Nick Jonas

CSC356 Quiz 1

1. (Pg 8) The term regression was coined by Francis Galton while he was studying the fact that children of tall people tend to be shorter than their parents. Because the children were shorter, he called this “regression to the mean”.
2. Attributes are the different groups of data that can be used to draw correlations. For example: “Number of rooms” is an attribute. Feature can also indicate a data type, but it can have several other meanings as well.
3. (Pg 39) Root Mean Square Error (RMSE) is a commonly used performance measure as it gives a higher weight for larger errors. It is therefore fairly sensitive to outliers and is good when they are rare. It is good for data that looks like a bell shaped curve.
4. (Pg 40) An alternate to RMSE is Mean Absolute Error (MAE). MAE is helpful when there are many outlier districts that may affect the result.
5. (Pg 20) Both utility and cost functions help define θ0 and θ1 as a performance measure, but in different ways. A utility function measures how good your model is whereas a cost function measures how bad it is. Cost functions are typically used more than utility functions, especially with linear regression.
6. (Pg 8-10) Classification is the distinguishing of data between classes. For example, a binary classifier of the mnist dataset might be able to identify if a number is or isn’t a 5. Another example is a spam filter, which distinguishes whether an email is spam or not. Clustering, on another note, attempts to detect similar groups of data. For example, if clustering was applied to the readers of a blog, one might be able to find out how many of their readers were female vs male.
7. (Pg 9-10) k-Nearest Neighbors is a supervised learning algorithm, K-Means is an unsupervised learning algorithm.
8. Later
9. Tail heavy distributions means that the tail end of a histogram fades slowly, making it not symmetrical. This makes it a bit harder for some machine learning algorithms to detect patterns
10. Grid search would look at all values within a set or range and find the best one. Randomized Search picks values at random in a range.
11. Nonzero values only on the main diagonal line through the matrix
12. Best case: 1.0