CS909 2020 Assignment 2: Regression

(by Fayyaz Minhas)

Due: **11Mar2020 12pm**

Inspired from the LYSTO challenge (https://lysto.grand-challenge.org/LYSTO/), In this assignment, the objective is to develop a regression model for calculating the number of certain type of cells (called lymphocytes) in a given histopathology image patch. For this assignment, all you have to know is that these cells appear in the given image (technically called a immunohistochemistry or IHC image) with a blue nucleus and a brown membrane. Your task is to develop a machine learning model that uses training data (patch images with given cell counts) to predict cell counts in test images.

The data 'breast.h5' can be downloaded from: http://shorturl.at/fuCEO

The subset of the challenge dataset that you have been given focuses on breast tissue images from a total of 18 different individuals. You can read the data as follows:

```
import h5py
import numpy as np
D = h5py.File('breast.h5', 'r')
X,Y,P = D['images'],np.array(D['counts']),np.array(D['id'])
```

Here, X, Y and P contain the Images, Cell Counts, and Patient IDs, respectively.

Training and Testing: Use data from patient IDs 1-13 for training and cross validation and 14-18 for testing. Be sure not to test on the images of patients you have used in your training. Each image is in RGB space so it is represented by an array of size 299x299x3 where the first two dimensions correspond to the width and height of the image and the last three correspond to the R,G and B channel.

Submission: You are expected to submit a **single Python Notebook** containing all answers and code.

Question No. 1: (Showing data) [20 Marks]

Load the training and test data files and answer the following questions:

i. How many training and test examples are there? [2 marks]

ii. Show some image examples using `plt.imshow`. Describe your observations on what you see in the images and how it correlates with the cell count (target variable). [2 marks]

iii. Plot the histogram of counts. How many images have counts within each of the following bins? [3 marks]

- 0 (no lymphocytes)
- 1-5
- 6-10
- 11-20
- 21-50
- 51-200
- >200

iv. Pre-processing: Convert and view a few images from RGB space to HED space and show the D channel which should identify the brown elements in the image. For this purpose, you can use the color separation notebook available here: https://scikit-image.org/docs/dev/auto_examples/color_exposure/plot_ihc_color_separation.html [5 marks]

v. Do a scatter plot of the average of the brown channel for each image vs. its cell count. Do you think this feature would be useful in your regression model? Explain your reasoning. [3 marks]

vi. What is the number of images for each patient? Do you think this can have an impact on your regression model? [2 marks]

vii. What performance metrics can you use for this purpose? Which one will be the best performance metric for this problem? Please give reasoning. [3 marks]

Question No. 2: (Feature Extraction and Classical Regression) [50 Marks]

i. Extract features from a given image. Specifically, calculate the:
   a. average of the "brown", red, green and blue channels
   b. variance of the "brown", red, green and blue channels
   c. entropy of the "brown", red, green and blue channels
   d. Histogram of each channel
   e. PCA Coefficients (you may want to use randomized PCA or incremental PCA, see: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html)
   f. Any other features that you think can be useful for this work. Describe your reasoning for using these features.

   Plot the scatter plot and calculate the correlation coefficient of each feature you obtain vs. the target variable (cell count) across all images. Which features do you think are important? Give your reasoning. [20 marks]

ii. Try the following regression models with the features used in part-I. You can do 3-fold cross-validation analysis (https://scikit-learn.org/stable/modules/cross_validation.html) to select feature combinations and optimal hyper-parameters for your models. Report your results on the test set by plotting the scatter plot between true and predicted counts for each type of regression model. Also, report your results in terms of RMSE, Correlation Coefficient and R2 score (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics). [30 Marks]
   a. Ordinary Least Squares (OLS) regression
   b. Multilayer Perceptron (in Keras or PyTorch).
   c. Ridge Regression (Required For MSc. Students only)
   d. Support Vector Regression (Required For MSc. Students only)

Question No. 3 (Using Convolutional Neural Networks) [30 Marks]

Use a convolutional neural network (in Keras or PyTorch) to solve this problem by directly in much the same was as in part (ii) of Question (ii). You are to develop an architecture of the neural network that takes an image directly as input and produces a count as the output. You are free to choose any network structure as long as you can show that it gives good cross-validation performance. Report your results on the test set by plotting the scatter plot between true and predicted counts for each type of regression model. Also, report your results in terms of RMSE, Correlation Coefficient and R2

score. You will be evaluated on the design on your machine learning model, cross-validation and final performance metrics. Try to get the best test performance you can.

Based on your models, you may want to participate in the challenge as well and report your challenge scores (optional but will you can get a bonus if your rank high in the challenge).