Experiment 5

Monte-Carlo Trials

Instruction: Create a new .Rmd file and write code for each section in separate R code block

Monte Carlo methods estimate quantities by repeatedly generating random samples from a model, computing an outcome each time, and averaging the results. With enough trials, these empirical averages approximate true probabilities, expectations, or distributions (Monte Carlo was a casino in Monaco).

Central Limit Theorem states that the sum (or average) of large number of independent random variables with finite variance becomes approximately normal, regardless of their original distribution. This theorem is used in approximating complicated distributions with normal, deriving confidence intervals and hypothesis tests, simplifying sampling distribution calculations and justifying normal based error analysis in Machine Learning.

1. Generate nsamples=100 samples from uniform distribution (X~U(0,5)). Find sample mean of these 100 samples. Repeat the experiment for ntrials=10000 trials and store sample means in a vector ‘mx’. Plot density histogram of mx and overlay a Gaussian curve.

2. Generate nsamples=100 samples from Exponential distribution (X~Exp(5)). Find sample mean of these 100 samples. Repeat the experiment for ntrials=10000 trials and store sample means in a vector ‘mx’. Plot density histogram of mx and overlay a Gaussian curve.

3. Generate nsamples=100 samples from Gaussian distribution (X~N(0,5)). Find sample mean of these 100 samples. Repeat the experiment for ntrials=10000 trials and store sample means in a vector ‘mx’. Plot density histogram of mx and overlay a Gaussian curve.

4. Generate nsamples=100 samples. Each sample can randomly come from one of the following three distributions (X~U(0,5), X~Exp(5), X~N(0,5)). Find sample mean of these 100 samples. Repeat the experiment for ntrials=10000 trials and store sample means in a vector ‘mx’. Plot density histogram of mx and overlay a Gaussian curve.

5. Exploratory: In each step 1-4, decrease the variable ‘nsamples’ and identify the value when the mean distribution stops being Gaussian (i.e. identify how much ‘nsamples’ should be considered as large).

6. Generate simulated data: Select x in the range (1,5,by=0.01); y=a+b*x+d where a=1.2, b=2.1, d~N(0,2). Plot a scatter plot y vs x.

7. Train a linear model on y vs x. Find the coefficients a, b and mean of error residuals (me).

8. Repeat steps 6-7 for 1000 trials. Evaluate a, b and me for each trial and store it in a vector. Plot three density histogram (a) a, (b) b, and (c) me. Overlay a Gaussian curve in

each case. Use the plot, skewness and Kurtosis values to identify if these variables are from Gaussian distribution.

9. Exploratory: Repeat steps 6-8. In this step assume that y=a+b*x+d where a=1.2, b=2.1, and d is from a t-distribution with degree of freedom as 3.