IS 733 Homework 2, Due 3/2/2021 (midnight through Blackboard)

Points

1. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

   (a) Calculate the covariance between age and body fat. Show your working. You can use a calculator for elementary operations, i.e. $+,-,\times$, /, and you don't need to write out all of the numbers. Recall that $Cov(A,B) = E[(A-\bar{A})(B-\bar{B})] = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n} = E[AB] - \bar{A}\bar{B}$. (10)

   (b) What does the answer to the previous question tell us about the relationship between age and body fat? (5)

   (c) List a possible source of noise or errors in this dataset. (5)

   (d) Calculate the 5 number summary for the age attribute, and use this to draw a boxplot for age. (10)

   (e) Perform min-max normalization on body fat to normalize the values to be between 0 and 1. Show your working. Note that in this case the formula simplifies to $v' = \frac{v-\min_A}{\max_A - \min_A}$. (10)

2. Suppose that a data warehouse consists of the four dimensions, *date*, *spectator*, *location*, and *game*, with the data cube measure *charge*, where charge is the fare that a spectator pays when watching a baseball game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

   (a) Design concept hierarchies for each of these dimensions that could be used to encode the levels of aggregation needed for typical analyses of this collection of data. (10)

   (b) How many cuboids are there in the data cube with the concept hierarchies that you've specified (including the base and apex cuboids)? (10)

   (c) Starting with the base cuboid *[date, spectator, location, game]*, what specific OLAP operations should you perform in order to list the total charge paid by student spectators at Camden Yards in 2015? (10)

   (d) List two (2) key differences between OLAP and OLTP, and illustrate the corresponding advantages from using a data warehouse in the scenario we have been considering in this question. (10)

3. Consider the following data set:

x	y	class $\in \{1, 2, 3\}$
0	0	1
4	0	2
1	3	3

   (a) Draw a scatter plot of these three instances. (10)

   (b) On your scatter plot, draw the decision boundary for the nearest neighbor classifier with these instances (using the standard Euclidean distance). (10)

Total: 100