ISyE 619 HW 3 Solution

Spring 2014

Due: 4/30 on class

Problem 1 (4 points)

The training data:

index:	0	1	2	3	4	5	6	7	8	9
x value	0	1	2	3	4	5	6	7	8	9
y value:	1	1	1	-1	-1	-1	1	1	1	-1

Assume the weak classifier has the form: $x < v$, or $x > v$. The threshold v is determined to minimize the probability of error over the entire data. Please use adaboost method with $M = 3$ to find out the final prediction and the classification error

Problem 2 (3 points):

Data Format of "PCA service.xls"

The columns:

1) Name of the company
2) Revenue (in millions)
3) Profits (in millions)
4) Assets (in millions)
5) Market value (in millions)
6) Profits as % of revenue
7) Profits as % of assets
8) Earnings per share (negative number implies loss)
9) 1987-97 annual growth of earnings per share (in %)
10) Total return to investors in 1997 (in %)
11) 1987-97 average annual rate of total return to investors (in %)

The rows: object of a company

Wal-Mart	119299	3526	45525	113730.8	3	7.8	1.56	18.9	74.8	20.5
AT&T	53261	4638	58635	105878.7	8.7	7.9	2.84	4.2	46.2	16.2
Sears	41296	1188	38700	22573.8	2.9	3.1	2.99	-3.7	0.1	17.3
Traveler	37609	3104	386555	69419.7	8.2	0.8	2.54	41.3	79.8	32.7
Kmart	32183	249	13558	8204.9	0.8	1.8	0.51	-11.3	10.8	1

Analyze these data sets using PCA and discuss your conclusions in following steps:

(a) Conduct the PCA based on covariance matrix and summarize the result

(b) Plot the percentage of variance explained by each PC, and the PC scores of the subjects (each company) in the first 2 PCs space.

(c) How many PCs will be retained? What conclusions can you draw from the results? Is this analysis reasonable? If yes, please explain the reason. If not, please do further analysis to draw your conclusion. (Hint: you may need to consider the PCA based on correlation matrix (i.e., first standardize the dataset))

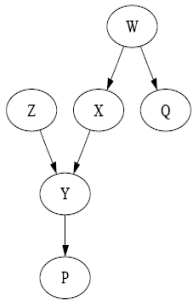
Hint:

In R you may refer to the function "princomp". You can take a look at some examples online before starting this question.

If you use matlab, following functions may be helpful:
[pc, zscores, pcvars] = princomp(data);
[pc, zscores, pcvars] = princomp(zscore (data));
zscore()
figure,scatter(),bar(),plot(), xlabel(), ylabel(), title()

Problem 3 (3 points)

For the BN below, please decide if the following statements are true or false. (Please use D-separation to explain and support your statements)



(1) Q is independent of {X, Y, Z, P} given W

(2) Z is independent of {X, W, Q}

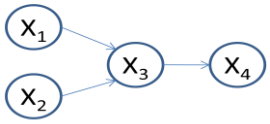
(3) Z is dependent of {X, W, Q} given P

(4) {Z, Y, P} is independent of {W, Q} given X

(5) {Z, Y, P} is dependent of Q

Problem 4 (3 points)

Given the true BN as below,



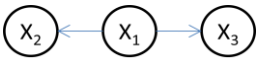
Please specify: by which statistical independence test the arc is removed or oriented at each step of PC algorithm. And also please draw the learned graph after each PC step. Assume infinite sample size, i.e., the statistical test will always identify the true independence or dependence relationship. (*Hint: you should start from a fully connected model. Then, you need to sequentially remove arcs. Finally, you need to orient the arcs.*)

Problem 5. (3 points)

Given the data on three variables, calculate the CH score for the following BN structure:

	Variable values for each case		
Case	x_1	x_2	x_3
1	*present*	*absent*	*absent*
2	*present*	*present*	*present*
3	*absent*	*absent*	*present*
4	*present*	*present*	*present*
5	*absent*	*absent*	*absent*
6	*absent*	*present*	*present*
7	*present*	*present*	*present*
8	*absent*	*absent*	*absent*
9	*present*	*present*	*present*
10	*absent*	*absent*	*absent*

BN structure



Problem 6. (4 points)

Simulate different sample sizes according to the "sprinkler" BN and use K2 to learn a BN for each sample size. Then, compare the learned BNs with the true model. Please tell me how much data it takes to recover the generating structure?