Tae Coding
Introduction to Data Science: CS61
Summer 2018
Class Exercise#6

Date Given: July 3, 2018 Due Date:

==========

There are 2 problems in this homework assignment. Please use Python's Scikit-Learn package to solve these 2 problems.

Text Book: "An Introduction to Statistical Learning" (ISLR).
By James, Witten, Hastie, Tibshirani
Chapter 2: Statistical Learning: Page 53/54, Problem#7.

There is no need to buy this text book. I have copied the problems from the PDF version of this book.

==========

Problem#1

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the *best* value for K to be large or small? Why?

Book: Fundamentals of Machine Learning for Predictive Data Analytics
By: Kelleher, MacNamee, D'Arcy

Chapter 5: Similarity-based Learning: Page 240: Problem#1
==========

Problem#2

1. The table below lists a dataset that was used to create a nearest neighbour model that predicts whether it will be a good day to go surfing.

ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Assuming that the model uses Euclidean distance to find the nearest neighbour, what prediction will the model return for each of the following query instances.

ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
Q1	8	15	2	?
Q2	8	2	18	?
Q3	6	11	4	?