IE 529
Homework set 3

1. Consider the basic single-swap algorithm discussed in class, where we will assume the *new* center is selected randomly. Determine the computational complexity required to complete one iteration of the algorithm. Explain your reasoning.

2. Consider 100 points evenly distributed on a **unit square** to form a 10×10 grid, and 100 points randomly chosen from the uniform distribution on the **unit square**. Suppose you use k-means to find the centroid of each data set (so you are applying k-means with $k = 1$).

   (a) Which of the two data sets is *likely* to have the smaller cost (sum of square error terms), and why? (You can use simulations to explain your answer if you like).

   (b) How do you expect these costs would compare as n, the number of points, gets larger? Provide a sketch of a proof for your answer.

3. As we have and will encounter Jensen's inequality and the *geometric-mean algebraic-mean* (GM-AM) inequality in our readings, we will work through the details of these in this homework problem.

   Preliminary: A subset D of a real vector space (e.g., $\mathbf{R}^d$) is convex (concave) if every convex (concave) linear combination of a pair of points of D is in D, i.e., if $x, y \in D$ and $0 < \alpha < 1$ imply that $\alpha x + (1 - \alpha)y \in D$. A function $f : D \to \mathbf{R}$ is similarly said to be convex (concave) if $f(\alpha x + (1-\alpha)y) \leq (\geq) \alpha f(x) + (1-\alpha)f(y)$. These notions can be extended to linear combinations of any finite number of points, with scalings α_i such that $\sum_i \alpha_i = 1$.

   Prove the following.
   Jensen's inequality: Suppose the function $f : D \to \mathbf{R}$ is a concave function. Assume $x_1, x_2, \ldots, x_n \in D$ and $0 < \alpha_i < 1$ for $i = 1, 2, \ldots, n$ with $\sum_i \alpha_i = 1$. Then

$$\sum_{i=1}^{n} \alpha_i f(x_i) \leq f\left(\sum_{i=1}^{n} \alpha_i x_i\right).$$

   Hints: First note for the case $n = 1$ there is nothing to prove and for $n = 2$ the statement follows immediately from the definitions. So consider $n \geq 3$ and an induction argument. That is, assume the statement is true for some small n, and show it holds for $n + 1$.

   When will equality hold?

4. Now using Jensen's show the **GM-AM inequality** holds:
   Let $\{x_i\}$, $i = 1, 2, \ldots n$, be a set of n non-negative real numbers. Show that the following inequality holds:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \leq \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right).$$

Hint: note that the function $f(x) = logx$ *is concave on* $(0, \infty)$.

5. Here you are given a simple set of data, from which you should formulate a linear model, a quadratic model and a logistic model to fit the data, and compare these three models using the AIC_c test. You can use existing functions in Matlab or Python for all pieces of this problem (for example, in Matlab to fit a logistic model consider the command 'glmfit').

   The data is from a study of the effectiveness of a pesticide (a gaseous agent) on a certain species of beetles. For a total of 481 beetles, the data indicates dose versus mortality: dose is given as $\log_{10}$ of concentration of gas in mg/l. Note that the result is binary, that is the beetle is either "dead" or "alive" following the trial; after n_i trials you have a sample proportion for each dose level. The data is below.

   Please provide your models (equations with the learned parameters), a plot of your models versus the data, and the AIC_c values for each of your models. Based on your simulation plots and AIC_c values, which model structure do you think is most appropriate for the data?

dose	no. dead beetles	total no. beetles
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60