CST 383: Intro to Data Science
Dr. Glenn Bruns

Reading, Geron chapter 2, part 2

Instructions. Please read Chapter 2 of "Hands-On Machine Learning" from the section 'Discover and Visualize the Data to Gain Insights' through the end of section 'Feature Scaling'. Answer the following questions by downloading and editing geron2-2.txt. Some of the questions come from class lectures.

1. Which measure of central tendency is less sensitive to extreme values in a data set? a) mean b) median

2. Which is the correct statement? a) variance is the square of the standard deviation b) variance is the square root of the standard deviation.

3. (Yes/No) If you have attribute values that are missing completely at random, it is generally okay to drop rows of data that contain missing attribute values?

4. What happens when you "impute" a missing value of a feature? a) you remove it, b) you assign a non-NA value to it, or c) you assign NA to it.

5. I have a categorical feature 'role', with possible values 'student', 'TA', and 'instructor'. If I transform this feature to numeric feature(s) using the 'dummy variable' (also known as 1-hot encoding) approach, how many features will I obtain?

6. Which scaling method is preferred if your data has outliers? a) 0-1 scaling b) z-score normalization.

Submission. Submit geron2-2.txt on iLearn.

Grading. 10 points/question.