CST 383: Intro to Data Science
Dr. Glenn Bruns

Reading: Data preprocessing

Purpose. The purpose of this assignment is to make sure you have a good understanding of the process of data cleaning. In real-life projects, data cleaning is usually more than half of the work done in the project, so it will benefit you to develop skill in this area.

Instructions. In Chapter 3 of the 'Python Data Science Handbook', read these sections:

- Handling Missing Data (about 8 pages)

In Chapter 2 of the 'Hands-on Machine Learning' text, read these sections:

- Discover and Visualize the Data to Gain Insights (about 6 pages)
- Prepare the Data for Machine Learning Algorithms (about 10 pages)

You can skip subsection 'Transformation Pipelines' in the 'Prepare the Data' section.

Answer the following questions by downloading and editing preprocessing.txt. Some of the questions come from class lectures.

1. When would you feel more comfortable deleting rows of training data containing missing data?
    a. when the data is missing at random
    b. when the data is not missing at random
2. Which of these statistical functions would you use for imputing values of a feature of type string?
    a. mean
    b. median
    c. mode
3. One way to process outliers is to "clip" very large or very small values. What happens to a very large value when clipping is used?
    a. its value is removed
    b. its value is set to 1
    c. its value is set to the maximum specified in the clipping operation
4. When feature scaling is used on a data frame, do we scale by the values in rows, or by the values in columns?
    a. rows
    b. columns
    c. both
5. Suppose you have a Pandas Series with values 2, -3, 22, 6. What is the new value of 2 after unit-interval scaling?
6. Suppose you have a Pandas Series with values 6, -3, 19, 2. What is the new value of 6 after z score normalization scaling?

7. You should pay special attention to a feature of a data set that is numeric but has only a few distinct values. Why?
    a. the feature should perhaps be treated as a categorical feature
    b. the values of the feature should perhaps be rounded
    c. the values of the feature should perhaps be scaled

Submission. Submit preprocessing.txt on iLearn.