MIS 21 Intersession 2016
Lab #1: Basic Ruby

Objectives

The purpose of this lab is to be able to apply the Ruby programming language in constructing an application to meet certain requirements. It will extensively test your ability to use variables, arrays, hashes, functions/methods, sorting and simple if else statements.

Instructions

Create a ruby application that implements the K Nearest Neighbors (KNN) algorithm. The algorithm takes in an integer parameter k that represents the number of closest neighbors to classify/label an unknown set of data. Data points can be represented by a label and a vector of numbers called features. An example of a data point can be represented by the following:

```
data_point_a = { :label =>"cat", :features => [0.25, 0.1, 5] }
data_point_b = { :label => "dog", :features => [0.3, 0.14, 7] }
```

To get the closeness of data_point_a to data_point_b, we need to implement a function that accepts 2 arrays and uses the euclidean distance to return a distance value. The euclidean distance can be computed by the following:

$$f(v1, v2) = \sqrt{\sum_{i=0}^{n-1} (v1_i - v2_i)^2}$$

Let's say you now have a collection of data points that represents cats and dogs. And someone shows you another data point but we're not sure if it is a cat or a dog. Using the KNN algorithm, we can do the following to classify the unknown data point by doing the following:

1. Define a value for k (let's say k = 5)
2. Determine the closeness the unknown data point's features against each data point in your collection by using the euclidean distnace.
3. Get the k number of data points that has the least distance value (in this case 5 nearest neighbors).
4. From the 5 nearest neighbors, the label with the highest count becomes the label of the unknown data point.

Suggested signature for the knn function:

```
def classify_with_knn( data, k, unknown_data_point)
end
```

Where:

- **data**: an array of data points with labels and features
- **k**: the number of nearest neighbors to get for classification
- **unknown_data_point:** the data point we want to classify

Test Data:

You can use the following to see if your KNN algorith works:

Label	Feature 0	Feature 1	Feature 2
cat	12	1	1
cat	13	3	1
cat	15	1	1
dog	3	12	2
dog	4	11	2
dog	1	11	2
cat	11	4	1
cat	13	2	1
dog	3	13	2
dog	7	12	2

See if you can classify a data point with features [14, 2, 1]. Is it a cat or a dog?