FREIE UNIVERSITÄT BOZEN
LIBERA UNIVERSITÀ DI BOLZANO
FREE UNIVERSITY OF BOZEN · BOLZANO

Fakultät für Informatik | Facoltà di Scienze e tecnologie informatiche | Faculty of Computer Science

Advanced Algorithms

Written Examination

17th February 2014

FIRST NAME		**LAST NAME**	
STUDENT NUMBER		**SIGNATURE**	

Instructions for students

Write First Name, Last Name, Student Number and Signature where indicated. If not, the examination cannot be marked.

Use a pen, not a pencil.

Write neatly and clearly.

Student Code Ethics

Students are expected to maintain the highest standards of academic integrity. Work that is not the students' own creation will receive no credit. Remember that you cannot give or receive unauthorized aid on any assignment, quiz, or exam. Students cannot use others' ideas and declare that they belong to them. Students are required to properly cite the original sources of the ideas and information used in their work.

Exercise 1 (4 points)
Explain the typical scenario of the RSA cryptosystem, in which Alice and Bob want to communicate in private and Eve attempts to decrypt the messages they exchange. In particular, explain why the encryption and decryption of messages is easy for Alice and Bob, while decryption is extremely difficult for Eve.

Solution
Suppose that Alice wants to send a message m to Bob. She encrypts m using Bob's public key, obtaining m_{enc} and then sends m_{enc} over the communication channel. Bob can then decrypt m_{enc} using his private key, obtaining back m.
Bob's public key is (N, e) and can be calculated in polynomial time by selecting two prime numbers p and q, setting $N=p*q$, and selecting e as a number relatively prime to $(p-1)*(q-1)$. Bob's private key d is the inverse of e mod $(p-1)*(q-1)$, also calculated in polynomial time using the extended Euclid algorithm. The encryption $m_{enc}=m^e \ mod \ N$ is calculated in polynomial time. The decryption $m = m_{enc}^d \ mod \ N$ is also calculated in polynomial time.
In order to decrypt m_{enc} without knowing d, Eve could experiment with all possible values of m, each time checking whether $m^e \equiv m_{enc} \ mod \ N$, but this would take exponential time. Or she could try to factor N to retrieve p and q, and then figure out d by calculating the inverse of e mod $(p-1)*(q-1)$, but factoring would also take exponential time.

Exercise 2 (2 points)
Make an example of set cover problem. Outline a greedy algorithm to solve it. Is the greedy algorithm assured to find an optimal solution?

Solution
An example of set cover problem is the following:

> "A county is in its early stages of planning and is deciding where to put schools. There are only two constraints: (1) each school should be in a town; and (2) no one should have to travel more than 30 miles to reach the closest school. What is the minimum number of schools needed?"

This example is an instance of the generic set cover problem, that is

SET COVER

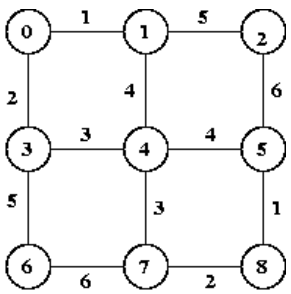
Input:	A set of elements B, sets $S_1, \ldots, S_m \subseteq B$
Output:	A selection of the S_i whose union is B.
Cost:	Number of sets picked.

In our example $B=\{t_1, \ldots, t_n\}$ is the set of towns and S_i is the subset of B of towns within 30 miles of t_i.
A greedy algorithm to solve the example set cover problem is:

Repeat
  Pick the S_i with maximum cardinality
Until all elements of B are covered

This greedy algorithm is not assured to find the solution with minimum cost, but it can be proved that the algorithm will use at most $k*ln \ n$ sets, where n is the cardinality of B and k is the cost of the best solution.

Exercise 3 (3 points)
For the graph below, show the steps in executing Kruskal's algorithm. Draw the graph with the MST edges identified at each step.



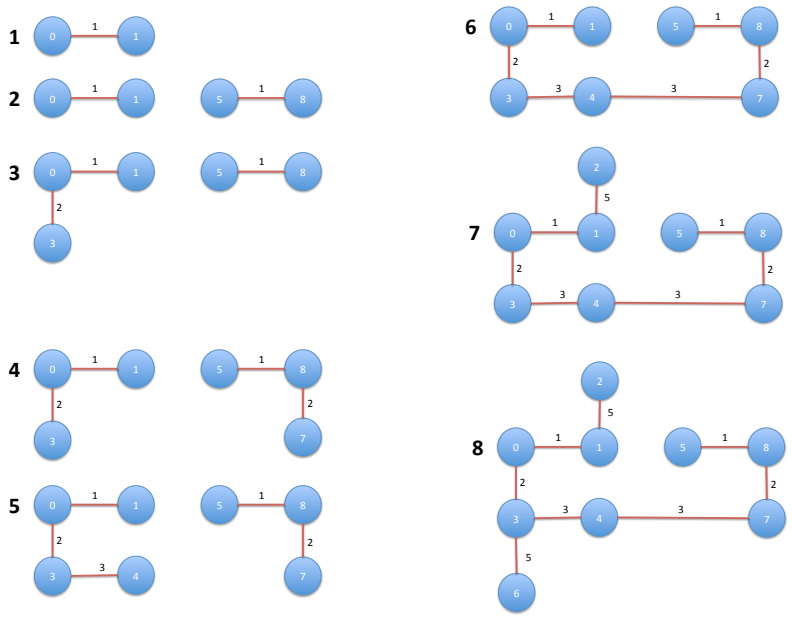
Solution

Let *G=(V,E)* be the graph above. We apply the Kruskal's algorithm in order to find a MST on *G*.

1. Let first order the edges in *E* by increasing weight

Weight		2	3	4	5	6
Edges	(0,1) (5,8)	(0,3) (7,8)	(3,4) (4,7)	(1,4) (4,5)	(1,2) (3,6)	(2,5) (6,7)

2. Now let's start from an empty tree and repeatedly add to the partial solution the next lightest edge from *E* that doesn't produce a cycle. The creation of the MST is outlined in the figure below, where the partial solution at each iteration of the algorithm is shown. The cost of the found MST is 22.



Exercise 4 (3 points)

Under a Huffman encoding of *n* symbols with frequencies $f_1, f_2, \ldots, f_n$, what could the maximum length of a codeword possibly be? Give a sample set of frequencies that would produce this case.

Solution

The longest codeword can be of length *n-1*. An encoding of *n* symbols with *n-2* of them having probabilities $1/2, 1/4, \ldots, 1/2^{n-2}$ and two of them having probability $1/2^{n-1}$ achieves this value.

Exercise 5 (3 points)

O'Hagan Bookworm Booksellers buys books from two publishers. Duffin House offers a package of 5 mysteries and 5 romance novels for $50, and Gorman Press offers a package of 5 mysteries and 10 romance novels for $150. O'Hagan wants to buy at least 2,500 mysteries and 3,500 romance novels, and he has promised Gorman (who has influence on the Senate Textbook Committee) that at least 25% of the total number of packages he purchases will come from Gorman Press. How many packages should O'Hagan order from each publisher in order to minimize his cost and satisfy Gorman? How much will the books cost him?

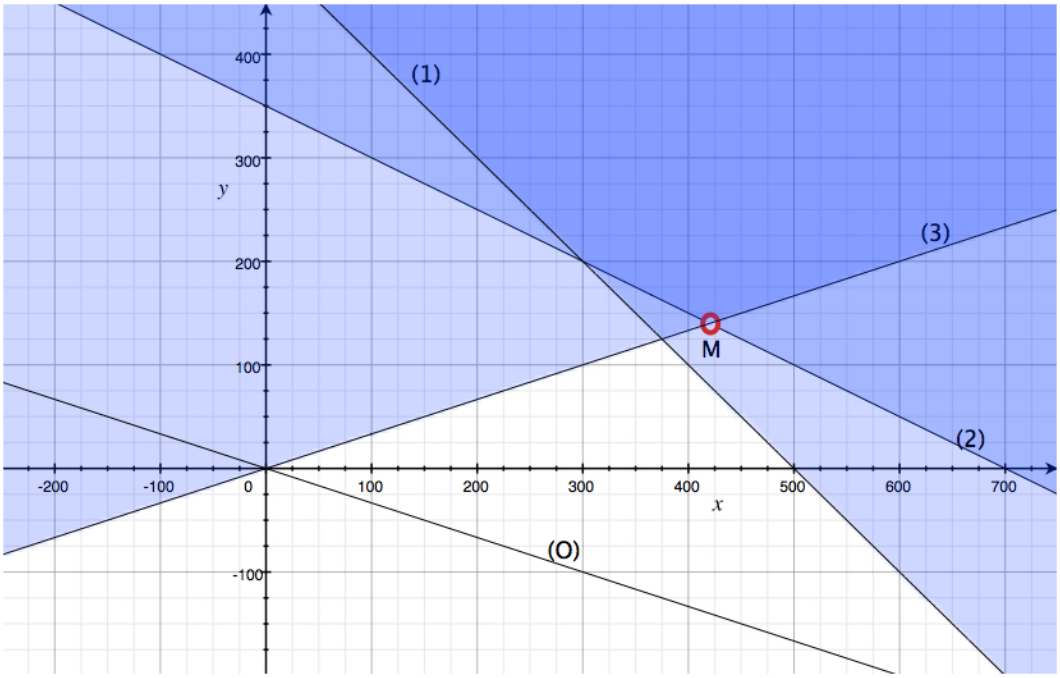
1. Formulate a linear programming model for this problem.
2. Use the graphical method to solve this model.

Solution

A linear programming model for this model is the following, where *x* is the number of packages bought from Duffin House and *y* is the number of packages bought from Gorman Press.

Minimize 50x+150y (O)
5x+5y >= 2500 (1)
5x+10y >= 3500 (2)
(x+y)/4 <= y (3)
x >= 0, y>=0

The graphical representation of this model is below. The darkest blue area is the feasible region (trivial constraints are not shown). The line corresponding to all the pairs *(x,y)* for which the objective is equal to 0 is denoted by *O*. Given the slope of this line, we can visually determine that the minimum of the objective function is the vertex of the feasible region indicated by *M*. The coordinates of *M* can be found calculating the intersection of lines (2) and (3), *M=(420, 140)*. The value of the objective function in M is 42000.



Exercise 6 (3 points)
Consider predicting a person's allergy intensity given the level of pollen (the substance that produces the allergy) in the air, and the amount of allergy treatment the person has taken. Assume that allergy intensity, level of pollen, and amount of allergy treatment have continuous values.

1. Describe this problem as a machine learning problem, indicating:
   - The learning task *T*;
   - The performance measure *P*;
   - The training experience *E*.
2. This problem can be solved using linear regression
   - Describe the hypothesis *h*;
   - Suppose to run the gradient descent algorithm in order to find the optimal values of the parameters of *h*. What do you think would reasonable values be for the parameters found by gradient descent?

Solution

- Learning task *T*: predicting the allergy intensity giving the level of pollen and the amount of treatment taken
- Performance *P*: the error in the prediction
- Experience *E*: learning set of triples <pollen, treatment, allergy> each of which indicates the experienced level of allergy corresponding to specific level of pollen and amount of treatment.

The hypothesis *h* has the form $h=\theta_0+\theta_1 p+\theta_2 t$ where *p* and *t* denote the level of pollen and the amount of treatment, respectively. The hypothesis *h* is used to predict the allergy intensity as a linear combination of *p* and *t*. Running the gradient descent we find the values for θ_0, θ_1, and θ_2 that minimize the error of the prediction on the training examples in the learning set *E*.

A reasonable value for θ_0 could be *0*, as a person does not generally have a baseline level of allergy. The parameter θ_1 would reasonable be greater than 0, as an increment of the pollen level in the air would produce an increase of the level of allergy. The parameter θ_2 would reasonable be less than 0, as an increment of the taken treatment amount would reduce the level of allergy.