GSCM 410/510 – Summer 2021
Solutions Week 2

Short-Answer Problems

These concepts can appear on the short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1 *What is data wrangling? Why is it important?*

Data almost never arrives ready for analysis. Many issues need to be addressed, such as inconsistent coding of responses, missing data, and superfluous variables. Even with those issues settled, the data usually needs pre-processing, including standardization (or similar) and conversion of categorical variables to indicator variables.

2 *Explain the concept of a row name in a data frame. Describe the default row names and the advantage of replacing them with a suitable column from the read data frame.*

Each row has a unique identifier. By default, the identifiers are the consecutive integers, starting from zero. However, the data file may contain a unique identifier as one of the already existing columns, such as Name for a data file of employees. In that situation, better to replace the default integers with the more meaningful names.

3 *What is a variable transformation for a continuous (numeric, not categorical) variable? In general terms, not specific code, how is a transformation implemented?*

A variable transformation defines a new variable, or creates new values for an exsiting variable, by performing an arithmetic operation on existing variables. To specify in Python, simply specify the arithmetic operation, realizing that variables are identified as part of the data frame in which they exist. For example:

d['Salary000'] = d['Salary'] / 1000

4 *What is a variable transformation for a categorical variable? In general terms, not specific code, how is a transformation implemented?*

5 *What is the distinction between the Pandas .loc and .iloc methods? What is their purpose?*
These two methods subset a data frame, by rows and/or columns. .loc subsets by row name or column name. .iloc subsets by index, i.e., the ordinal position of the row or column, starting with (unfortunately) 0.

Write the Python expression for referring to variables x1, x2 and x3 in the df data frame. most general, df.loc[:, 'x1','x2','x3'], or, sometime works, df['x1','x2','x3']

6 *Explain the following reference to a data frame named data: data[rows, columns]*

A data frame is a 2-D object, rows and columns. Any one data value in a data frame is identified by its row and column coordinates. This notation specifies the name of the data frame and then references data values in one or more rows and columns.

7 *What does it mean to filter rows by data values?*

Filtering subsets a data frame by rows, selecting only those samples that satisfy some logical criterion, such as Gender == 'F', which reduces a data frame down to only those rows of data marked with F as the gender.

8 *What is the purpose of the variable type category? When should it be used?*

A Pandas object refers to a non-numerical object, which is necessarily a categorical variable. But categorical variables can also be integer variables. A category is a newer data type meant specifically to refer to any categorical variables. By default, non-numerical objects are objects, but they can be declared as type category. Same for integer scaled variables that are an integer type, but can, and should, be declared as type category.

9 *What is the purpose of an indicator (or dummy) variable?*

An indicator variable is a numerical representation of a categorical variable. The number of indicator variables formed is equal to the number of levels or categories. A dummy variable is an indicator variable that assumes the value 0 if the category level is not present, and a 1 when present.

10 *Consider an item on a survey with three possible responses D (disagree) N (neutral) A (agree). What indicator variables would be defined and how are the values of those variables determined?*

Three indicators would be defined, one for each potential response: D, N, and A. The values of these variables would be 0's and 1's, with the indicator variables getting a '0' where that response was not present in the initial categorical variable, and a '1' where that response was present in the initial categorical variable.

11 *What are quartiles of a distribution and how are they computed?*

For a given variable, the first quartile (Q1) is the middle number between the smallest number and the median of the data set. The second quartile (Q2) is the median of the data. The third quartile (Q3) is the middle value between the median and the highest value of variable.

12 *How is the inter-quartile range analogous to the standard deviation in terms of both being summary statistics of a distribution of a continuous variable?*

The more variable the values of a distribution, the more extreme are the first and third quartiles of the distribution. The IQR is the positive difference between the first and third quartiles. So the larger the variability of the values of a variable, the larger are both the standard deviation as well as the IQR.

13 *Define outliers of a distribution in terms of its inter-quartile range.*

The traditional definition is that an outlier is beyond 1.5 IQR's from the first or third quartile of the distribution.

14 *What does it mean to say that the median and inter-quartile range are robust to outliers?*

If all the values of a distribution remain the same except that the largest value of the distribution changes from 10 to 10,000,000,000, the median and IQR remain unchanged. On the contrary, the mean and standard deviation will be drastically affected.

15 *a. What is standardization to z-scores?*

A z-score indicates how many standard deviations the original value is from the mean of the distribution. The distribution of z-scores has the same shape as the original distribution, but with a different scaling.

b. What are the mean and standard deviation of a distribution of z-scores?

The mean of a distribution of z-scores is 0, with a standard deviation of 1.

c. What is their range if the distribution is normal.

For a normal distribution, little more than 95% of the values fall within two standard deviations of the mean.

16 *What is the only way to know for sure which rescaling is best for the data values of the predictor variables (features) – MinMax, Standardization, or Robust Scaling?*

A theme of machine learning is to see what works. Keep testing data separate from training data, and do whatever you want with the data, choosing what ultimately works best. Most machine learning algorithms perform better, that is, more accurate forecasts, when the data have about the same scales. Experience is that often it makes no difference which method is chosen, but one does not know in advance if working with new data. Try them out and see if looking for ways to increase the forecast.