CSE445 Assignment #1

There are 5 different questions given below. The total marks are 15 (each question has 3 marks).

Q 1. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in prediction or not. Give explanation for your answers.

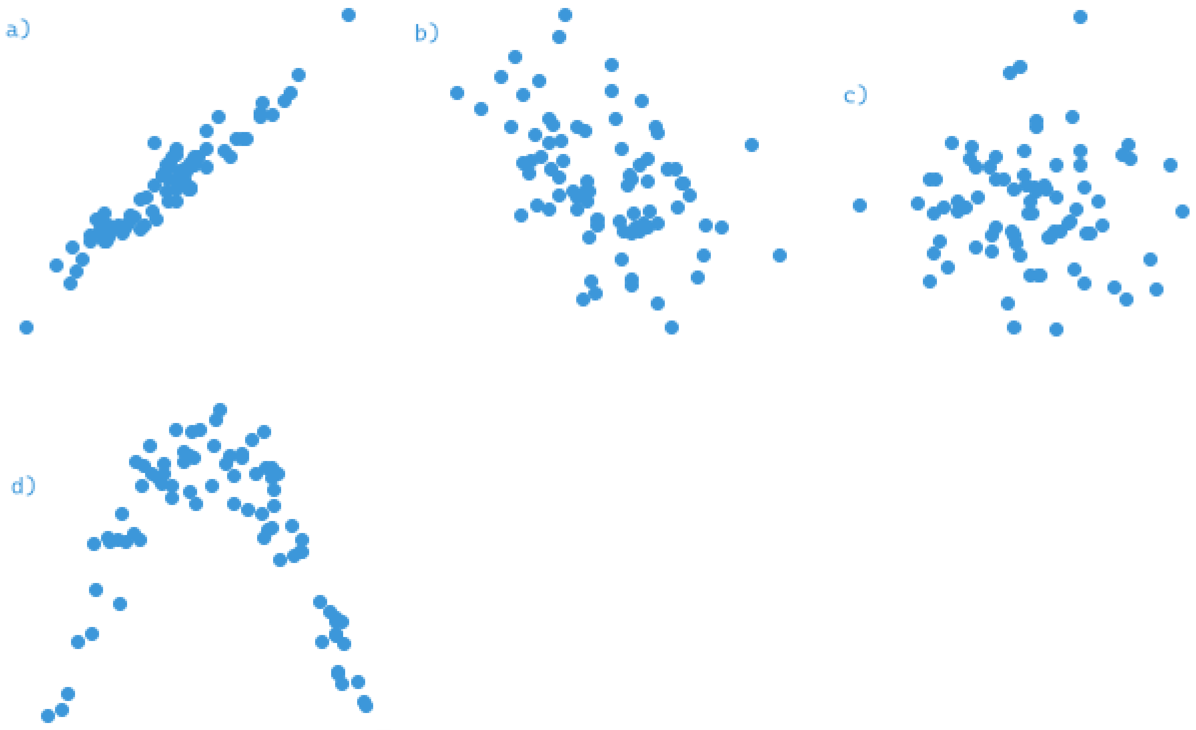
(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Q 2. When do we use regression analysis for machine learning? When is it not appropriate to use? Give some examples of where it is used other than the ones discussed in class.

Q 3. How would you fit your regression model (how would it look like) to the scatter plots of different data shown in the graphs below? How accurate would these models be for prediction? Give reasons for your answers.



Q 4. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

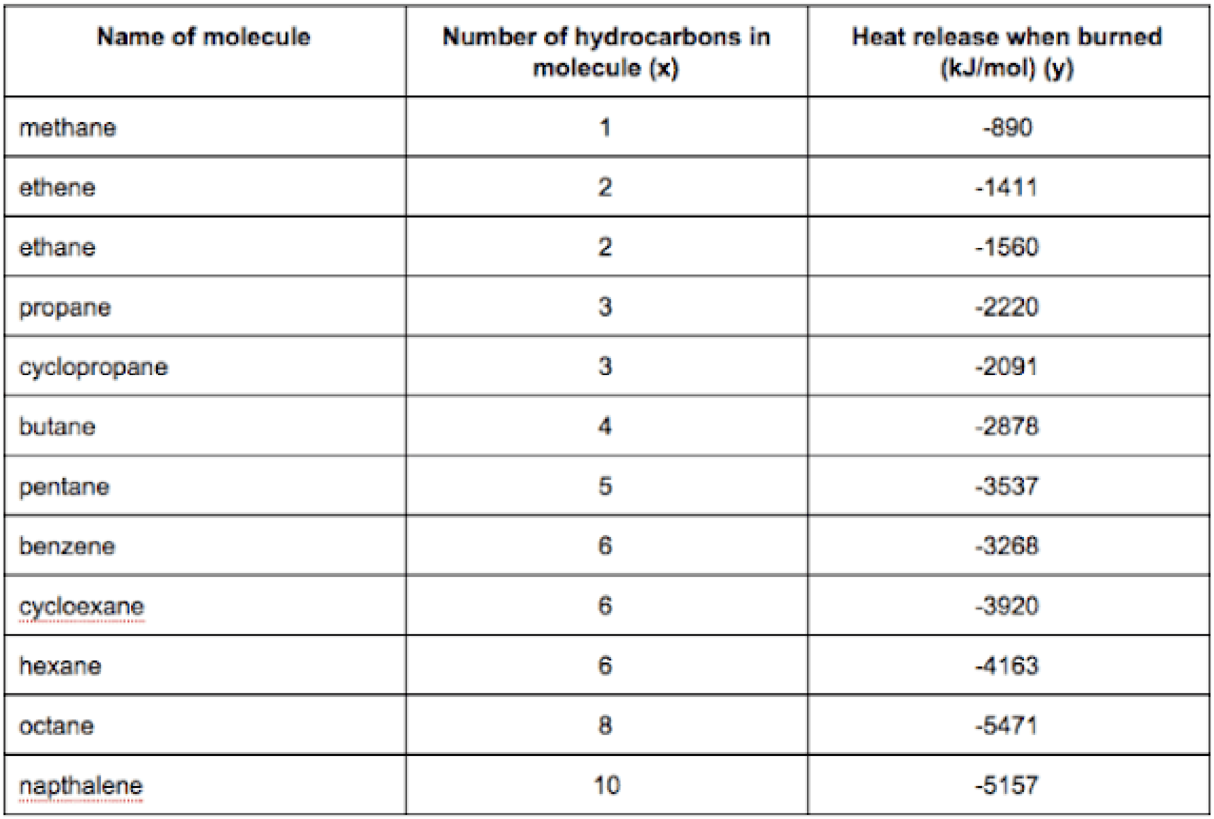
(a) Given historical data of children's ages and heights, predict children's height as a function of their age.

(b) Given 50 articles written by male authors, and 50 articles written by female authors, learn to predict the gender of a new manuscript's author (when the identity of this author is unknown).

(c) Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".

(d) Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.

Q 5. Many substances that can burn (such as gasoline and alcohol) have a chemical structure based on carbon atoms; for this reason, they are called hydrocarbons. A chemist wants to understand how the number of carbon atoms in a molecule affects how much energy is released when that molecule combusts (meaning that it is burned). The chemist obtains the dataset below. In the column on the right, "kJ/mol" is the unit measuring the amount of energy released.



Name of molecule	Number of hydrocarbons in molecule (x)	Heat release when burned (kJ/mol) (y)
methane	1	-890
ethene	2	-1411
ethane	2	-1560
propane	3	-2220
cyclopropane	3	-2091
butane	4	-2878
pentane	5	-3537
benzene	6	-3268
cycloexane	6	-3920
hexane	6	-4163
octane	8	-5471
napthalene	10	-5157

You would like to use linear regression ($y = \beta_0 + \beta_1 X$) to estimate the amount of energy released (y) as a function of the number of carbon atoms (x). Which of the following do you think will be the values you obtain for β_0 and β_1? You should be able to select the right answer without actually implementing linear regression. Give reasons for your answer choice.

(a) β_0 = −569.6, β_1 = 530.9
(b) β_0 = −1780.0, β_1 = −530.9
(c) β_0 = −569.6, β_1 = −530.9
(d) β_0 = −1780.0, β_1 = 530.9