Data Mining 2022 Assignment 2: Prediction of Cellular Composition

(by Fayyaz Minhas)

Due: 12 noon Wed 23th March 2022 (UK Time)

In this assignment, the objective is to develop a regression model for predicting the number of six different types of cells in a given histological image patch. **Your task is to develop a machine learning model that uses training data (patch images with given cell counts) to predict cell counts of each type in test images. Counts of different types of cells in a given image patch is called its cellular composition.**

Cell Type	Count
T1	0
T2	0
T3	410
T4	41
T5	3
T6	292

We shall be using the data for the CoNIC: Colon Nuclei Identification and Counting Challenge (see: https://github.com/TissueImageAnalytics/CoNIC and https://arxiv.org/abs/2111.14485 for details about the challenge). However, you are not required or expected to participate in the formal challenge. The data files required for this coursework (counts.csv, images.npy) can be downloaded from: http://shorturl.at/fsGNU. This link also contains other files including a readme and binding licensing information for this dataset. You will also need to download the fold split file (split.txt) from https://github.com/foxtrotmike/CS909/blob/master/2022/A2/split.txt.

You can read the data as follows:

```
import numpy as np
import pandas as pd
X = np.load("images.npy")#read images
Y = pd.read_csv('counts.csv')#read cell counts
F = np.loadtxt('split.txt')#read fold information
```

The data consists of 3 folds (in variable F). Here, X is a matrix containing N 256x256x3 Red-Green-Blue (RGB) channel images of size 256x256 pixels and Y is an Nx6 matrix with each row corresponding to a single image patch and each column corresponding to the 6 cell types (called T1: neutrophil , T2: epithelial T3: lymphocyte, T4: plasma , T5: eosinophil and T6: connective). Thus, for a single image, you are given the counts of each of the six cell types.

Training and Testing: Use data from the first two folds for training and validation and the third fold for testing. Wherever applicable, performance metrics for the test fold (3rd Fold) are two be reported unless otherwise specified.

Submission: You are expected to submit a **single Python Notebook** containing all answers and code. Include all prediction metrics in a presentable form within your notebook and include the output of the execution of all cells in the notebook as well so that the markers can verify your output. **Also submit a consolidated table of your performance metrics within the notebook to indicate which model performs the best (MANDATORY).**

Question No. 1: (Data Analysis) [20 Marks]

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

i. How many examples are there in each fold? [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 counts of different types of cells especially T3 cells. [2 marks]

iii. For each fold, plot the histogram of counts of each cell type separately (6 plots in total). How many images have counts within each of the following bins? [4 marks]

    0
    1-5
    6-10
    11-20
    21-20
    31-40
    41-50
    >50

iv. Pre-processing: Convert and show a few images from RGB space to HED space and show the H-channel which should indicate cellular nuclei. 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 H-channel for each image vs. its cell count of a certain type for images in Fold-1 (6 plots in total). Do you think this feature would be useful in your regression model? Explain your reasoning. [4 marks]

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

Question No. 2: (Feature Extraction and Classical Regression) [40 Marks]

For the following questions, use only T3 type of cells as the output prediction variable.

i. Extract features from a given image. Specifically, calculate the:
    a. average of the "H", red, green and blue channels
    b. variance of the "H", red, green and blue channels
    c. entropy of the "H", red, green and blue channels
    d. Any other features that you think can be useful for this work. Describe your reasoning for using these features.
    HINT/Suggestion: You may want to use PCA Coefficients of image data (you may want to use randomized PCA or incremental PCA, see: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html). In case of computational complexity, you can reduce the number of images being used in determining the PCA basis. You can also look at other features such as GLCM features (https://scikit-image.org/docs/dev/auto_examples/features_detection/plot_glcm.html) or transfer learning features. You can resize the images if needed.

Plot the scatter plot and calculate the correlation coefficient of each feature in Q(2i,a-c) 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. Plot the scatter plot between true and predicted counts for each type of regression model for the test data. Also, report your prediction performance in terms of RMSE, Pearson Correlation Coefficient, Spearman Correlation Coefficient and R2 score (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics) on the test data. [20 Marks]
   a. Ordinary Least Squares (OLS) regression
   b. Support Vector Regression

Question No. 3 (Using Convolutional Neural Networks) [40 Marks]

a. Use a convolutional neural network (in Keras or PyTorch) to solve this problem in much the same was as in part (ii) of Question (2). You are to develop an architecture of the neural network that takes an image directly as input and produces a count as the output corresponding to T3 cells. You are free to choose any network structure as long as you can show that it gives good performance. Report your results on the test examples by plotting the scatter plot between true and predicted counts on the test data. Also, report your results in terms of RMSE, Pearson Correlation Coefficient, Spearman Correlation Coefficient and R2 score. You will be evaluated on the design of your machine learning model and final performance metrics. Try to get the best test performance you can. Please include convergence plots in your submission showing how does loss change over training epochs. [20 Marks]
b. Use three fold cross validation with your optimal network architecture to predict the counts of T3 cells. Do 3-fold cross-validation with the given folds and report the results for each test fold in the form of separate predicted-vs-actual count scatter plots (3 folds so 3 plots in total) using your model and report your results in terms of RMSE, Pearson Correlation Coefficient, Spearman Correlation Coefficient and R2 score for each fold separately. [5 Marks]
c. Use a convolutional neural network (in Keras or Pytorch) to predict the counts of 6 types of cells simultaneously given the image patch as input and perform 3-fold cross-validation using the given folds. You are free to choose any network structure as long as you can show that it gives good cross-validation performance. Do 3-fold cross-validation using the specified folds and report the results for each test fold for each cell type in the form of separate predicted-vs-actual count scatter plots (3 folds, 6 cell types so 18 plots in total) using your optimal machine learning model and report your results in terms of RMSE, Pearson Correlation Coefficient, Spearman Correlation Coefficient and R2 score for each cell type and each fold separately along with the average of each cell type across the 3 folds. [15 Marks]

   HINT:
   A naïve (but possibly effective) strategy can be to be simply try the same network architecture in 3(a) and train 6 different models separately for each cell type.

Restrictions:

Please do not share these data files or the assignment solutions publicly. Each student needs to submit a single solution and a solution should be developed by the student without assistance from other students. Students must adhere to the licensing information for this dataset.

Useful hints:

Feel free to resize the images to reduce the amount of required compute. However, if you do this, please ensure that the code for doing this is included in your submission notebook.

Look at glob (https://docs.python.org/3/library/glob.html) to get list of all file names in a given folder.

Look at skimage for image reading and image resizing (https://scikit-image.org/) and converting images to HED Space (see code below)

```
from skimage.color import rgb2hed
import skimage
from skimage.io import imread
print('skimage version',skimage.__version__)
import matplotlib.pyplot as plt
I = X[0]/255.0 #read sample image and rescale pixel range in it
I_hed =  rgb2hed(I) #convert to HED
plt.imshow(I);plt.title('Original Image');plt.show()
I_h = I_hed[:,:,0]; plt.figure(); plt.imshow(I_h,cmap='gray');plt.colorbar();plt.title('H
Channel');plt.show()
I_e = I_hed[:,:,1]; plt.figure();  plt.imshow(I_e,cmap='gray');plt.colorbar();plt.title('E
Channel');plt.show()
I_d = I_hed[:,:,2]; plt.figure();  plt.imshow(I_d,cmap='gray');plt.colorbar();plt.title('D
Channel');plt.show()
```

For calculating various regression metrics, please see: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics