ST 495 Advanced computing for statistical methods
Homework problem set 1

January 11, 2022

No `R` packages are permitted for use in this assignment.

1. Find a real data set for your course projects. If you are having trouble finding a data set, then consider the data sets available at `https://www.kaggle.com/datasets`. Provide a link for your data set, and a brief description (4-5 sentences) of why you are interested in these data.

2. Load your data set in your `R` script file. Clean, format, print summary statistics, and present a variety of exploratory plots. What are some population features you might be able to learn from your data set?

3. (a) Write a function to return the value of a polynomial of order n, with coefficients $a_0, a_1, \ldots, a_n$, evaluated at a point x. Precisely,

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

   The function must take arguments `(x, a)`, where `x` is a numeric scalar and `a` is an $(n+1)$-dimensional vector (note that $n+1$ is implicitly defined by `length(a)`).

   (b) Vectorize your function for the argument `x` (see, `https://stat.ethz.ch/R-manual/R-devel/library/base/html/Vectorize.html`).

   (c) Using your vectorized function from part (b), plot the polynomial with coefficients $a_0 = 4, a_1 = 1$, $a_2 = -.5$, and $a_3 = 1.9$ over $x \in [-2, 2]$.

   (d) Write the algorithm for Newton's method (from your Calculus 1 course: `https://en.wikipedia.org/wiki/Newton%27s_method`) for finding the roots of a differentiable function, and add this as an optional item to return in your polynomial function from (a). Specifically, your new function takes arguments `(x, a, roots)`, where `roots` is a logical argument with the value `TRUE` indicating that the function should also return a root of the polynomial. Note, your algorithm will need to determine the first derivative of the polynomial. Use your function to return the root of the polynomial in part (c).

4. Monte Carlo experiments.

   (a) Use a Monte Carlo approximation to evaluate $P(X > .5)$ for $X \sim \text{uniform}(0,1)$ (hint: write the probability as an expectation of an indicator function). How many Monte Carol samples do you need to approximate the true value of $P(X > .5)$ to 4 decimal places?

   (b) Law of large number (commonly abbreviated "LLN") results establish that sample means of independent and identically distributed (commonly abbreviated "iid") random variables $X_1, \ldots, X_n$ (with $E(|X_i|) < \infty$) converge to the common mean $\mu := E(X_i)$ as $n \to \infty$. Generate synthetic data sets from 3 different probability distributions and verify that the LLN holds. In each case, how large does n need to be to approximate μ within 4 decimal places of the sample mean?

   (c) Central limit theorem (commonly abbreviated "CLT") results establish that sample means of iid random variables $X_1, \ldots, X_n$ (with a finite second moment), when properly scaled, converge *in distribution* to a Gaussian distribution; precisely,

   $$\sqrt{n}(\overline{X}_n - \mu) \longrightarrow \mathrm{N}(\mu, \sigma^2),$$

   as $n \to \infty$, where $\mu := E(X_i)$ and $\sigma^2 := E[(X_i - \mu)^2]$. Generate synthetic data sets from 3 different non-Gaussian probability distributions and verify that the CLT holds by overlaying the $\mathrm{N}(\mu, \sigma^2)$ density function on top of a histogram of the synthetic data.