Solutions

10-418/10-618 ML for Structured Data **Name:** ________________________
Fall 2022
Practice Problems 1
October 10, 2022
Time Limit: – minutes **Andrew ID:** ________________________

Instructions:

- Please fill in your name and Andrew ID above. Be sure to write neatly, or you may not receive credit for your exam.
- This exam contains 17 pages (including this cover page). There are 22 questions. The total number of points is 63.
- You are allowed to use one page of notes.
- If you have made a mistake, cross out the invalid parts of your solution, and circle the ones which should be graded.
- Look over the exam first to make sure that none of the 17 pages are missing. The problems are of varying difficulty, so you may wish to pick off the easy ones first.
- No electronic devices may be used during the exam.
- Please write all answers in pen or *darkly* in pencil.
- You have – minutes to complete the exam. Good luck!

Instructions for Specific Problem Types

For “Select One” questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

~~●~~ Noam Chomsky

For “Select all that apply” questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don’t know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

~~■~~ I don’t know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-418

10-~~4~~618

1 Learning to Search

Suppose we are using DAgger to train a model on the search space giving in Figure 1. We assume there is a one-to-one correspondence between states and actions (e.g. from state B there are two actions D and E). State A is the start state. The cost of each leaf node is given in parentheses (e.g. the cost of H is 7). The current model's score of each action is labeled on the edge (e.g. at state B, the score of action D is 4). Below assume the model policy is the greedy policy induced by the model scores. *(Hint: for score higher is better and for cost lower is better.)*

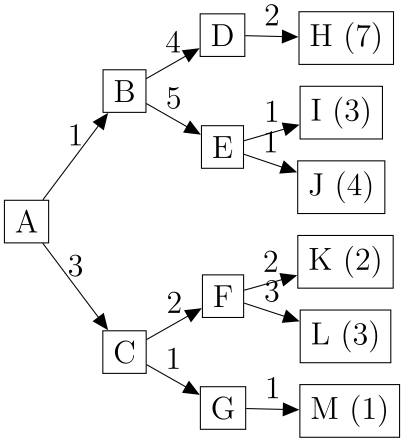


Figure 1

1. (2 points) **Numerical answer:** What is the oracle completion cost of node A?

   □

   1

2. (2 points) **Numerical answer:** What is the oracle completion cost of node B?

   □

   3

3. (2 points) **Short Answer:** What is the sequence of actions taken by the oracle policy from state A?

   □

   C, G, M

4. (2 points) **Short Answer:** What is the sequence of actions taken by the oracle policy from state B?

E, I

5. (2 points) **Short Answer:** What is the sequence of actions taken by the model policy from state A?

   C, F, L

6. (2 points) **Short Answer:** Suppose we are training with the Traditional Supervised Approach, what is the sequence of (state, action) pairs collected for training the model? *Give only one trajectory.*

   (A, C), (C, G), (G, M)

7. (2 points) **Short Answer:** Suppose we are training with DAgger, what is the sequence of (state, action) pairs collected for training the model? *Give only one trajectory. Assume the model scores are fixed during collection. Assume DAgger follows the model policy with probability 1.*

   (A, C), (C, G), (F, L)

8. (2 points) **Select all that apply:** Which of the following classification methods could we use to define and learn a model policy for DAgger?

   □ Decision Tree

   □ Multinomial Logistic Regression

   □ One-vs-All Classifier using SVM

   □ None of the above

   All of them.

9. (1 point) **True or False:** One advantage of Teacher Forcing is that (unlike Scheduled Sampling) it exposes the model to its own mistakes at training time, so that the model can learn to correct for them.

   ○ True

   ○ False

   False. It's the other way around: SS exposes the model to its own mistakes.

2 Sequence to Sequence

1. Suppose you wish to use a recurrent neural network language model (RNNLM) to compute the log-likelihood of a sequence of words $w_1, w_2, \ldots, w_T$.

   (a) (1 point) **True or False:** To properly compute the log-likelihood of the sequence, we take the most probable word at the current timestep to be the input for the next timestep.

   ○ True

   ○ False

   False. We only feed in the sentence itself.

   (b) (1 point) **True or False:** The first token fed into the RNNLM is w_1.

   ○ True

   ○ False

   False. The first token fed into the RNNLM must be a start-of-sentence token or the like.

   (c) (1 point) **Fill in the blank:** *The output softmax at timestep t represents ____.*

   ○ the conditional probability distribution over w_t given w_{t-1}

   ○ the conditional probability distribution over w_t given $w_{t-1}, w_{t-2}, \ldots, w_1$

   ○ the marginal probability distribution over w_t

   ○ the marginal probability distribution over w_{t-1}

   ○ None of the above

   conditional probability distribution over w_t given $w_{t-1}, w_{t-2}, \ldots, w_1$

2. A recurrent seq2seq model feeds an input sequence $[\mathbf{a}_1, \mathbf{a}_2, \ldots]$ to a uni-directional LSTM encoder to obtain a hidden state sequence $[\mathbf{b}_1, \mathbf{b}_2, \ldots]$. The last hidden state of $\mathbf{b}_{LAST}$ is fed in as the zero'th hidden state of an RNNLM decoder (i.e. $\mathbf{c}_0$). The RNNLM decoder produces the hidden sequence $[\mathbf{c}_1, \mathbf{c}_2, \ldots]$, a sequence of output probability distributions $[\mathbf{d}_1, \mathbf{d}_2, \ldots]$, and a sequence of predictions $[\hat{y}_1, \hat{y}_2, \ldots]$. The ground truth output sequence is $[y_1^*, y_2^*, \ldots]$. Assume we train to minimize cross-entropy loss, given by the sum of a sequence of losses, one per output timestep, $[\ell_1, \ell_2, \ldots]$. All sequences have finite length.

   (Note: for full credit in the questions below, you must list the smallest correct set of intermediate quantities.)

   (a) (2 points) **Fill in the blank:** The encoder hidden state at time t, $\mathbf{b}_t$, is a function of ____.

$\mathbf{b}_{t-1}, \mathbf{a}_t$

(b) (2 points) **Fill in the blank:** The decoder hidden state at time t, $\mathbf{c}_t$, is a function of ____.

$\mathbf{c}_{t-1}, y^*_{t-1}$

(c) (2 points) **Fill in the blank:** The decoder output probability distribution at time t, $\mathbf{d}_t$, is a function of ____.

$\mathbf{c}_t$

(d) (2 points) **Fill in the blank:** The loss at time t, ℓ_t, is a function of ____.

$\mathbf{d}_t, y^*_t$

3 Probabilistic Graphical Models

1. (1 point) **True or False:** All valid potential functions for an undirected graphical model can be expressed as $\psi(\mathbf{x}) = e^{f(\mathbf{x})}$ for $f : \mathcal{X} \to \mathbb{R}$.

   ○ True

   ○ False

   False, because potential functions are allowed to equal zero.

2. Consider the pairwise Markov Random Field (MRF) with the following joint distribution:

$$p(a, b, c, d) = \frac{1}{Z}\psi(a)\psi(b)\psi(c)\psi(d)\psi(a, b)\psi(a, c)\psi(b, c)\psi(c, d)\psi(d, a)$$

   (a) (2 points) **Drawing:** Draw the undirected graphical model corresponding to this pairwise MRF.

   a UGM with all pairs connected except for (b,d)

   (b) (2 points) **Drawing:** Draw the factor graph corresponding to this pairwise MRF.

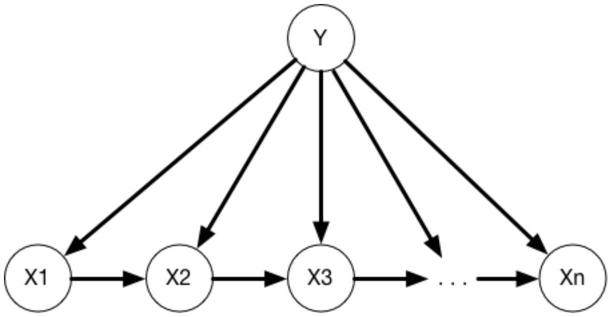
   a pairwise factor between all pairs except for (b,d), unary factor on each variable.

3. Suppose we want to design a classifier that predicts a single class y for the features $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$. The model is a Bayesian Network. The underlying assumption of this model is a feature x_i is conditionally independent of all other features given y, x_{i-1}, and x_{i+1}. Assume that feature x_1 is conditionally independent of all other features given just y and x_2.

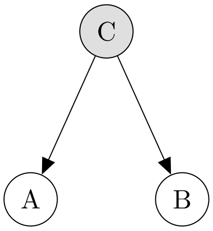
   (a) (2 points) **Short Answer:** Write down the joint probability of y and $x_1, \ldots, x_n$ as product of conditional probabilities that preserves the conditional independence assumptions provided. Each conditional probability term should rely on as few variables as possible.

$p(y, x_1, \ldots, x_n) = p(y)p(x_1|y) \prod_{i=2}^{n} p(x_i|y, x_{i-1})$

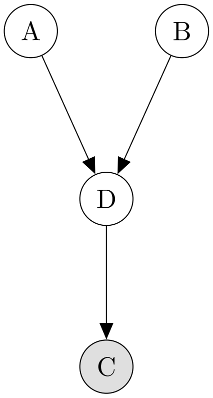
(b) (2 points) **Drawing:** Draw a graphical model that represents the conditional independence assumptions of this classifier.



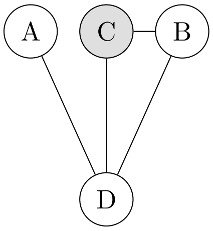
4. (3 points) Consider the following graphical models.



Graph (a)



Graph (b)



Graph (c)

(a) **True or False:** In graph (a), **A** and **B** are conditionally independent given **C**.

○ True

○ False

True, because nodes are conditionally independent given their Markov Blanket.

(b) **True or False:** In graph (b), **A** and **B** are conditionally independent given **C**.

○ True

○ False

False.

(c) **True or False:** In graph (c), **A** and **B** are conditionally independent given **C**.

○ True

○ False

False.

4 Marginal Inference

1. Consider the behavior of the Variable Elimination algorithm on the factor graph with a lattice-like structure in Figure 4.

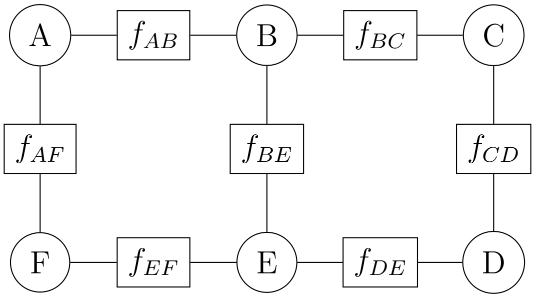


Figure 2

(a) (2 points) **Short Answer:** Write the equation for the factor computed when eliminating A first.

$\psi_{FB} = \sum_A \psi_{AB}\psi_{AF}$

(b) (2 points) **Short Answer:** Write the equation for the factor computed when eliminating B first.

$\psi_{AEC} = \sum_B \psi_{AB}\psi_{BC}\psi_{BE}$

(c) (1 point) **Select One:** Based on the above equations, which node can be eliminated first more efficiently?

○ A

○ B

○ Doesn't matter which node is eliminated first

A

2. Consider an application of belief propagation to the acyclic factor graph in Figure 3 and answer the following questions.

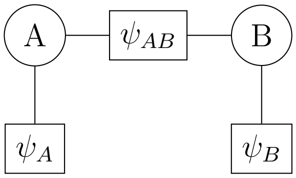


Figure 3

a	$\psi_A(a)$
0	1
1	2

b	$\psi_B(b)$
0	2
1	1

a	b	$\psi_{AB}(a,b)$
0	0	1
0	1	1
1	0	1
1	1	2

(a) (2 points) **Numerical Answer:** What is the message from ψ_{AB} to A?

a	$\mu_{\psi_{AB}\rightarrow A}(a)$
0	3
1	4

(b) (2 points) **Numerical Answer:** What is the node belief at A? *(Your answer should be the unnormalized belief.)*

a	$b_A(a)$
0	3
1	8

(c) (2 points) **Numerical Answer:** What is the marginal distribution $P(A = a)$?

a	$P(A = a)$
0	$3/11$
1	$8/11$

(d) (2 points) **Numerical Answer:** What is the partition function Z?

11

3. Answer the following questions based on your understanding of the differences between variable elimination and belief propagation:

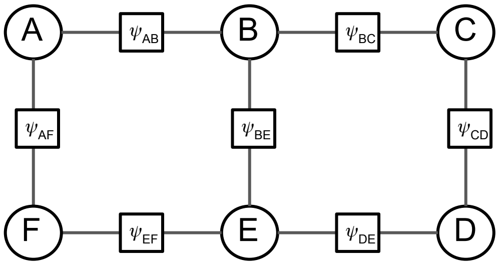


Figure 4

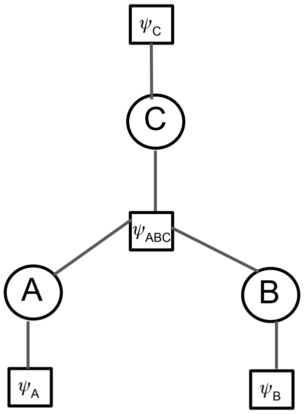


Figure 5

(a) (1 point) **Select One:** For a graphical model with a tree structure, belief propagation can be used to compute exact marginal probabilities.

○ True

○ False

○ Only true for binary trees

True

(b) (1 point) **Select One:** To compute the marginal distribution for variables D, E, and F respectively in the lattice factor graph from figure 4, which method would you prefer?

○ Variable Elimination

○ Belief Propagation

○ Neither method has an advantage over the other

Variable Elimination since graph is not acyclic

(c) (1 point) **Select One:** To compute the marginal distribution for variables A, B, and C respectively in the tree factor graph from figure 5, which method would you prefer?

○ Variable Elimination

○ Belief Propagation

○ Neither method has an advantage over the other

Prefer BP: BP will just be more efficient since it will simultaneously compute marginals for all the nodes.

5 Learning for BayesNets, MRFs, and CRFs

1. Suppose we are learning a fully observed Bayesian Network over 10 binary variables, where the conditional probability tables are parameterized by the minimum required number of parameters.

   (a) (1 point) **True or False:** If the maximal number of edges is added to this model, then the total number of parameters in this model is $2^{10} - 1$.

   ○ True

   ○ False

   True. Suppose we have a fully left connected BayesNet with variables arranged $x_1, \ldots, x_{10}$. Then we have CPTs, $p(x_{10}|x_9, \ldots, x_1)p(x_9|x_8, \ldots, x_1) \ldots p(x_1)$. The first of these would have 2^9 parameters, the second 2^8, and so on down to 2^0 for the last. This adds up to 2^{10-1}.

   (b) (1 point) **True or False:** Given $N = 200$ training examples, we can write down the maximum likelihood estimates of the parameters of the BayesNet in closed form.

   ○ True

   ○ False

   True

2. (2 points) **Short Answer:** For a CRF model described by the equation

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x}, \lambda, \mu)} \exp(\sum_{i=1}^{n} (\lambda \boldsymbol{f}(y_i, y_{i-1}, \boldsymbol{x}) + \mu \boldsymbol{g}(y_i, \boldsymbol{x})))$$

with two types of functions $\boldsymbol{f}$ and $\boldsymbol{g}$ and the corresponding weights λ and μ, write down the equation for the partition function Z.

$$Z(x, \lambda, \mu) = \sum_{y} \exp(\sum_{i=1}^{n} (\lambda \boldsymbol{f}(y_i, y_{i-1}, \boldsymbol{x}) + \mu \boldsymbol{g}(y_i, \boldsymbol{x})))$$

3. (2 points) **Short Answer:** In an arbitrary graph structured CRF, why is learning generally considered intractable?

Computing partition function Z is intractable as it requires summing over an a set $\mathcal{Y}(\mathbf{x})$ that is exponentially large in the size of $\mathbf{x}$.

4. (1 point) **True or False:** An MRF is a discriminative model whereas a CRF is a generative model.

   ○ True

   ○ False

   False