NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp Fall 2015

Missingness, Imputation, & KNN

Question #1: Missingness & Imputation for the Titanic Dataset

Load the `titanic3` dataset from the `PASWR` library; this dataset describes the survival status of individual passengers from the Titanic voyage. There are 14 different variables, some of which have quite a bit of missingness.

1. How many variables contain at least one missing value?
    a. What are these variables?
    b. For each variable, what is the extent of missingness (how many missing values are there and what is the percentage of missingness)?
2. How many observations contain at least one missing value?
    a. What is the percentage of missingness from an observation standpoint? (Hint: the `complete.cases()` function might be useful here).
3. How many cells in the data are missing values?
    a. What is the percentage of missingness from a dataset standpoint?
4. What are the different combinations of missingness in the dataset?
5. What kind of missingness do you have for each variable that contains missing values? Give a reason and scenario as to why you believe this.
6. Impute using mean value imputation the age variable.
    a. Graph the distributions of the age variable before and after mean value imputation. Describe what you see. What problems may arise?
7. Impute using simple random imputation for the age variable.
    a. Graph the distributions of the age variable before and after simple random imputation. Describe what you see. What problems may arise?

Question #2: K-Nearest Neighbors with the Titanic Dataset

Continue with the `titanic3` dataset from the `PASWR` library.

1. Impute using the single missing value of the fare variable using simple random imputation. What value was imputed?
2. Plot the simple random imputation of fare against the simple random imputation of age; color this plot by pclass. Describe any trends.
3. Add two points to your plot representing the following passengers:
    a. A 50 year old who paid $400 for their ticket.
    b. A 10 year old whose parents paid $100 for their ticket.
4. What classes would you think these new individuals would belong to?
5. Impute the missing class values for the new passengers using 1 Nearest Neighbor. What were the predicted classes for each passenger?
6. Impute the missing class values for the new passengers using the $\sqrt{n}$ Nearest Neighbor rule. What were the predicted classes for each passenger? Why did they change/not change?

Question #3: Minkowski Distances with the Titanic Dataset

Continue with the `titanic3` dataset from the `PASWR` library.

1. Create a new data frame that includes:
    a. The pclass, survived, sex, age, sibsp, and parch variables from the original `titanic3` dataset.
    b. The simple random imputation of the fare variable you created above.
2. Separate this new data frame into two separate data frames as follows (note that there should be no observations that appear in both data frames):
    a. For observations that are totally complete: all variables.
    b. For observations that are missing a value for age: all variables except age.
3. Use 1 Nearest Neighbor to impute using:
    a. Manhattan distance.

b. Euclidean distance.
c. Minkowski distance with $p = 10$.

4. Overlay and label four separate density curves: one for each of the three 1 Nearest Neighbor imputed age values, and one for the original complete age observations. Describe what you see any why this might be occurring.
5. Use the $\sqrt{n}$ Nearest Neighbor rule to impute using:
    a. Manhattan distance.
    b. Euclidean distance.
    c. Minkowski distance with $p = 10$.
6. Repeat part 4 with the $\sqrt{n}$ Nearest Neighbor solutions. What is happening here?