Exam Data Mining January 11, 2023, 17.00-19.30 hrs

Question 1 MIXED SHORT QUESTIONS

Part 1 (Bagging and Random Forests)

(a) FALSE (the order of examples doesn't matter)
(b) FALSE (sample per split, not per tree)
(c) TRUE
(d) TRUE

Part 2 (Frequent Pattern Mining)

(a) TRUE
(b) FALSE (3 times)
(c) FALSE (candidates are generated by rightmost extension)
(d) TRUE

Part 3 (Classification Trees)

If we predict class c, the probability of making a wrong prediction is 1-p(c|t).
We predict class c with probability p(c|t), so the probability of making an error is
\Sum_{c=1}^C p(c|t)(1-p(c|t)), which is the formula for the gini-index.

Part 4 (Graphical Models)

The correct answer is (b). Model (b) fits the independence model, and x_1 and x_2 are exactly independent in the data,
so the independence model gives a perfect fit of the observed counts. The saturated model (c) fits equally well, but uses
one more parameter, so has a worse BIC score. Model (a) fits the uniform table of counts 25,25,25,25.
It has two less parameters than the independence model, but a much worse fit.

Part 5 (Logistic Regression)

The correct answer is (d). The odds are multiplied by exp(0.14) which is approximately 1.15.
Multiplication by 1.15 is the same as an increase of 15%.

Question 2 CLASSIFICATION TREES

(a) i(t1) = 1/2 * 1/2 = 1/4. i(t2) = i(t3) = 1/5 * 4/5 = 4/25. Reduction is 1/4-(1/2*4/25+1/2*4/25)=9/100.

(b) The smallest minimizing subtree (SMS) for a1=0 is obtained by pruning in t2.
    Next we compute g(t1) = 21/100 and g(t3)=2/100. We prune in t3 and set a2=2/100.
    Next we recompute g(t1)=3/10, and set a2=3/10.
    Summarizing: T1 is obtained by pruning in t2, and it is the SMS for a in [0,2/100).
                 T2 is obtained by pruning T1 in node t3, and it is the SMS for a in [2/100,3/10).
                 The root node is the SMS for a >= 3/10.

(c) sqrt(2/100 * 3/10) = 0.077 (approximately)

Question 3 CLOSED FREQUENT ITEM SET MINING

(a)

LEVEL 1:			LEVEL 2:		
	sup	gen?		sup	gen?
A	2	v	AB	2	x
B	4	v	AC	2	x
C	5	v	AD	0	x
D	3	v	BC	4	x
E	1	x	BD	2	v
			CD	3	x

E and AD are pruned due to insufficient support, all other pruning because the itemset has a subset with the same support.

(b)

gen	closure	sup
A	ABC	2
B	BC	4
C	C	5
D	CD	3
BD	BCD	2

Question 4 BAYESIAN NETWORKS

(a) Yes, the resulting model has the same skeleton and v-structures.

(b) The current score of lipo is -1208. After deleting the edge mental --> lipo, lipo has one parent left, which is smoke.
    The score of lipo then becomes:
    598 log 598/961 + 363 log 363/961 + 463 log 463/880 + 417 log 417/880 = -1245.855, which we will round to -1246.

    The change in log-likelihood score is -1246 + 1208 = -38, so a decrease of 38.

(c) The penalty per parameter is log(1841)/2 = 3.76. Deleting the edge mental --> lipo reduces the number of parameters by 2,
    so the change in BIC score is -38 + 2 * 3.76 = -30.48, or -30 after rounding. The BIC score decreases by 30.

Question 5 MULTINOMIAL NAIVE BAYES

(a) |V|=10

    P(good|Pos) = (2+1)/(8+10) = 1/6
    P(good|Neg) = (0+1)/(8+10) = 1/18
    P(teacher|Pos) = (1+1)/(8+10) = 1/9
    P(teacher|Neg) = (2+1)/(8+10) = 1/6

(b) The class priors are P(Pos) = P(Neg) = 1/2
    We ignore "very" in the test document, because it didn't occur in the training set

    P(Pos)P(good teacher|Pos) = 1/2 * 1/6 * 1/9 = 1/108
    P(Neg)P(good teacher|Neg) = 1/2 * 1/18 * 1/6 = 1/216

    P(Pos|good teacher) = (1/108) / (1/108 + 1/216) = 2/3

(c) One training document of class A containing the word "a", and one training document of class B containing the word "b"
    Test document containing the words "a" and "b".