K-means Classification

For this project you will be implementing the K-means clustering algorithm to classify flowers. The Iris dataset (http://archive.ics.uci.edu/ml/datasets/Iris?ref=datanews.io) is provided for you in the form described below. *Do not use any modules that implement kmeans for you!*

Every line of the complete Iris dataset looks like this (150 samples in total):

```
sepal length, sepal width, petal length, petal width, class
```

Each of these values are a real number, except for class which is one of three strings: "Iris Setosa, "Iris Versicolor", or "Iris Virginica."

I have randomly selected 120 samples with their associated labels and placed them into two files: iris-train.data and iris-train.labels. These samples are the training set: the samples that you are to use to train your k-means clustering algorithm. The test set and associated labels exist in iris-test.data and iris-test.labels and consist of the remaining 30 samples. These will be used for testing your model for accuracy. I have given you exactly how I will test your code in a file called test_proj4.py, but when the tests on mimir show up I will not be using the same version of the training/test set that you have.

Now for the functions that you have to implement...

train(data, labels, k=3)
Train takes three arguments as follows:

- data: A list of individual samples of flowers. Each sample should be a list of real numbers.
- labels: A list of labels of flowers. Each label corresponds with the piece of data at the same index in the list of data points.
- k: optional argument, default value is 3. this changes how many clusters you try to make, 3 makes the most sense in terms of this problem, but if you are curious you are welcome to change it.

The train function should return a model that can be used by the classify function. I don't care how you do it, it just has to work when passed into your classify function.

classify(x, model)
The classify function takes two arguments as follows:

- x: a single sample of data without a label. It has the same format as what you trained on.
- model: a model that you trained with the training set

This should return one of the three possible class labels for a sample: "Iris Setosa, "Iris Versicolor", or "Iris Virginica."

I hope that those explanations clear up the confusion from last project. Please put all questions you have on Piazza, I'm pretty quick to answer there.

Hints
In order to do K-means classification and not just clustering, you need add calculation of the majority label in each cluster that you find. The label you assign a sample that you are trying to classify is then just the label of the cluster

it is closest to.

Assigned on 5/03/2017 2:01:00 AM | Due on 5/17/2017 2:01:00 AM 5000 Submissions Remaining

Looks like you don't have any submissions yet

Click submit to submit your code!

Submit Download Starter Code Open IDE