IS5152 Data-driven Decision Making
SEMESTER II 2018-2019
Assignment 1
Suggested solution

1. (10 points) The respondents in a survey were asked which of the following 8 criteria are true about a newly opened fast food restaurant and whether they would patronize the restaurant again in the near future:

   - A1. Clean and tidy
   - A2. Fresh ingredients are used
   - A3. Easy to find seats/tables
   - A4. Serves breakfast items I prefer
   - A5. Wide variety of items on menu
   - A6. Fast and efficient service
   - A7. Already had a good experience/impression
   - A8. A place I feel familiar with

   The following data were collected from 10 respondents:

Respondent	A1	A2	A3	A4	A5	A6	A7	A8	Decision
1	0	0	0	1	1	1	0	0	0
2	0	1	0	1	1	0	0	1	0
3	1	1	0	1	0	1	0	0	0
4	1	0	0	1	1	1	0	0	0
5	1	0	0	0	1	1	0	1	0
6	0	0	1	1	0	1	1	1	1
7	1	0	1	1	0	1	0	1	1
8	1	0	1	1	0	1	0	1	1
9	1	1	1	1	0	1	0	1	1
10	1	1	0	1	0	1	1	1	1

   Note:

   - Columns A1 to A8: 1 = true, 0 = not true
   - Column *Decision*: 1 = yes, will patronize again soon, 0 = otherwise

   (a) (6 points) A univariate binary decision tree is to be used to distinguish between the respondents who say they will patronize the restaurant again soon and those who do not.

      i. Compute the heterogeneity of the data using Gini index.
      $P_0 = 0.5, P_1 = 0.5, Gini = 1 - (0.5)^2 - (0.5)^2 = 0.5$

      ii. Suppose we split the data using the input A3, what is the heterogeneity after splitting?

- A3 = 0 : 5 samples Class 0, 1 sample class 1, Gini =$1-(5/6)^2-(1/6)^2 = 10/36 = 5/18$
- A3 = 1 : 4 samples Class 1, Gini = 0
- Gini index after split = $(6/10)(5/18)+(4/10)(0) = 1/6 = 0.167$

iii. With A3 as the first split in the tree, build a tree that correctly classifies all the training data samples.

```
           A3
         /   \
       =0     =1
       /        \
     A7       Class 1
    /  \
  =0    =1
  /       \
Class 0  Class 1
```

iv. (4 points) Are the data samples linearly separable? Explain your answer briefly.

Data samples are linearly separable:

If $\sum_{i=1}^{8} w_i A_i \geq c$, then Decision = 1,
Otherwise Decision = 0.

Let $w_1 = w_2 = ... w_8 = 1$ and $c = 4.5$
If $\sum_{i=1}^{8} A_i \geq 4.5$, then Decision = 1,
Otherwise Decision = 0.

Alternatively:
If $A_3 + A_7 \geq 0.5$, then Decision = 1,
Otherwise Decision = 0.

2. (10 points) Consider a hypothetical data set from a very small local financial institution shown below.

It contains information about the customers of this institution as described by the values of three attributes: Job, Marital status and Education. The possible values for these attributes are as follows:

- Job: Unemployed, Student, Blue-collar, Professional.
- Marital status: Single, Married, Divorced.
- Education: Primary, Secondary, Tertiary.

We are interested in building a classifier to predict if the customers have or have not taken up a personal loan with the institution.

Define Misclassification Index $= 1 - \max_h P_h$, where P_h is the proportion of data samples that belong to class h.

(a) (2 points) What is the misclassification index of this data set? Misclassification index = $1 - \max\{\frac{7}{15}, \frac{8}{15}\} = \frac{7}{15}$.

				taken
Customer no.	Job	Marital status	Education	Personal Loan?
1	Blue-collar	Married	Primary	NO
2	Blue-collar	Married	Primary	NO
3	Blue-collar	Single	Tertiary	YES
4	Professional	Divorced	Primary	NO
5	Professional	Divorced	Tertiary	YES
6	Professional	Married	Secondary	NO
7	Professional	Married	Tertiary	YES
8	Professional	Single	Tertiary	YES
9	Professional	Single	Tertiary	YES
10	Student	Married	Tertiary	NO
11	Student	Single	Secondary	NO
12	Student	Single	Tertiary	NO
13	Unemployed	Divorced	Secondary	YES
14	Unemployed	Married	Primary	NO
15	Unemployed	Married	Tertiary	YES

(b) (2 points) Suppose we are building a decision tree that allows multi-split of a non-leaf node. How much of the impurity (as measured by the misclassification index) can be reduced if the values of the attribute Education are used to split the data?

- Education = primary: 4 NO, 0 Yes, misc. index $= 0$
- Education = secondary: 2 NO, 1 Yes, misc. index $= 1 - \max\{\frac{1}{3}, \frac{2}{3}\} = \frac{1}{3}$.
- Education = tertiary: 2 NO, 6 Yes, misc. index $= 1 - \max\{\frac{2}{8}, \frac{6}{8}\} = \frac{1}{4}$.
- Index after split = $(4/15)0 + (3/15)(1/3) + (8/15)(1/4) = 0.2$.
- Reduction in index = $\frac{7}{15} - 0.2 = \frac{4}{15} = 0.267$.

(c) (6 points) The rules obtained by C4.5Rules are as follows:

- If Education = Primary, then Personal Loan = NO,
- else if Job = Student, then Personal Loan = NO,
- else if Education = Tertiary, then Personal Loan = YES,
- else Personal Loan = NO.

Compute:

- the accuracy
- the true positive
- the true negative

of the above rules.

Note: assign (Personal Loan = YES) as positive class, and (Personal Loan = NO) as negative class.

- If Education = Primary, then Personal Loan = NO (4 NO correcttly classified, 1,2,4,14)
- else if Job = Student, then Personal Loan = NO (3 NO correcttly classified, 10,11,12)

- else if Education = Tertiary, then Personal Loan = YES (6 YES correctly classified, 3,5,7,8,9,15)
- else Personal Loan = NO (1 NO correct no 6, 1 YES incorrect no 13).
- Answer:
  - the accuracy = 14/15 = 93.33%.
  - true positive = 6, true positive rate = 6/7 = 85.71%.
  - true negative = 8, true negative rate = 8/8 = 100%.

3. (10 points) The Good Furniture Company produces inexpensive tables, chairs, beds and cupboards. The production process for each require a certain number of carpentry work and a certain number of labor hours in varnishing department. These requirements are summarized below:

	Hours required to produce one unit				
Department	Table (X_1)	Chair (X_2)	Bed (X_3)	Cupboard (X_4)	Available hours per week
Carpentry	4	3	6	8	800
Varnishing	2	1	3	4	300
Profit per unit	\$60	\$50	\$80	\$160	

Note: Let X_1, X_2, X_3, X_4 be the number of tables, chairs, beds and cupboards to produce, respectively.

(a) (2 points) Formulate a linear programming problem to find the optimal number of tables, chairs, beds and cupboards that must be produced weekly to achieve maximum profit.

$$\max \; 60X_1 + 50X_2 + 80X_3 + 160X_4$$

subject to

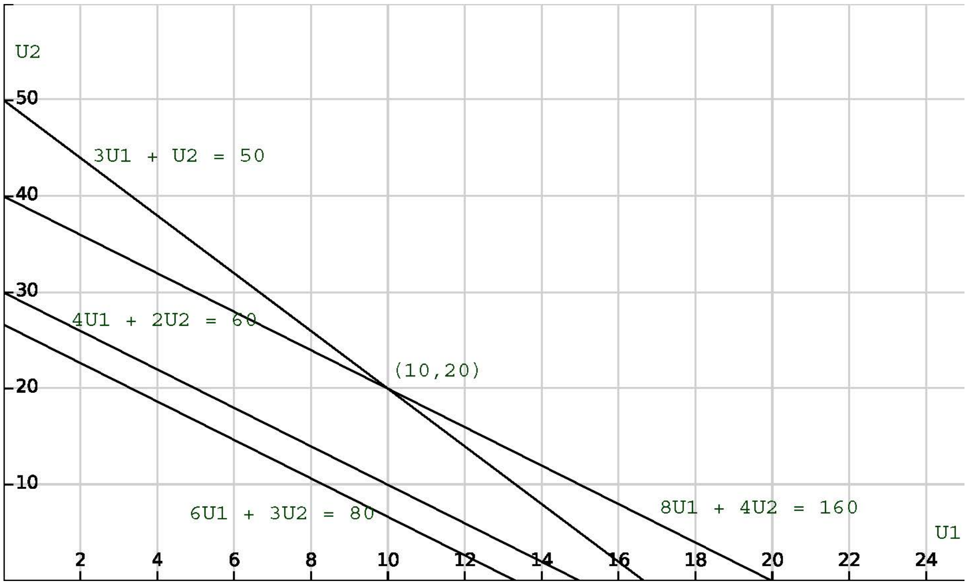
$$\begin{aligned} 4X_1 + 3X_2 + 6X_3 + 8X_4 &\leq 800 \\ 2X_1 + 1X_2 + 3X_3 + 4X_4 &\leq 300 \\ X_1, X_2, X_3, X_4 &\geq 0 \end{aligned}$$

(b) (3 points) State the dual of your linear program in part (a) above.

$$\min \; 800U_1 + 300U_2$$

subject to

$$\begin{aligned} 4\,U_1 + 2\,U_2 &\geq 60 \quad \text{C1} \\ 3\,U_1 + 1\,U_2 &\geq 50 \quad \text{C2} \\ 6\,U_1 + 3\,U_2 &\geq 80 \quad \text{C3} \\ 8\,U_1 + 4\,U_2 &\geq 160 \quad \text{C4} \\ U_1, U_2 &\geq 0 \end{aligned}$$



(c) (3 points) Find the optimal solution of the problem using the graphical approach. State clearly how many tables, chairs, beds and cupboards that must be produced and the maximum profit.

$U_1 = 10, U_2 = 20$. C1 is not binding, hence $X_1 = 0$. C3 is not binding, hence $X_3 = 0$. U_1 and U_2 are strictly positive, hence both primal constraints are binding: $3X_2 + 8X_4 = 800$ and $X_2 + 4X_4 = 300$. Solution $X_2 = 200, X_4 = 25$, Primal objective function value = $60X_1 + 50X_2 + 80X_3 + 160X_4$ = dual objective function value = $14000 = 800U_1 + 300U_2$

(d) (2 points) Suppose you can buy one additional hour of carpentry or one additional hour of varnishing for the same price, which one (carpentry or varnishing) would you get? Explain your decision.

Varnishing, since $U_2 > U_1$.

4. (10 points) The table below shows data samples from a credit scoring database.

customer number i	no. of cards x_1	monthly income ($000) x_2	Status	target value d_i
1	0	2	Bad	-1
2	2	2.4	Bad	-1
3	2	5	Good	$+1$
4	3	4	Good	$+1$
5	1	4.5	Good	$+1$
6	1	4	Good	$+1$
7	1	2	Bad	-1
8	0	1.2	Bad	-1

Status is the class label (target attribute), while *no. of credit/debit cards* and *monthly income* are the input attributes.

(a) (4 points) Find the hyperplane that separates Good and Bad credits with the maximum margin. Hint: Plot the data samples.

Support vectors for the two classes:

$$\begin{aligned} w_1x_1 + w_2x_2 + b &= 1 \text{ for } d_i = 1 \\ w_1x_1 + w_2x_2 + b &= -1 \text{ for } d_i = -1 \end{aligned}$$

If the data samples # 4,6 and 2 are support vectors:

$$\begin{aligned} 3w_1 + 4w_2 + b &= 1 \\ 1w_1 + 4w_2 + b &= 1 \\ 2w_1 + 2.4w_2 + b &= -1 \end{aligned}$$

The solution is $w_1 = 0, w_2 = \frac{5}{4}, b = -4$

(b) (6 points) Show that your hyperplane in part (a) above is optimal by checking all the KT conditions.

The multipliers for the 3 support vectors $\alpha_2, \alpha_4, \alpha_6$ must satisfy the following conditions:

$$\begin{aligned} -\alpha_2 + \alpha_4 + \alpha_6 &= 0 \\ -2\alpha_2 + 3\alpha_4 + \alpha_6 &= 0 \\ -2.4\alpha_2 + 4\alpha_4 + 4\alpha_6 &= \frac{5}{4} \end{aligned}$$

KT conditions: Solution: $\alpha_2 = \frac{50}{64}, \alpha_4 = \alpha_6 = \frac{25}{64}$.

i. Feasibility: all samples are on the correct side of the decision hyperplane.

ii. Complementarity conditions: set $\alpha_1 = \alpha_3 = \alpha_5 = \alpha_7 = \alpha_8 = 0$, then all conditions $\alpha_i[d_i(w_1x_1 + w_2x_2 + b) - 1], i = 1, 2, \ldots 8$ are satisfied.

iii. $\mathbf{w_i} = \sum_{i=1}^{8} d_i\alpha_i\mathbf{x_i}$ and $\sum_{i=1}^{8} \alpha_i d_i = 0$.

iv. $\alpha_i \geq 0, i = 1, 2, \ldots 8$.