CS 365: Introduction to Scientific Modeling
Assignment 2
Probability Distributions and Regression
Part I Due: Wed. Oct. 8, 2014 1:00 pm
Part II Due: Mon. Oct. 13, 2014 1:00 pm

1 Introduction

In this assignment, we will be working with an existing data set, reading the data into MATLAB, making scatter plots and histograms, and using simple curve fitting techniques to test for trends and general patterns (e.g., power laws). We will be using a data set, available from the class web site (`http://cs.unm.edu/~forrest/classes/cs365`). All figures should have their axes labeled in a way that clearly indicates what the x- and y-variables mean, have a helpful legend if appropriate (e.g., if more than one data series is displayed), have axes that range over all of the displayed data but not much more, and use fonts are large enough to read easily.

The assignment is designed with MATLAB in mind, but you are free to use any language you choose. Before writing a program to complete the various sections of the assignments, don't forget to consult the manual to see if there is a built-in function that will give you the information you need.

Parts 2.1 and 2.2 are due at the beginning of class, Oct. 8. Parts 2.3 and the optional 2.4 are due at the beginning of class Oct. 13. Please hand in a written copy of your report and the code you wrote to complete the assignment.

2 Assignment

We will be working with a data set containing information about spam messages sent from various Internet Service Providers (ISPs) around the world in 2013. The data provide information about the number of spam messages sent, the number of IP addresses sending spam, the number of customers each ISP has, the continent on which the ISP is located, and the year and quarter that the data were collected.

2.1 Import Data and Preliminary Analysis

The data are stored in the file `spam.csv`. The file contains information on 295 ISPs in each of the four calendar quarters of 2013. Each row contains 6 different values described in the table.

Column	Variable	Notes
1	ID	ID number for ISP
2	Calendar Quarter	
3	Originating Continent	NA=0, EU=1, AS=2, SA=3, AF=4, Oceania=5
4	Total Spam Messages Sent	
5	Average Number of Spam Sending IPs per week	
6	Number of Subscribers	

1. Download the data file from the class website and import the data into a matrix. Hint: Use the MATLAB function `csvread`, but read the documentation carefully.

2. What is the size of imported matrix (number of rows and number of columns)? Does this match the previous description of the data set? If not speculate why not?

3. How many total messages were sent from each continent during each quarter in the dataset? Format your results in a nicely organized legible table. Which continent sent out the most spam messages (total number)? Are these values a good indicator of which countries are the worst spammers? Why or why not?

4. Next, incorporate the number of subscribers into your analysis to make a more valid comparison, and report the results in a new table.

2.2 Distributions

1. Make a histogram of how many ISPs are located on each continent. Chose one of the four quarters for this calculation. Which continent has the most ISPs?

2. Plot the histogram of the spam messages column. This will be a sorted frequency plot showing the probability of each given number of messages being sent out by a single ISP. Does the histogram resemble any of the distributions we have discussed in class? Why or Why not?

3. Calculate the Complement Cumulative Density Function *ccdf* for the spam messages similar to above (Hint: use `ecdf`). Plot the *ccdf* on regular, semilogy, and loglog axes. Does it seem to follow a power-law or an exponential? Or something else?

4. Compute the least-squares fit for *ccdf* and report the slope (scaling exponent) of the line.

5. Compute the Maximum Likelihood Estimate(MLE) for b (assuming a power law) using the formula given in class:

$$b = 1 + n \times (\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}})^{-1}$$

   How does it match the value you found in 5? Hint: Some ISPs sent 0 spam messages. These should be excluded from the estimate.

6. Find the Maximum Likelihood Estimate (MLE) for the parameter of λ assuming an exponential distribution for spam messages. Hint: The MLE for an exponential distribution with parameter λ is just the reciprocal of the mean of the data: $\lambda = 1/\mu$.

7. Find the Maximum Likelihood Estimate (MLE) for μ (mean) and σ (standard deviation) assuming a log normal distribution. Hint: First calculate the log of the number of spam messages, then find the mean(μ) and standard deviation(σ) of these data.

8. Find the log likelihood of all three estimates. Which is closer to 0? What does this mean? The Log Likelihood is given by:

$$\mathcal{L} = \sum_{i=1}^{n} \log(pdf(x_i))$$

where the

$$pdf(x) = \lambda e^{-\lambda x}$$

for an exponential distribution,

$$pdf(x) = \frac{b-1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-b}$$

for a power-law distribution, and

$$pdf(x) = \frac{1}{x\sigma\sqrt{2pi}} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$$

for the log normal distribution. Hint: Be careful calculating the log likelihood for the log normal distribution. You may need to use some algebra on $\log(pdf(x))$, before you calculate it.

2.3 Curve Fitting

Next, we will conduct some preliminary data analysis, using the following steps:

1. Make a scatter plot of the number of IP addresses sending out spam versus the number of spam messages sent by that ISP in that quarter. That is, include all of the quarters in one plot. Is there a recognizable relationship? Next, redo this plot using logged axes. Is there an easily identifiable relationship?

2. Compute a linear regression for the *log* number of spam messages, using the *log* number of IP addresses as the independent variable. Report the slope and intercept, and plot the line along with the data. Hint: Use `polyval` to help draw the line, once you have the slope and intercept computed.

3. How good is the fit? Compute the coefficient of determination (R^2) for your linear fit. If y_i is the data (*log* (number of messages)) and f_i is the value predicted by the best-fit line at $x = i$ then $R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2}$

4. Do you think this is a good fit for the data? Explain why these two quantities (IP addresses and Messages) are or are not related.

2.4 Extra Credit

1. Using your linear regression, calculate the percent increase in the number of spam messages if the number of IP addresses sending spam increases by 10%.

2. MATLAB has many built-in tools for fitting distribution parameters and calculating likelihoods. Use these tools to find the parameters you found in Section 2.2. Are they the same or different? Why?

3. Fit a multivariate linear regression to characterize the number of spam message sent, using both the *log* of the number of source IP addresses and the *log* of the number of subscribers. Is this a better fit than what you observed in Section 2.3?

4. Fit the linear regression from Section 2.3 for each of the different continents. How do the results differ for each continent? Discuss what you think this means.

3 What to hand in

Hand in a short report (not more than 5 pages). For each section of the assignment, outline the procedures you used, display the plotted data, and then give a short (2-3 sentence interpretation of the results), including your speculations about any anomalies you find in the data.

4 Late Policy

You are allowed three free "late days" to be used at your discretion throughout the semester. After you have used up your late days, I will deduct 10% per day from the grade you would have received on any late work.