Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
FINAL EXAM Winter 2019 (12/9/2019)

60 points, 2 hours (2 pts/min); closed book. *Answer question on exam, and put **your name** on them.* L^2 = LR. You can use Γ(a, b...) = –a log a – b log b –…Don't use red ink. You can attach pages of your work, but it isn't obligatory.

---

1. [10 points]

(a) Let the level of knowledge of a student about information theory be a dichotomous variable, K, whose value is high (h) or low (l). Assume that an exam testing students on their knowledge makes only a binary discrimination, i.e., the grade, G, is also h or l. The probabilities of possible situations are in the table below, where a+b+c+d = 1. Write an (equality or inequality) equation involving *only* the terms H(K), H(G), and/or T(K:G) (use as many of these terms as necessary) which would hold if the exam resolved all uncertainty about K but was *more detailed* than needed to resolve this uncertainty. (2 pts)

		G: l	G: h
K	l	a	b
	h	c	d

(b) For the data given by the table below, with known probabilities, a…h, give an expression for T(A:Z) in terms of parameters, a…h, using the definition of T as a sum of p log (p/q) terms. Give only the *first two terms* of this sum. (2 pts)

	Z:	0		1	
	B:	0	1	0	1
A:	0	a	b	c	d
	1	e	f	g	h

(c) Let time be a variable, as follows. Consider ABC_1 and ABC_2 distributions below, where C_1 = t and C_2 = t+1, with probabilities, a+b+c+d = 1 and e+f+g+h = 1. The AB distribution at time t is the reference. I wish to test the hypothesis that AB at t+1 is the same as this reference distribution, i.e., the same as AB at time t. Using the p log (p/q) expression for transmission, write the T in terms of a…h that will let me test this hypothesis. (2 pts)

C_1

	B_1	B_2
A_1	a	b
A_2	c	d

C_2

	B_1	B_2
A_1	e	f
A_2	g	h

T =

(d) For some ABZ data, $T_B(A:Z) = 0$. Circle all models where T(model) is 0. (2 pts)

i. ABZ	ii. AB:AZ:BZ	iii. AB:AZ	iv. AB:BZ	v. AZ:BZ
vi. AB:Z	vii. AZ:B	viii. BZ:A	ix. A:B:Z	x. none of these

(e) Consider the contingency table, where a+b+c+d = 1

	y_1	y_2
x_1	a	b
x_2	c	d

Let x be a word in a message; let y be the word that follows it. Write an expression, in terms of parameters, a, b, c, & d, for the amount of information which receiving y provides beyond what one already knows from having previously received x. (2 pts)

Information(y | x) =

2. [12 points]

(a) Suppose A is a quantitative variable that needs to be binned. We want to maximize its power to predict Z but use degrees of freedom *efficiently*, since this might allow us to also use B to predict Z. One reasonable heuristic is to choose the number of bins for A which maximizes: (circle one; 2 pts)

i. H(Z|A)	iv. [H(Z) – H(Z|A)] * |A|
ii. H(Z) – H(Z|A)	v. H(Z|A) / |A|
iii. H(Z|A) * |A|	vi. [H(Z) – H(Z|A)] / |A|

(b) Suppose our DV is Z and our IVs are A, B, and C. The predictive effect of A on Z, *controlling for* B and C (how much knowing A tells us about Z beyond what B and C tell us about Z) is (circle one; 2 pts)

	i. T(ABCZ)	
	ii. T(ABC:ABZ:ACZ:BCZ	
iii. T(ABC:ABZ:ACZ)	iv. T(ABC:ABZ:BCZ)	v. T(ABC:ACZ:BCZ)
vi. T(ABC:AB:CZ)	vii. T(ABC:ACZ:BZ)	viii. T(ABC:BCZ:AZ)
ix. T(ABC:ABZ)	x. T(ABC:ACZ)	xi. T(ABC:BCZ)
	xii. T(ABC:AZ:BZ:CZ)	
xiii. T(ABC:AZ:BZ)	xiv. T(ABC:AZ:CZ)	xv. T(ABC:BZ:CZ)
xvi. T(ABC:AZ)	xvii. T(ABC:BZ)	xviii. T(ABC:CZ)
	xix. T(ABC:Z)	

(c1) To evaluate the model AB:AZ:BZ relative to the *independence model* as the reference, the appropriate information distance equals (circle all that are true; 2 pts)

i. T(AB:Z) – T(AB:AZ:BZ)
ii. T(AB:AZ:BZ)- T(AB:Z)
iii. T(ABZ) – T(AB:AZ:BZ)
iv. T(AB:AZ:BZ) – T(ABZ)
v. H(AB:Z) – H(AB:AZ:BZ)
vi. H(AB:AZ:BZ) – H(AB:Z)
vii. H(ABZ) – H(AB:AZ:BZ)
viii. H(AB:AZ:BZ) – H(ABZ)

(c2) To evaluate the model AB:AZ:BZ relative to the *data* as the reference, the appropriate information distance equals (circle all that are true; 2 pts)

i. T(AB:Z) – T(AB:AZ:BZ)
ii. T(AB:AZ:BZ)- T(AB:Z)
iii. T(ABZ) – T(AB:AZ:BZ)
iv. T(AB:AZ:BZ) – T(ABZ)
v. H(AB:Z) – H(AB:AZ:BZ)
vi. H(AB:AZ:BZ) – H(AB:Z)
vii. H(ABZ) – H(AB:AZ:BZ)
viii. H(AB:AZ:BZ) – H(ABZ)

(d1) True or false? (circle one; 2 pts): T(A:B:C) = T(A:B) + T(AB:C)

(d2) True or false? (circle one; 2 pts) T(ABC:Z) = T(A:Z) + T_A(B:Z) + T_{AB}(C:Z)

3. [12 points]
(a) Reference = top.

(a1) A Type I error will probably result in a model that is (circle one; 2 pts)

i. more complex than necessary
ii. too simple to fit the data

(a2) A Type II error will result in a model that is (circle one; 1 pt)

i. more complex than necessary
ii. too simple to fit the data

(b) Reference = bottom.

(b1) A Type I error will result in a model that is (circle one; 2 pts)

i. less complex than is statistically justified
ii. too complex to be statistically justified

(b2) A Type II error will probably result in a model that is (circle one; 1 pt)

i. less complex than is statistically justified
ii. too complex to be statistically justified

(c) A test is given to detect a disease. A 'negative' test result means that this condition is not detected, i.e., the patient is judged to be free of the disease; a 'positive' result means that the condition is detected, i.e., the patient is judged to have the disease. The actual condition of the patient and the test conclusions are summarized in these frequencies: TN = true negatives, FP = false positives, FN = false negatives, TP = true positives. The frequency marginals of the actual distribution are N = TN + FP, P = FN + TP. Assume that the null hypothesis is 'negative.'

	Test	(-)	(+)	
Actual	(-)	TN	FP	N
	(+)	FN	TP	P

Which of these is true (circle one; 2 pts)?

i. FP and FN are both Type I errors
ii. FP and FN are both Type II errors
iii. FP are Type I errors; FN are Type II errors
iv. FP are Type II errors; FN are Type I errors
v. none of the above

(d) Suppose the reference is *independence*. For some m_j, one calculates $\Delta df(m_j \rightarrow m_{ind})$ and $\Delta L^2(m_j \rightarrow m_{ind})$. One also decides on some α_c. Say that going into the Chi-square table with Δdf and α_c one gets some L^2_c and that the model $\Delta L^2 > L^2_c$. One would then say that (circle; 2 pts)

i. $m_j = m_0$, so one must use the bottom, m_{ind}, as one's model
ii. $m_j \neq m_{ind}$, so one is happy with m_j (*or* one *tries* to go higher in the lattice)
iii. $m_j \neq m_{ind}$, so is *un*happy with m_j and *must* go lower in the lattice

(e) When the reference is the *data*, one calculates $\Delta df(m_0 \rightarrow m_j)$ and $L^2(m_j)$ and decides on some α_c. Say that going into the Chi-square table with Δdf and α_c one gets some L^2_c and suppose that the model $L^2(m_j) > L^2_c$. One would then usually (circle; 2 pts)

i. not reject the null, and consider m_j as possibly a good model
ii. not reject the null, and use the data, m_0, as one's model
iii. reject the null, and consider models higher in the lattice than m_j
iv. reject the null and consider models lower in the lattice than m_j

4. [10 points] Consider on the left an *observed* (data) probability table (p) for a *directed* system, with sample size N. None of parameters (a...h) is 0. Let the *calculated* table (q) for model AB:BZ be the table on the right.

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	a	b	e	f
A_2	c	d	g	h

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	q_1	q_2	q_3	q_4
A_2	q_5	q_6	q_7	q_8

(a) Solve for q_2 *algebraically* for model AB:BZ in terms of a…h. (2 pts)

q_2 =

(b) Use IPF to obtain q_3^{AB} (not q_2!) and then $q_3^{AB:BZ}$ in terms of parameters a…h. $q_3^{initial} = 1/8$.

(b1) After imposing AB, we get (1 pt; write only the factor that multiplies $q_3^{initial}$)

$q_3^{AB} = q_3^{initial} *$ _____

(b2) After imposing also BZ, we get (3 pts: write only the factor that multiplies q_3^{AB})

$q_3^{AB:BZ} = q_3^{AB} *$ _____

(c1) Given the Occam fit output for model m = AB:AZ:BZ immediately below,

A	B	p(AB)	p(Z=1|AB)	p(Z=2|AB)	q_m(Z=1|AB)	q_m(Z=2|AB)
1	1	a	b	c	d	e
1	2	f	g	h	i	j
2	1	k	l	m	n	o
2	2	r	s	t	u	v
		marginals	w	x	y	z

circle the parameters in the table that are necessary and sufficient to evaluate p_{margin} for (A=2, B=2); circle the <u>smallest</u> number of necessary and sufficient parameters (2 pts).

(c2) The uncertainty reduction that one gets for (A=2, B=2) is (circle one; 2 pts)

i. $H(Z) - H(Z|A_2B_2)$
ii. $H(Z|A_2B_2) - H(Z)$
iii. $H(Z) - [\, H(Z|A_2) + H(Z|B_2)\,]$
iv. $H(Z|A_2) + H(Z|B_2) - H(Z)$
v. none of the above

5. [6 points]
(a) For the following table, state-based models are shown in italics. The reference is the top. |A|=|B|=|Z|=2

#	Model	T	%I	df	L^2	p
1	ABZ	- - -	100%	7	--	1.000
2	*AB:Z:a0BZ*	*0.0002*	*100%*	*6*	*0.3*	*0.603*
3	*AB:Z:a0b1Z*	*0.0696*	*61%*	*5*	*120.3*	*0.000*
4	*AB:Z:a0b0Z*	*0.0876*	*51%*	*5*	*151.4*	*0.000*
5	AB:AZ:BZ	0.1478	17%	6	255.5	0.000
6	AB:BZ	0.1482	17%	5	256.2	0.000
7	*AB:Z:a1b1Z*	*0.1610*	*10%*	*5*	*278.4*	*0.000*
8	*AB:Z:a1b0Z*	*0.1720*	*3%*	*5*	*297.4*	*0.000*
9	AB:AZ	0.1777	0%	5	307.2	0.000
10	AB:Z	0.1780	0%	4	307.6	0.000

(a1) The model # for the best variable based model is _____ (2 pts).

(a2) The model # for the best state based model is _____ (2 pts)

(b) Assume that the AB projection of the ABZ data is nearly what I would expect if A and B were independent, and so I want a (non-standard-RA) model where the hypothesis of independence for A and B is included in the model. The model will have a relation written as $AB_{A:B}$ and I want the model $AB_{A:B}$:AZ:BZ, which means that q(AZ) = p(AZ) and q(BZ) = p(BZ), but q(AB) = p(A)*p(B), *instead of* q(AB) = p(AB).

df($AB_{A:B}$:AZ:BZ) = _____ (2 pts)

6. [10 points]

(a) (4 pts) The set-theoretic mapping below maps values of A, B, C onto values of Z.

A	B	C	Z
0	0	0	1
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	0

(a1) The ABCZ mapping is (circle one):

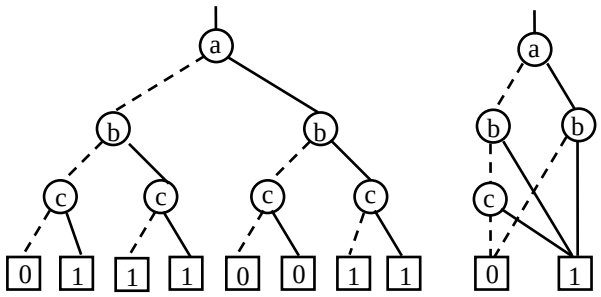
i. decomposable (without constraint loss)
ii. not decomposable (without constraint loss)

(a2) If you chose *decomposable,* the simplest structure equivalent to ABCZ is _____. Its *predicting* components are (circle one):

i. all deterministic ii. all stochastic iii. some deterministic, some not

If you chose *not decomposable,* the # of extra tuples of the 1st decomposition is ____

(b) *Introduction to binary decision diagrams*: The set-theoretic relation R = {001, 010, 011, 110, 111} can be represented by the tree on the *left side* of the figure below, where lines coming out of and down from a variable indicate its possible values, where a dashed line means 0 and a solid line means 1. In the square boxes at the bottom of the tree, 1 indicates that the tuple is in the relation; 0 indicates it is not. So reading down the tree on its left side, the three dashed lines culminating in a boxed 0 mean that 000 (i.e., a=0, b=0, c=0) is not in the relation. But 001 is. Etc. By information-preserving operations one can transform the tree on the left to the graph on the right. Reading the graph downwards in the same way, we see that 000 is not in the relation, but 001 is; also 01* is in the relation, as is 11*, but 10* is not. The right hand graph – called a binary decision diagram – specifies a 'compressed' relation, R_{BDD} = {001, 01*, 11*}.



(b1) Suppose I don't know R, but I am given R_{BDD}. To get the set of abc tuples implied by R_{BDD}, I replace 01* by {010, 011} and 11* by {110, 111}. These replacements expand R_{BDD}, and together with 001, make R_{BDD} = R. The *theoretical justification* for making these replacements is (circle one; 2 pts)

i. model is loopless	ii. model has loops
iii. law of subscripting	iv. law of distribution
v. minimum entropy	vi. maximum entropy

(b2) The compressed relation R_{BDD} is a set-theoretic analog of information-theoretic (circle one; 2 pts)

i. k-systems analysis	ii. Bayesian networks
iii. latent variable modeling	iv. state-based modeling

(c) I have two separate frequency distributions AB and BC which I *think* are samples of the same population, but I am not sure. I am interested in the relationship between A and C, so I want to merge these distributions by composition into $ABC_{AB:BC}$ and then take the AC projection of the result of this composition. What do I need to do first to establish whether such a merging is possible, i.e., statistically legitimate? (2 pts)