CSC591: Foundations of Data Science
HW5: Resampling, Bayesian Inference, Missing Data Analysis

Released: 11/26/18
Due: **12/02/18 (23:55pm);** (One day late: -25%; -100% after that).

Student Name:
Student ID:

Notes

- Submit single zip file containing: (1) all solutions as single pdf file (Filename: Lastname_StudentID.pdf);
- You can also submit scanned hand written solution (should be legible, TA's interpretation is final).
- This h/w is worth 4% of total grade
- You can discuss with your friends, but solution should be yours.
- Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
- All submissions must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
- You should attempt all questions, but all questions may not be graded.
- No makeups; for regarding policies, refer to syllabus and 1st day lecture slides.

Q#	Max Points	Your Score
1	15	
2	15	
3	20	
4	25	
5	10	
6	25	

Note: For each question (and subparts) please also list how much time it took to solve it.

Q1. Resampling, Cross-Validation (3 x 5 = 15 points)

(**a**) Discuss (at least one; more is better) advantages and disadvantage(s) of validation set, leave-one-out, and k-Fold cross-validation methods.

(**b**) Define bias-variance tradeoff and what is its implication for classification.

(**c**) Define true positive, false negative, false positive, true negative, overall accuracy, individual class accuracy, precision, recall, and f-measure using error matrix.

Q2. Sampling and Cross-Validation (15 points)

(a) (5 points) A non-data scientist, who is not fully familiar with the cross-validation techniques, designed the following simple classifier for two-class classification problem.

1. Starting with a 2000 attribute data and 100 samples with class labels, selected a subset of 10 attributes having largest correlation with the class labels.
2. Built a classifier (logistic regression) using only those 10 attributes selected in the 1st step.

In order to estimate the test performance, the non-data scientist applied cross-validation technique at step-2 and concluded that his procedure achieved minimum required accuracy.

Your objective is to **identify the problem** with the above procedure and **suggest a correct solution** (be clear and concise).

(b) (5 points). LOOCV technique.

Perform LOOCV on the given data, using the following classifier: for a given input x, predict 1 if x_i is greater than the mean(x) from the training data, and 0 if it is less than or equal to the mean. Report the CV accuracy.
Note: $MSE_i = 1$ if the predicted and test labels are same, else 0.

x	y
1	0
2	0
3	0
4	1
5	1

(c) (5 points) Accuracy assessment.
For the given table, compute the (i) construct contingency table; (ii) compute overall accuracy, (ii) precision, (iv) recall, and (v) F-measure.

Classification Prediction 1	Gound-truth
1	1
2	2
1	1
1	1
1	1
2	1
2	1
2	2
2	1
2	2
2	1
2	2
2	1
1	1
1	2

Q3. Bayesian Inferencing (20 points)

One of the standard measures (say "acceleration") reported for cars is the time (in seconds) required to reach 0-60 mph. A company determines that the acceleration for their new car is a normal r.v. with a s.d of 0.8 sec. Assume a normal prior distribution, $N(8, 0.2)$. If 10 of the production cars are tested and determined that the average acceleration is 9 sec, then find the 95% Bayesian interval for μ.

Q4. **(This question is optional)** The following table summarizes two exam scores. Left half of the table gives complete scores and right half gives an example of missing data. (25 points)

Complete Data		Missing Data	
mt1	mt2	mt1	mt2
74	66	74	66
70	58	70	58
66	74	66	74

55	47	55	47
52	61	52	61
47	38	47	38
45	32	45	
38	46	38	
33	41	33	41
28	44	28	

Answer the following (using data given in the above table):

(1) Based on the missing data, determine missing data pattern and justify your answer (5 points)
(2) Compute Mean and Standard Error (SE) for (i) complete data, and (ii) missing data using list-wise deletion. (5 points)
(3) Comment on bias of the estimates of (2.ii) as compared to estimates from complete data (2.i). (5 points)
(4) Impute missing data using simple regression (see lecture slides). (5 points)
(5) Compute Mean and SE on imputed data, and comment on bias of the estimates. (5 points)

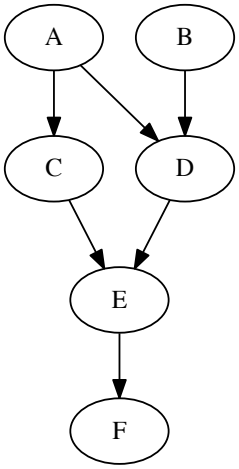
Q5. **(This question is optional)** (**10 points**)

(a) Describe missing data patterns and missing data mechanisms (**5 points**)
(b) Describe various traditional methods for dealing with missing data, highlight advantage and disadvantages of each method (**5 points**)

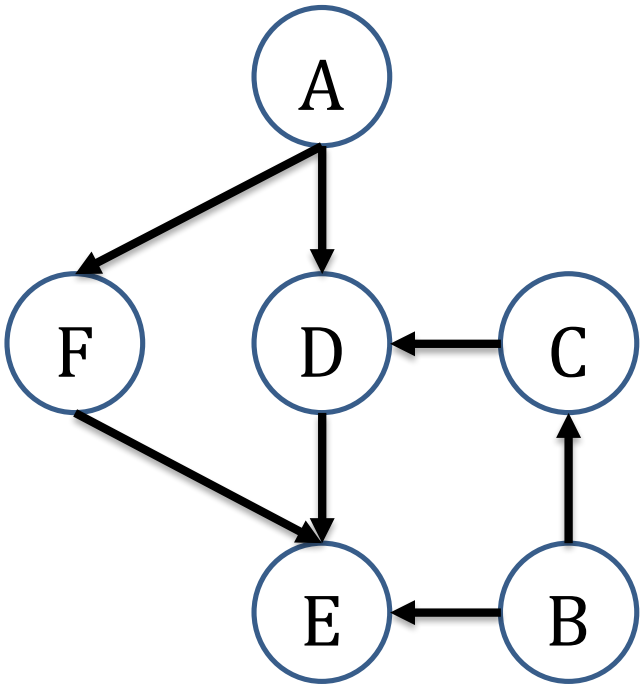
Q6. Bayes Networks (25 points)

(a) [10 points] Given the following Bayes net (all variables are binary) and probabilities; compute P(A=0, B=1, C=0, D=1, E=0, F=1)

P(A = 1) = 0.6
P(B = 1) = 0.4
P(B = 0) = 0.7
P(C = 1 | A = 0) = 0.2
P(C = 1 | A = 1) = 0.7
P(D = 1 | A = 0, B = 0) = 0.3
P(D = 1 | A = 0, B = 1) = 0.7
P(D = 1 | A = 1, B = 0) = 0.6
P(D = 1 | A = 1, B = 1) = 0.3
P(E = 1 | C = 0, D = 0) = 0.8
P(E = 1 | C = 0, D = 1) = 0.6
P(E = 1 | C = 1, D = 0) = 0.2
P(E = 1 | C = 1, D = 1) = 0.5
P(F = 1 | E = 0) = 0.9
P(F = 1 | E = 1) = 0.6



(**b**) [15 points] Given the following network



(i) [5 points] Compute P(A,B,C,D,E,F), and (**ii**) [10 points] using the result from (i) show that B and D are conditionally independent given A and C.