# CS 567 Machine Learning – Programming Assignment 2
# Finding Erratic Ants

Nazar Khan, PUCIT
Fall 2016

| | |
|---|---|
| **Assigned** | Monday, January 9, 2017 |
| **Due** | Monday, January 16, 2017 before 5:30 pm |

## 1 Data

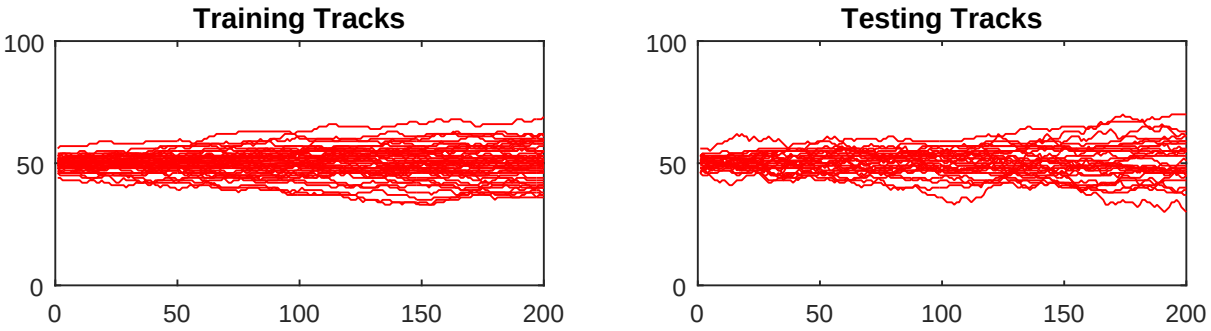


Figure 1: Some samples from training and testing tracks. The test set contains some anamolous tracks.

We will model ant walks and find ants with anamolus tracks. The file tracks.mat contains two variables

1. 'training_tracks' is a 200 by 1000 matrix containing synthetic tracks of 1000 ants consisting of 200 steps each.
2. 'testing_tracks' is another 200 by 1000 matrix containing synthetic tracks of 1000 ants consisting of 200 steps each.

## 2 Model

All tracks in 'training_tracks' come from the same model. The model determines two things

1. starting point
2. location of next step

The starting point of each track follows $\mathcal{N}$(mean_sp,std_sp$^2$) where 'mean_sp' is a fixed location. In simpler words, each ant starts from a location that is normally distributed around 'mean_sp' with standard deviation 'std_sp'.

Given the location at the previous step, the new location can be modelled as previous location + $\mathcal{N}$(0,std_step). In simpler words, each ant moves forward but in somewhat erratic steps and the amount of erratic walk depends on the standard deviation 'std_step'.

Most tracks in 'testing_tracks' come from the same model as 'training_tracks'. However, some tracks are from a different model with greater erraticity 'std_step'.

## 3 Anamoly Detection (10 marks)

Your task is to find outliers in 'testing_tracks'. That is, tracks that do not come from the ant walk model used for generating 'training_tracks'. Plot the outlier tracks in red and the remaining in blue.

An outlier can be defined as one with low probability density p(track) computed using the training tracks. For example, you can use kernel density estimation by implementing the following function

$$\text{p}(\mathbf{x}) = \text{evaluate\_kernel}(\mathbf{x},\text{training\_points});$$

where $\mathbf{x}$ can be the 200 dimensional vector representing a track.

You will also need to implement a function

$$\text{outlier\_indices} = \text{find\_outliers}(\text{training\_points},\text{testing\_points});$$

Notice the use of the variable name 'training_points' instead of 'training_tracks' so that your implementations are generic and can be used for any anamoly detection problems in the future.

Helpful Matlab commands:

- load tracks.mat; loads the variables stored in the .mat file named tracks.mat.
- plot(testing_tracks,'b'); will plot the all testing tracks in blue color.
- plot(testing_tracks(:,17),'r'); will plot the 17th testing track in red color.
- plot(testing_tracks(:,[3 19 589]),'r'); will plot the 3rd, 19th and 589th testing track in red color.
- hold on; will prevent Matlab from overwriting previous plot commands. Subsequent plots will be displayed without removing previous ones.
- print(gcf,'-dpng','results'); stores the currently displayed Matlab figure as an image in the file 'results.png'.

## 4 Bonus (5 marks)

Use testing_tracks to find the models that generated tracks for the

1. normal ants
2. erratic ants

Explain your solution.

## 5 Submission

Submit your_roll_number_PA2.zip in `\\printsrv\Teacher Data\Dr.Nazar Khan\Teaching\Fall2016\CS 567 Machine Learning\Submissions\PA2`. The .zip file should contain

- all relevant .m files,
- result.png image showing the outlier tracks in red and the remaining ones in blue,
- outlier_inds.txt containing the indices of the outlier tracks, and
- solution and explanation of bonus task.