Question 1 [50%]. Linear regression is the process for determining a hyperplane that best represents the general trend of a data set. The simplest form of linear regression involves two variables: y being the dependent variable (output) and $\mathbf{X}=[x_1,x_2,...,x_n]$ being the vector of n independent variables (observations). The equation developed is of the form $y = \mathbf{W}^T\mathbf{X} + b$, where $\mathbf{W}=[w_1, w_2,...,w_n]$ is the vector of the regression coefficients, and b is an offset.

Given a dataset $\mathcal{D} = \{\mathbf{Xi}, yi\}_{i=1}^{M}$ of M annotated samples, you are asked to find $\mathbf{W}$ and b that best represent $\mathcal{D}$.

a. Formalize this problem as an optimization problem. Based on your formulation, indicate if it is a free optimization, constraint satisfaction or constrained optimization problem. Why?

b. If you are asked to solve this problem using an evolutionary algorithm (EA), how would you represent an individual solution? Give an example!

c. What would be the fitness function? Define formally!

d. How could you apply cross-over and mutation? Give one example for each!

e. Assume that the mutation step size, *s*, is the only parameter in your EA design. Develop (and define formally) a parameter control strategy with the following features:
    - changes in *s* are based on feedback from the search progress,
    - some user control of *s* can be imposed,
    - a given *s* acts on all individuals of the population.

Question 2 [15%]. Given the fitness function $f(x) = x^2$, and three individuals in a population: x_1=1, x_2=2, x_3=3:

a. Calculate the fitness for each individual.

b. If the selection is implemented using the roulette wheel algorithm with Fitness Proportional Selection for a transposed fitness function $f'(x) = f(x) + 100$, what is the probability that x_2 is **not** selected ? Write any assumptions you have made.

c. If selection is implemented using the tournament selection, what is the probability that x_2 is **not** selected? Write any assumptions you have made.

Question 3 [15%]. Fill in the spaces with either of the following basic evolutionary algorithm variants: Genetic Algorithm/ Evolution Strategies/ Evolutionary Programming/ Genetic Programming/ Differential Evolution. They can be used more than once.

- No recombination is used in _____.

- In _____, mutation is possible but not necessary.

- _____ scheme uses crossover OR mutation (chosen probabilistically) at every iteration, while _____scheme uses crossover AND mutation sequentially.

- In _____, populations are list rather that sets.

- In _____, recombination creates one child, which can be realized either averaging parental values or selecting one of the parental values.

- In _____, chromosomes are non-linear structures such as trees or graphs.

Question 4 [10%]. What is the difference between experimental design-based tuning and search-based tuning? Max 3 sentences!

Question 5 [10%]. The algorithms in evolutionary computing (or pattern recognition in general) can be tested with (1) real data, (2) benchmark datasets in public repositories, or (3) simulated data. Compare the advantages and disadvantages of these three approaches.