INFO 7390
Advances in Data Sciences and Architecture
Exam 1

Student Name: _____
Professor: Nik Bear Brown

Due: Sunday July 22, 2018

All the questions REQUIRE an explanation of the question as well as python code to execute or simulate the question? Yes or no responses get no credit. Any text or code from the Internet MUST be cited.

Q1 (20 Points) The stock market is often modeled as a normal distribution; but is closer to a Weibull distribution. How does a normal distribution differ from a Weibull distribution?

Plot a normal distribution with mean 700 and standard deviation 100. What is the probability of a value being greater than 900?

Use a 2-parameter Weibull distribution to approximate a normal distribution (i.e. find a shape parameter that has a "bell shape" and a scale parameter to is close the normal above.)

See

https://www.johndcook.com/blog/distributions_scipy/

http://www.niar.wichita.edu/coe/NCAMP_Documents/Publications/NCAMP_Techincal_Presentations/Distribution_for_small_sample_sizes_and_fatigue_data.pdf

https://stackoverflow.com/questions/17481672/fitting-a-weibull-distribution-using-scipy

http://faculty.washington.edu/fscholz/DATAFILES498B2008/WeibullBounds.pdf

Plot a normal distribution with mean 700 and standard deviation 100 in the same graph as its Weibull approximation. What is the probability of a value being greater than 900 using the Weibull approximation?

Q2 (20 Points) What is a hypothesis test? What is meant by the null hypothesis and alternative hypothesis?

How does a z-test differ from a t-test?

Create an example problem using a hypothesis test. It MUST use data that you used in assignment 1, 2 or 3.

Show how to solve the problem with hand calculations.

Show how to solve the problem with python code.

Q3 (20 Points) Write an equation for linear regression. Explain how the error is dependent on either the dependent variables? What is the distribution of the error?

Read to following page on GLMs https://onlinecourses.science.psu.edu/stat504/node/216/

How does generalized linear model (GLM) relate to ordinary linear regression?

What is the relationship between the linear predictor(s) and the mean of the response distribution function if one models ordinary linear regression as a GLM?

Linear regression assumes that the error of the dependent variable is normally distributed. What would one do if it is not?

Use a GLM to create generalized linear model assuming the error of the dependent variable is normally distributed on the same data that you used in the linear regression assignment. How does it compare to the results in the linear regression assignment?

Relationship between the dependent and independent variables need not be of the simple linear form. What would one do if it is not?

What is regularization? Why does one use it? Can it be used with a GLM?

Q4 (20 Points) Consider the supervised learning algorithms support vector machines, random forests and multilayer perceptrons.

What are hyper-parameters?

What are the hyper-parameters for the algorithms: 1) support vector machines, 2) random forests and 3) multilayer perceptrons?

Using data, write python code to show the effect of hyper-parameters on each of these algorithms.

Q5 (20 Points) What is the difference between bagging, boosting and stacking?

In python, show an example of bagging, boosting and stacking.

Do gradient-boosted trees, and random forests use bagging, boosting or stacking?

If so, how do gradient-boosted trees differ from random forests?

Create a meta-algorithm to implement a stacked ensemble super-learner in python. It must use at least three base learners.

Did your super-learner help?