CSC591: Foundations of Data Science
HW4: Feature Selection, Resampling and Bootstrap, Bayesian Inference

Released: 11/13/15
Due: **11/23/15 (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); (2) separate R code file for Q4-Q5 with appropriately named readme files.
- You can also submit scanned hand written solution (should be legible, TA's interpretation is final).
- This h/w is worth 5% of total grade + **Bonus questions account for 2% of 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)
- No makeups; for regarding policies, refer to syllabus and 1st day lecture slides.

Q#	Max Points	Your Score
1	10	
2	20	
3	30	
4	15	
R5	25 (R mini project)	
B6	2% of grade (Bonus)	

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

Q1. Information theory (10 points)

Using the data provided in the following table, rank the attributes based on information gain.

a1	a2	a3	Class
T	T	5	Y
T	T	7	Y
T	F	8	N
F	F	3	Y
F	T	7	N
F	T	4	N
F	F	5	N
T	F	6	Y
F	T	1	N

Q2. Dimensionality Reduction (20 points)

(a) Find principal components for the following two-dimensional data: (20 points)

x1 = 1, 0, -1

x2 = -1, 1, 0

Q3. Nonparametric Tests (30 points)

(a) (10 points) A CNN meteorologist states that the median daily high temperature for the month of January in Raleigh is 67° Fahrenheit. The high temperatures (in degrees Fahrenheit) for 18 randomly selected January days in Raleigh are listed below. At $\alpha = 0.01$, can you reject the meteorologist's claim?

77 74 72 72 70 75 70 72 78 74 73 72 74 78 79 75 73 80

(b) (10 points) A drug company postulates that their vaccine will decrease the number of colds. The following table shows the results (number of colds before and after vaccination) of an experiment involving 14 subjects. At $\alpha = 0.05$, verify if the company's claim is true.

Subject	Before vaccine	After vaccine
1	0	2
2	1	1
3	2	0
4	2	2
5	2	3
6	3	2
7	3	1
8	3	3
9	3	2
10	4	1
11	4	3
12	5	2
13	5	4
14	6	3

(c) (10 points) Readings from a study conducted to measure impact of a food supplement on blood pressure (before and after the supplement) is given in the following table. At $\alpha = 0.01$, can you reject the claim that there is no reduction in blood pressure after taking supplement?

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Before	120	109	108	112	111	117	135	124	115	118	130	129
After	105	115	99	115	117	108	122	120	106	126	128	116

Q4. The following table summarizes the results of logistic regression classification results on 3 datasets construed using LOOCV method. (**15 points**)

(a) (5 points) Compute the (i) construct contingency table, (ii) compute overall accuracy, (ii) precision, (iv) recall, and (v) F-measure for each classification prediction.

(b) (10 points) Construct a fourth classifier by taking majority vote on 3 classifiers, and then compute all the measures asked in (a) for the fourth classifier.

Classification Prediction 1	Classification Prediction 2	Classification Prediction 3	Gound-truth
1	1	2	1
1	2	1	1
2	1	1	1
1	1	2	1
1	1	1	1
2	2	2	2
2	2	1	1
2	1	2	2
2	2	1	2
2	2	1	2
2	2	1	1
2	1	2	2
2	2	1	1
1	1	2	1
1	2	2	2

Mini R project (25 points)

Note: You should provide answers (e.g., plots or explanation) in the homework solution pdf file, but submit R scripts as separate text files (zip them together).

R5. PCA and Validation (25 points)

(**a**) (10 points) Implement Q5 from the book (under 5.4 exercises)

(**b**) (15 points) Implement PCA based Eigenface generation (see class slides). Use the data given in the faces-corrected.zip)

(Briefly describe your steps and include top 10 eigenface images; submit code separately).

Bonus Questions (2% of grade)

B6. (**a**) (Chapter 5) Answer Q2 from the book under 5.4 exercises (1%)

(**b**) (Chapter 6) Answer Q8 (only parts (a) to (c) from the book under section 6.8 exercises (1%).