Student: Edwin Meyers
Date: 01-Aug-20

Machine Learning Engineering Nanodegree
Udacity

Capstone Proposal:
Classifying the severity of prostate cancer from microscopy scans of prostate biopsy samples using different machine learning techniques

Dear reader,

This Capstone Project is based off of a challenge from Kaggle.com named “Prostate cANcer graDe Assessment (PANDA) Challenge” (link).

Please find below details regarding how I plan to approach solving the challenge:

1. Domain background[1]:

Radboud University Medical Center, Karolinska Institute, and Tampere University teamed up to create a Kaggle competition focused on diagnosing pancreatic cancer in biopsies of pancreatic tissue samples.

2. Problem statement[2]:

Over 1.2 million new cases of prostate cancer (PCa) are reported each year.[3] PCa represents 7% of all new cancer cases in men.[4] The key to lowering the number of PCa deaths is developing more precise diagnostics. Diagnosis of PCa is based on the grading of prostate tissue biopsies. These tissue samples are examined by a pathologist and scored according to the Gleason grading system. Two pathologists may diagnose two sample tissues differently, which could have potentially life-impacting effects on patient treatment and outcome.

In more project-based terms, this is a classification problem where a model will have to be trained such that if given a prostate biopsy image as an input, it would be able to adequately predict a rating (ie. ISUP score) on how affected by cancer it is.

3. Datasets and inputs[5]:

File: [train/test].csv

- image_id: ID code for the image and image file.
- data_provider: Name of the institution that provided the data. Both the Karolinska Institute and Radboud University Medical Center contributed data. They used different scanners that resulted in slightly different maximum microscope resolutions. As expected, they also worked with different pathologists for labeling their images.
- isup_grade: Only available for training set. The target variable. The severity of the cancer on a 0-5 scale.
- gleason_score: Only available for training set. An alternate cancer severity rating system with more levels than the ISUP scale. For details on how the gleason and ISUP systems compare, see the Additional Resources tab.

Folder: [train/test]_images

---

[1] From the Kaggle challenge description
[2] From the Kaggle challenge description
[3] US National Library of Medicine and National Institute of Health (link)
[4] US National Library of Medicine and National Institute of Health (link)
[5] From the Kaggle challenge description

- These are the images of the biopsies. Each is a large multi-level tiff file. Slightly different procedures were in place for the images used in the test set than the training set. Some of the training set images have stray pen marks on them, but the test set slides are free of pen marks. Each individual image is quite large (~20MB)
- Labels are imperfect because determining ground truth in this field is difficult. Even expert pathologists with years of experience do not always agree on how to interpret a slide. As such, the labels aren't always necessarily correct. This will make training models more difficult, but increases the potential medical value of having a reliable alternate opinion. Here are additional details about how consistently the pathologist's labels matched.

Folder: train_label_masks:
- Segmentation masks showing which parts of the image led to the ISUP grade. Not all training images have label masks, and there may be false positives or false negatives in the label masks for a variety of reasons. Mask values depend on the data provider. We can already foresee special considerations that will have to be taken when handling these different methods of labeling the masks:
  - Radboud: Prostate glands are individually labelled. Valid values are:
    0: background (non tissue) or unknown
    1: stroma (connective tissue, non-epithelium tissue)
    2: healthy (benign) epithelium
    3: cancerous epithelium (Gleason 3)
    4: cancerous epithelium (Gleason 4)
    5: cancerous epithelium (Gleason 5)
  - Karolinska: Regions are labelled. Valid values are:
    0: background (non tissue) or unknown
    1: benign tissue (stroma and epithelium combined)
    2: cancerous tissue (stroma and epithelium combined)

File: sample_submission.csv: Example of a valid submission file.

4. Solution statement

The objective is to develop and train a model that can predict a reliable USIP score for new PCa biopsy images.

5. A benchmark model

This Capstone project is based on a Kaggle competition. Kaggle has a leaderboard that displays the precision of other participants' models. Please find the leaderboard in the following url: *https://www.kaggle.com/c/prostate-cancer-grade-assessment/leaderboard*

There are over submissions listed in the leaderboard. Submissions achieved an evaluation score of 71.9 on average and a maximum of 93.4 (see "Section 6. Evaluation metrics" for more information on significance of metric). Given that this is my first Kaggle competition, I will be happy if my submission can score higher than half of the leaderboard.

6. Evaluation metrics[6]

The competition requests submissions include predictions of the `isup_grade` for each `image_id`. Submissions are scored based on the quadratic weighted kappa, which measures the agreement between two outcomes. Since the hidden test pictures have been evaluated by multiple pathologists, this metric can vary from 0 (random agreement) to 1 (complete agreement). In the event that there is less agreement than expected, the metric may go below 0. More details on how the quadratic weighted kappa is calculated is available in the Evaluation section of the Kaggle challenge webpage.

7. Project design outline:

Preliminarily speaking, the solution is expected to be reached by implementing the steps described below (feedback and recommendations would be appreciated for matters relating to picture manipulation and use in learning models).

Step 1: Exploratory data analysis:

A. Understand contextual information regarding the project and data available:
   - How many pictures come with masks? Will it be enough to develop an initial model that can detect segments to focus on?
   - Discover any pre-processing that may be needed in the *train.csv* and *test.csv* files. For example, different labeling practices will have to be standardized.
   - Determine how are Gleason and ISUP scores distributed in total and by source (ie. Karonlinska or Radboud).
   - Understand what are the core components of masks and how masks segmentations may vary one from another.

B. Developing preliminary approaches to better understanding image data and potential methods to effectively estimate ISUP scores:
   - Define a portion of images to perform Exploratory Data Analysis (EDA). Perhaps 10% will be a good enough sample size for image EDA. Running tests on the entire dataset would be inefficient given how large it is.
   - Developing preliminary functions to better explore image data:
     - Background identification and labeling function/model - Although some segmentation masks contain information that will help train such a model, it may be worth exploring unsupervised clustering methods to separate the background from the sample. (mentor feedback appreciated)
     - Tissue contour identification function/model - How to differentiate background pixels outside the tissue sample from those that are inside the tissue samples (ie. part of tissue formations and cell structures). (mentor feedback appreciated)
     - Color variance of tissue sample (color variance values including background pixels within contour and values excluding such pixels, always excluding pixels outside of sample contour.
   - Perform EDA on image data and hopefully uncover potential tendencies that would serve as additional input for a model.

[6] From the Kaggle challenge description

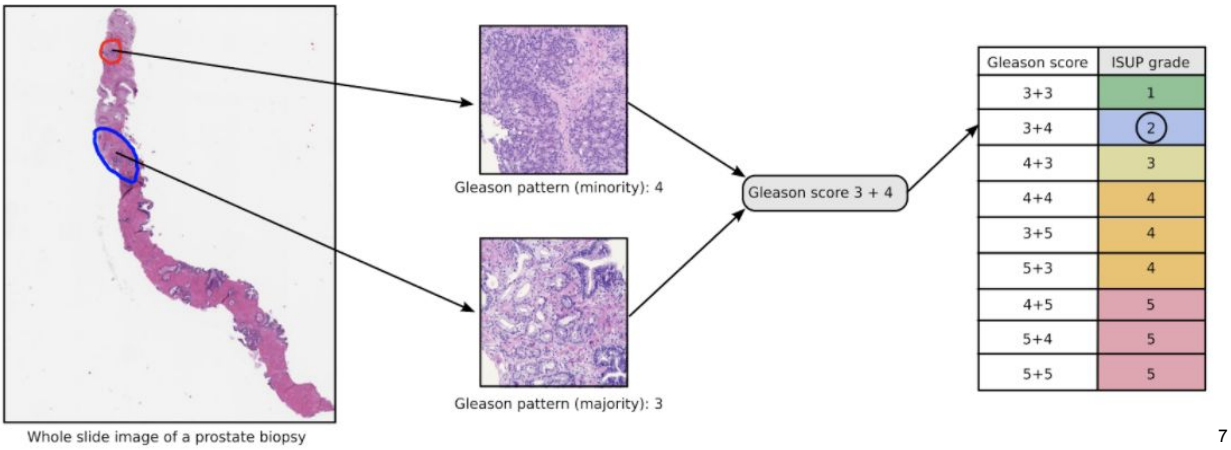
- Color variance of tissue samples plotted against respective ISUP Scores (sources used different biopsy preparation methods, resulting in inconsistent colorations between biopsies. Perhaps relative biopsy coloring variance is not affected by such processes)
- Tissue densities plotted against respective ISUP Scores.

Step 2: Defining areas of interest in the biopsies:

A. Develop a model that can binarily distinguish tissue samples from background with 100% certainty and apply corresponding masks to the entire database.
B. Develop method to focus on areas in sample tissue that may be cancerous
   - Define training, validation, and testing datasets from train_images images. Test images (3 in total) are too few to conduct appropriate testing.
   - A first model should be trained to separate benign and unidentified tissue portions from areas of concern (mentor help appreciated). Given that only a portion of the images have their corresponding segmentation masks, we will likely have to rely on unsupervised learning methods to train a model to separate benign from non-benign areas. Preliminarily speaking, it seems that using clustering or classification methods would be a good way to identify areas of interest. Color and formational features of the biopsies will serve as good input for a clustering model to correctly classify areas of interest from benign areas binarily. Areas identified as not benign are to be identified but not yet graded at this stage.

Step 3: Assign an ISUP score to defined areas of interest:

A. Identified “non-benign” areas will have to be zoomed in and/or labeled intelligently. Data regarding average size and dimensions of segmentation masks could be useful in defining the shape and size of the focus frame/mask. Certain or absolute consistency on how an area of interest is marked off will be crucial for developing a neural network or other classification model in the next step. (mentor help appreciated)
B. Train different models to determine ISUP grade:
   - Using segmentation masks and corresponding training image data to develop a preliminary classification model (segmentation masks show which parts of an image led to an ISUP grade) to determine ISUP grade. Functions previously developed to calculate relative color variance and inside-tissue-contour density should be applied to areas of interest and used as additional inputs for the trained models. Preliminarily, it is expected that a neural network model (eg, vgg16) or a random forest classifiers could be good models to use in this step. Different types of models will be tested to determine which determines the most accurate diagnosis (mentor help appreciated). As mentioned before, special consideration has to be given to inconsistent labeling.
C. Looking forward, the model that will be used to identify and evaluate areas of interest could be replaced with a potentially better model by using the method described in Step II (iii) further below.



[7]

Detailed examples of biopsy close-ups and their respective Gleason indices (aka ISUP) (in reference to four sample biopsies below. Pictures and description obtained directly from Kaggle competition's website):

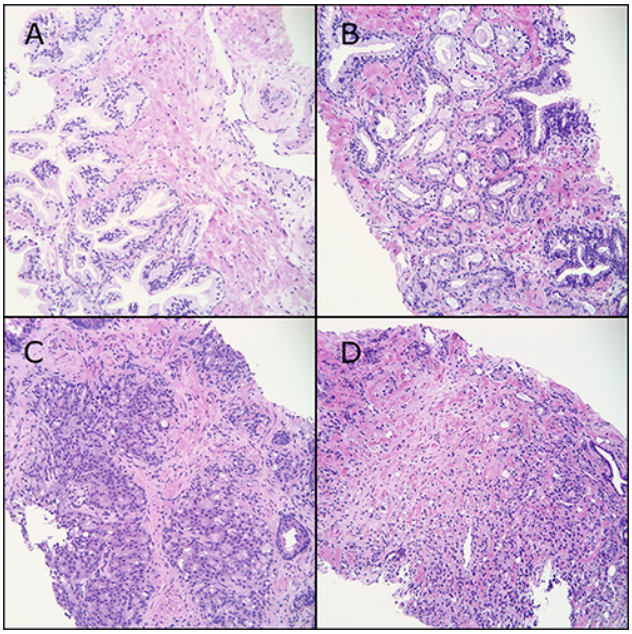
A. Benign prostate glands with folded epithelium. The cytoplasm is pale and the nuclei small and regular. The glands are grouped together.

B. Prostatic adenocarcinoma - Gleason Pattern 3 has no loss of glandular differentiation. Small glands infiltrate between benign glands. The cytoplasm is often dark and the nuclei enlarged with dark chromatin and some prominent nucleoli. Each epithelial unit is separate and has a lumen.

C. Prostatic adenocarcinoma - Gleason Pattern 4 has partial loss of glandular differentiation. There is an attempt to form lumina but the tumor fails to form complete, well-developed glands. This microphotograph shows irregular cribriform cancer, i.e. epithelial sheets with multiple lumina. There are also some poorly formed small glands and some fused glands. All of these are included in Gleason Pattern 4.

D. Prostatic adenocarcinoma - Gleason Pattern 5 has an almost complete loss of glandular differentiation. Dispersed single cancer cells are seen in the stroma. Gleason Pattern 5 may also contain solid sheets or strands of cancer cells. All microphotographs show hematoxylin and eosin stains at 20x lens magnification.

[7] Picture from the Kaggle challenge description



Step 4: Hyperparameter tuning to improve desired prediction outcomes:
Self explanatory

Additional pre-/post-processing steps should be taken in parallel when working with this data:

Step I: Given that there are two data sources (ie. Radbound and Karolinska) with slightly differing image qualities, special consideration should be taken. Mixing both sources when training the data may potentially be such a consideration that should be taken to make sure there are no biases implicit to image qualities that could potentially affect the quality of a prediction. Additionally, it may be beneficial to rotate and flip the images to help the model, especially if using neural networks, so as not to overfit.

Step II: Given that there two doctors can give different diagnoses, special consideration should be taken. Training a model with existing doubt on what is ground truth will be challenging and could require a combination of (i) eliminating pictures with doubtful ground truths before training models, (ii) creating average Gleason scores for conflicting diagnoses before training models, (iii) once a first model is created, use it to predict label on both test and train data, and only keep pictures that were correctly predicted by the model, with a given error margin. Then use those pictures to train a new model that will theoretically be trained on pictures that are known to generate greater consensus between pathologists and the predictor.[8]

Step III: If time permits, it would be useful to build the model in the cloud and to create an application that would allow users to upload the images directly through an internet portal.
Self explanatory. Time restrained step.

[8] This idea is borrowed from the Kaggle competition's winning team's strategy (link)