Optimization Techniques Final Project

Pre-reading Note: Try to read this assignment *twice* to make sure you grasp the full picture of what is required.

Objective: In this project, we aim to use the optimization tools you learnt to understand the causes promoting Covid-19 spread.

Problem Solving process we should follow:

1. Problem understanding
2. Solutions design
3. Solutions implementation.

Problem understanding is the first step of the problem solving process. One of the best tools for problem understanding is **model fitting**. If you can develop a mathematical model that fits available data (relating inputs to outputs or causes to outcomes) then you know you correctly understand the problem you are studying.

Dataset: We will focus on data-based model fitting. You can use the data at https://www.worldometers.info/coronavirus/. To understand the problem, we will develop **two classes of models**.

Model Class One: Number Of Daily Cases

A model is prepared *for each country* relating a number of candidate factors that can be **measured daily** in this country (model inputs) to the number of daily cases found in this country. Several factors that you may consider is temperature, humidity, number of days since the start of the curfew in this country, etc.

For this class we will try **two model types:**

1) Try to fit the number of daily cases to an **exponential curve** (simply *relate time to the number of daily cases*). This will be our **base model.** If the prediction from other more sophisticated models (involving more factors) is no better than that from the base model, this is an indication that you are not considering the right factors.

Note: Another quick test to avoid wasting time over inadequate factors, is to **compute the correlation coefficient** between the daily factor measurements and daily discovered number of cases.

2) The second model type is a **neural network**, whose inputs are the considered factors and output is the number of daily factors.

Model Class One: Number Of Total Cases

We will consider factors that can be measured for each country vs the **total number of cases** discovered in this country. These may include average temperature, population density, measures reflecting the strength of the economy of the country, measures reflecting the degree of education of the population, the density of a certain animal (suspect as a host) in the country.

A **neural network** is trained to relate the factors values as inputs to the total number of cases as output.

Note: Once again try to compute first the **correlation** between the series containing the factor value for each country and the series containing the total number of cases in each country. If the correlation coefficient is *too low*, there is *no need to consider* the corresponding factor. If your model fitting is successful, this is an indication that you understand the correct causes promoting the disease spread.

Important points to note for efficient, successful model development

1. **Available data should be divided into two groups:**

    a. **Training data**: That is used to develop the model (find the value of its unknown parameters)

    b. **Testing data**: That will be used to test the model accuracy.

2. **Exponential Model Parameters:**

    Optimization methods are used to find the exponential model parameters that best fit the training data (a curve fitting problem).

    With the exponential model, it is instructive to use **plotting and trial and error** to get a reasonable ***initial guess*** of the model parameters. Try a guess, plot the curve with your guess against the actual data.. Continue trying until you get a reasonable starting curve.

You may **use your initial guess as input to Newton's method** to get more accurate estimates of model parameters.

3. **Neural Networks Parameters:**

   Neural networks are typically trained (optimized) using a **backpropagation algorithm** (which is a gradient descent variant).

   For neural networks, be careful with your choice of **initial weights** and **inputs scaling**. Note that conventional neural nets employ sigmoids. For too large sigmoid inputs and too small ones, **sigmoids saturate** i.e. all inputs to them below a small threshold or above a large threshold give 0 or 1, respectively. The curve is flat for too small or too large inputs, the derivative is near zero and **no training is possible.**

4. **Avoid over-fitting.** Too small errors on training data can lead to poor generalization. The network simply memorizes the training examples but responds poorly on testing data.

5. **Credit Assignment Problem** *If the model fails to fit the test data, then to improve your model you need to solve a "credit assignment" problem*. Credit assignment means you need to determine the cause behind the model poor performance: several causes are possible: try to investigate them in this order (or think yourself about the best order to investigate them):

   a. Is the **method** you are using for model fitting (**training algorithm**)suitable? Does the method succeed in finding model parameter values that fit the training data with low error? If not, does trying an alternative optimization method (training algorithm) improve the situation?

   b. Are you using a **suitable model**? Your model may be *too simple or too complex* (too complex is equally bad). There are alternative neural network types as well as alternative modelling strategies to neural networks. Does trying these alternatives improve the situation?

   c. The **data you are using** may not be sufficient or cover all possible scenarios adequately. Does increasing the data (training examples) improve the situation? The examples you are using may not be representative of all situations.

d. Are you considering the **correct inputs**? you may not be considering all **relevant factors** that actually promote the spread of the disease. Also, some factors may not be independent from other factors. For example, the effect of population density depends on the number of curfew hours. Also, some factors may be delayed i.e. the effect of a factor's value may appear a few days later. Think of these points and try to overcome them.

e. **Data Inaccuracy.** Is the data accurate? Number of daily cases/ total cases reported for some countries may not be accurate enough.

Report And Delivery Details:

Important notes regarding report preparation:

1. Try to well organize the report:

   First: An **introduction section** describing report motivation.

   Second: A **results section** describing each model you tried (data, code and results) and reporting its results.

   Third (and most important): **Discussion and Learnt Lessons section** commenting on the results you obtained, particularly on how you solved the credit assignment problem whenever it met you. **Comment on any results** you obtained that appeared to be **inconsistent** or different from what you believe to be logical and **how you resolved the inconsistencies** (how you found why the results are different from what you expected).

2. You may use optimization and neural network toolboxes (ready packages). You may try neural net alternatives that were not part of the course if you wish.

3. Organize yourselves into groups (**maximum 5 members**).

   Deliver the group members to your representative eng. Amr El-zawawy. Amr would give each group an ID and give me the full list by next Saturday. (Name the project folder when you deliver it neuralnet_"groupid".

4. **Deadline:** You should deliver the project before 15 May 2020
