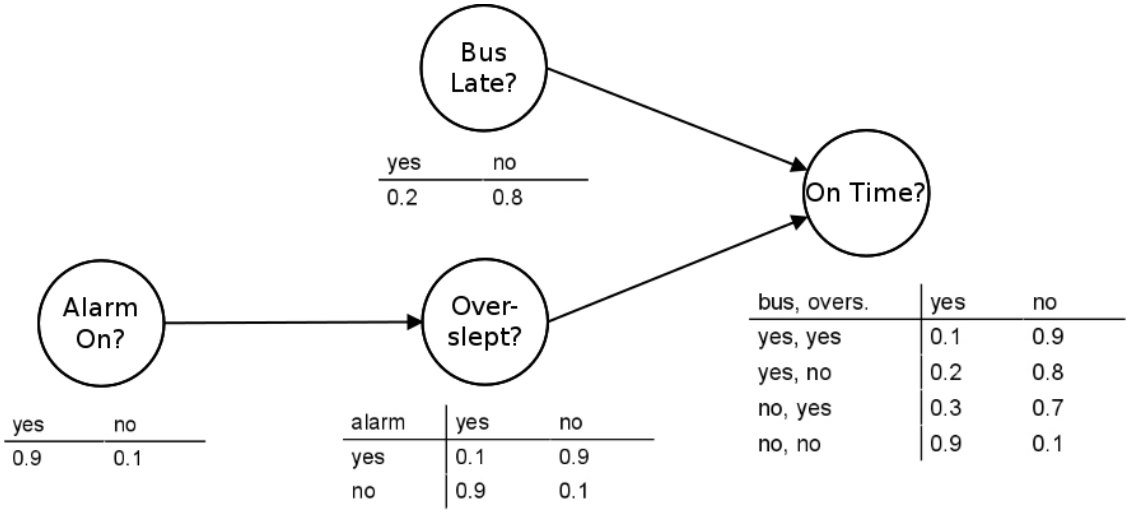
1. $p(A = 0)$.
2. $p(B = 1 \mid A = 0)$.
3. $p(B = 1)$.
4. $p(D = 1)$.
5. $p(B = 1 \mid D = 1)$.
6. $p(B = 1 \mid C = 1)$.

7. $p(B = 1 \mid A = 0, C = 1, F = 1)$.

Hints: some of the above quantities can be read from the table, some require using that probabilities sum to 1, some require the marginalization rule, some require Bayes rule, some require using [conditional] independence, and some will be simplified using calculations from previous sub-questions.

## 2.3 Learning in DAGs

The file *onTime.jld* contains a matrix $X$ containing samples from the following DAG model:[2]




The first column of $X$ is the "alarm" variable, the second is "bus late", the third is "over-slept", and the last is "on time".

1. Assuming the DAG structure above, give code for computing the MLE of the parameters in the model, and report the MLE values up to 2 decimal places.
2. Give code for generating samples using the MLE parameters.
3. The combination $(0, 1, 0, 1)$ does not occur in the training data, so if we fit the MLE general discrete distribution to this data we would get that it has a probability of 0. This is in contrast to the true model above, where we have $p(0, 1, 0, 1) = 0.0004$. Does your model give a better or worse estimate of the true proability of this event than the general discrete distribution? Why do you think it gives a better/worse estimate?

[2] from here: `https://www.uib.no/en/rg/ml/119695/bayesian-networks`

# 3 Undirected Graphical Models

Cancelled.

# 4 Very-Short Answer Questions

Give a short and concise 1-sentence answer to the below questions.

1. What is the difference between computing marginals and computing the stationary distribution of a Markov chain.
2. What is the inverse transform method used for?
3. Describe how we could use ancestral sampling to sample from the joint density over $x$ and $y$ defined by a Gaussian discriminant analysis model.
4. Suppose you had a black box that could generate IID samples from a distribution. Describe how you could use a Monte Carlo method to approximate $p(x \leq c)$ for this distribution.
5. What is the cost of generating a sample from a Markov chain of length $d$ with $k$ possible states for each time? What is the cost of decoding?

6. What is the difference between inference and decoding in Markov chains?
7. Suppose we are using a hidden Markov model to track the location of a submarine using sonar measurements. What would $x_j$ and $z_j$ represent in this example?
8. What is "explaining away"?
9. If two variables are not d-separated, are they necessarily depenent? If two variables are d-separated, are they necessarily independent?
10. What is an advantage and a disavantage of using logistic regression to parameterize the CPDs in DAGs compared to a tabular representation?

11. When decoding a DAG, why does the order that we compute the messages matter?
12. What is the relationship between multivariate Gaussians and UGMs?
13. What is the relevance of the Markov blanket in ICM?
14. Why might we do "thinning" of the samples when we use Gibbs sampling?

# CPSC 540 Relevant Papers for Project

This section is only to be done by students enrolled in CPSC 540. For students in CPSC 440, the description for your project will come with Assignment 4.

## Finding Relevant Papers

To help you make progress on your project, for this part you should hand in a list of 10 academic papers related to your current project topic. Finding related work is often one of the first steps towards getting a new project started, and it gives you an idea of what has (and has not) been explored. Some strategies for finding related papers are:

1. Use Google: try the keywords you think are most relevant. Asking people in your lab (or related labs) for references is also often a good starting point.
2. Once you have found a few related papers, read their introduction section to find references that these papers think are worth mentioning.
3. Once you have found a few related papers, use Google Scholar to look through the list of references that are *citing* these papers (particularly for recent ones). You may have to do some sifting if there are a lot of citations. Reasonable criteria to sift through large reference lists include looking for the ones with the most citations or focusing on the more recent ones (then returning to Step 2 to find the more-relevant older references).

For this question you only need to provide a list, but in the final assignment you will have to do a survey of at least 10 papers. So it's worth trying to identify papers that are both relevant and important at this point. For some types of projects it will be easier to find papers than others. If you are having trouble, post on Piazza.

Although the papers do not need to all be machine learning papers, the course project does need to be related to machine learning in some way, so at least a subset of the papers should be machine learning papers. Here is a rough guide to some of the most reputable places to where you see machine learning works published:

- The International Conference on Machine Learning (ICML) and the conference on Advances in Neural Information Processing (NeurIPS) are the top places to publish machine learning work. The Journal of Machine Learning Research (JMLR) is the top journal, although in this field conference publications are usually viewed as more prestigious.
- Other good venues include AISTATS (emphasis on statistics), UAI (emphasis on graphical models), COLT (emphasis on theory), ICLR (emphasis on deep learning), ECML-PKDD (European version of ICML), CVPR and ICCV/ECCV (emphasis on computer vision), ACL and EMNLP (emphasis on language), KDD (emphasis on data mining), AAAI/IJCAI (emphasis on AI more broadly), JRSSB and Annals of Stats (emphasis on statistics more broadly), and Science and Nature (emphasis on science more broadly).

## Paper Review

Among your list of 10 papers, choose one paper and write a short review of this paper. It makes sense to choose a paper that is closely-related to your project or to choose one of the most important-looking papers. The review should have two parts:

1. A short summary of the contributions of the paper. Say what problem the paper is addressing, why this is an important problems, what is proposed, and how it is being evaluated.
2. A list of strengths and weaknesses of the paper, and whether the claims are appropriately evaluated. For ideas of what issues to discuss, see the JMLR guidelines for reviewers:

`http://www.jmlr.org/reviewer-guide.html`

Note that you should include a non-empty list of strengths *and* weaknesses. Many students when doing their first reviews focus either purely on strengths or purely on weaknesses. It's important to recognize that all papers have weaknesses or limitations (even ones written by famous people or that are published in impressive places or that proved to be historically important) and all papers have strengths or at least a motivation (the authors must have thought it was worth writing for some reason).