Machine Learning 4/M coursework 2017

Ke Yuan

09 February, 2017

1. Introduction

Tissue microarray (TMA) is a recent innovation in the field of pathology. A TMA (Figure 1) contains many small representative tissue samples from hundreds of different cases assembled on a single histologic slide, and therefore allows high throughput analysis of multiple specimens at the same time. Beck et al (2011) constructed a machine learning framework that extract features from TMAs and predict patient survival. The core of the framework is a classifier (Figure 2) that distinguish between epithelial (where cancer cells live) and stromal (where immune cells and other normal cells live) regions. The classifier identifys epithelial and stromal regions from images in large patient cohorts, allowing of quntification of the interaction between cancer cells and normal cells. In `epi_stroma_data.tsv`, you will find data to train the classifier.

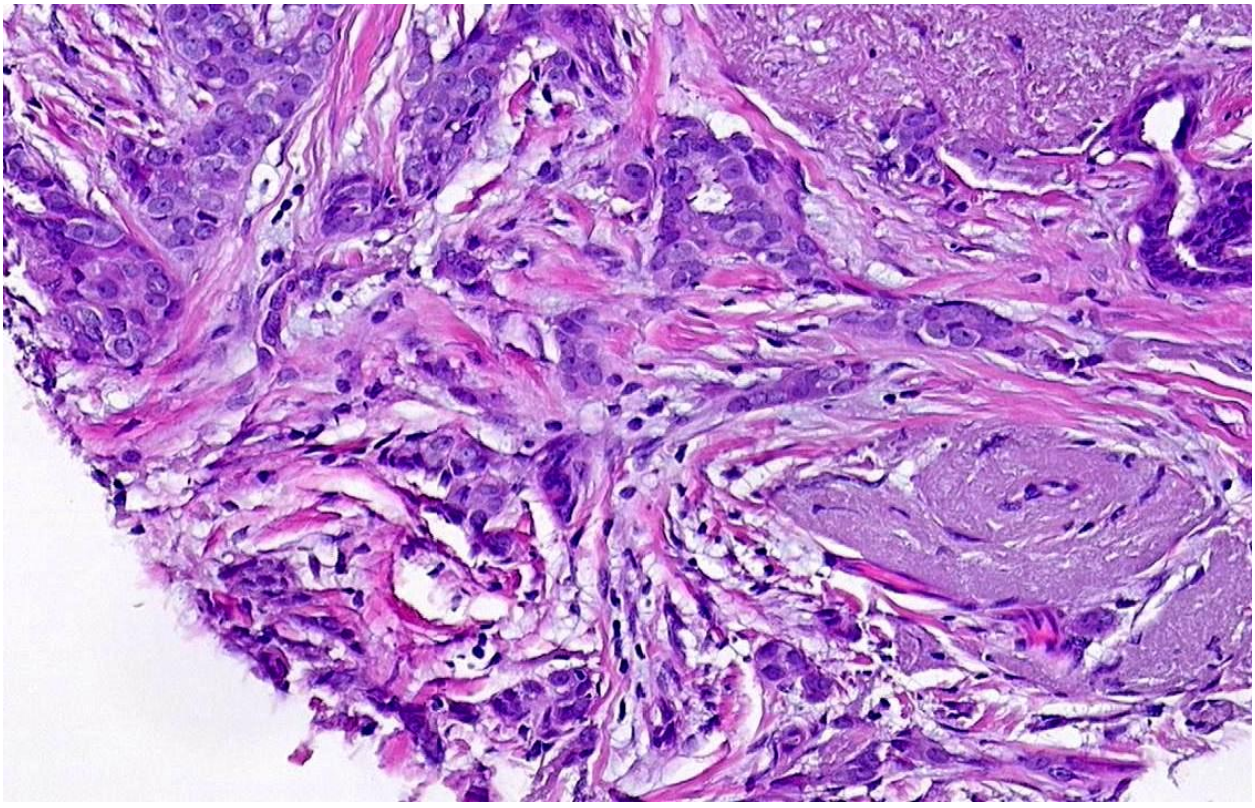

Figure 1: An example of a TMA from a breast cancer patient.

The file contains 7981 data points and 112 features. Each data point is constructed from small regions of coherent appearance known as superpixel. The first column EpiOrStroma is the class for each object with 1 representing epithelial region (red in Figure 2) 2 representing stromal region (green in Figure 2). The remaining 112 columns are features extracted from TMAs (Figure 1) using standard computer vision pipelines (e.g. image segmentation, edge detection, texture features, etc). Figure 3 shows the data in features `GLCM.Ang..2nd.moment..quick.8.11..Layer.1..all.dir..` and `Ratio.Layer.3`.

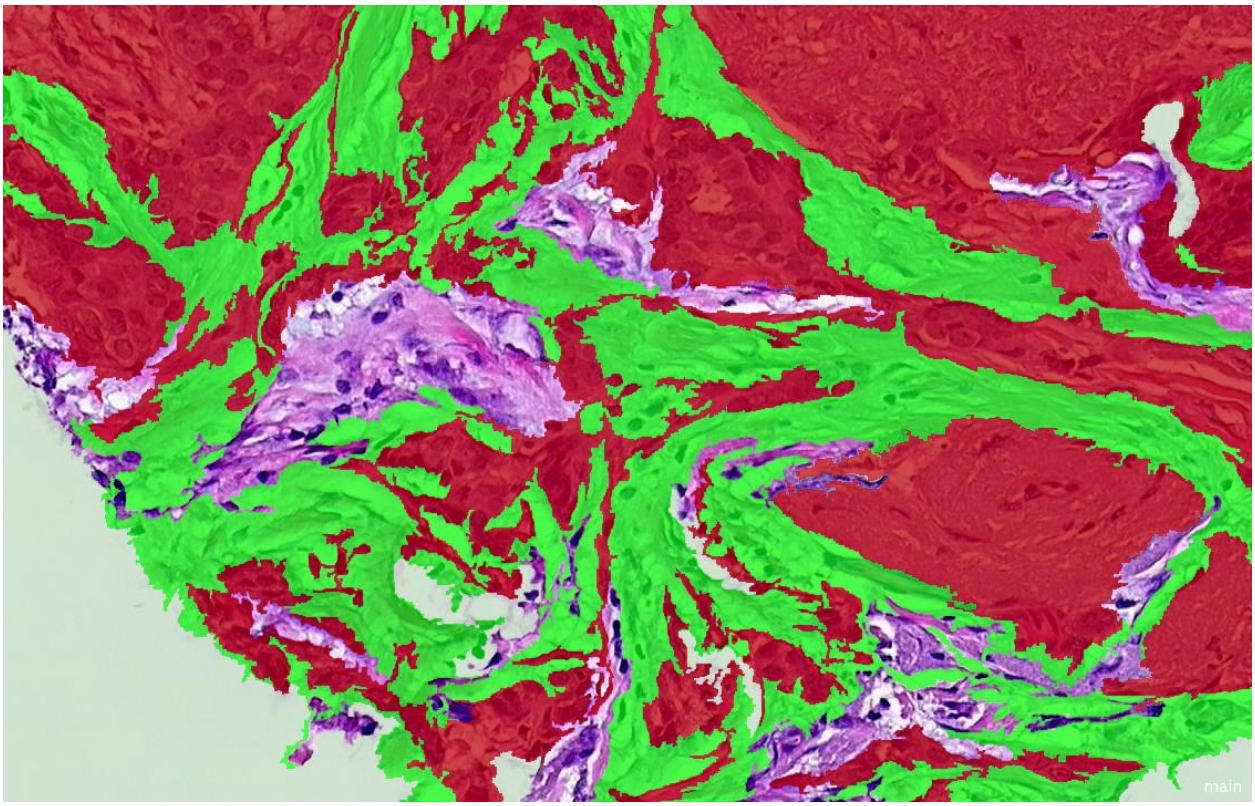

Figure 2: The same image with epithelial (red, 1s in the column EpiOrStroma) stromal (green, 2s in the column EpiOrStroma) superpixels identified.

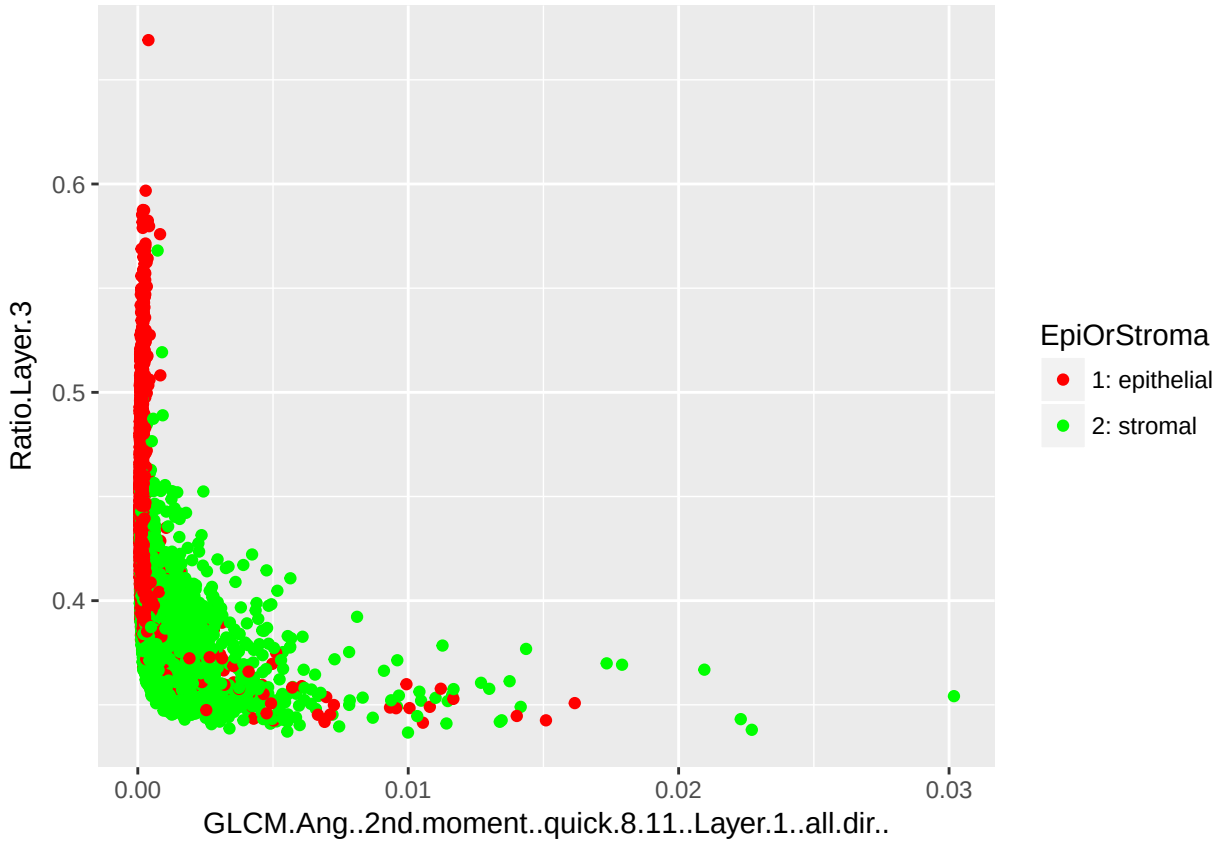

Figure 3: Scatter plot of features `GLCM.Ang..2nd.moment..quick.8.11..Layer.1..all.dir..` and `Ratio.Layer.3`.

2. Coursework task

The task of this coursework is to construct a epi vs stroma superpixels classifier. You're free to split data into train and test however you like. You will use the data to answer following two questions:

- Compare the preformace of two classification algoritm of you choosing (Using metrics you learned in the course)
- Can you obtain better performance by using only a subset of the features? Note: be careful about using test data to train

You can write classifier code from scratch or use any available libraries. However, if using third party libraries it must be clear from your report that you know how the algorithms work.

3. How will it be assessed

3.1 All students

You will submit code and a report. Ideally, your code will be of the form of an jupyter notebook (and not need any packages other than numpy and matplotlib etc). Your report should be in the style of a paper – describing the research question, describing why you took the approach you did, how your chosen models work, how you optimised any parameters and ultimately answer the research questions. Note: level 4 students, your report does not need to include a literature review.

3.2 Level M students

You need to additionally include a short (approx 1 page) literature review. This should cover general work in the area of identifying cell types from images using machine learning. Fuchs and Buhmann (2011) is a good starting point. Review on computer vision features and links to the features in the datasets is encouraged.

4 Deadline

The deadline for this work is **4pm on Monday 6th March 2017**.

5 Mark scheme

- (20%) Code that reproduces your results. Note that you can use external libraries but you must be able to explain how the models work in your report. Please make it very clear if your code relies on any non-standard external libraries. Use any language you like. Code should be clear and well commented.
- (20%, Level M students only) Literature review covering why identifying different cell types from images is problematic and the state of the art in Machine Learning in this area (1 to 2 pages). Include this in your report (below) – do not submit as a separate document.
- (80% (L4), 60% (M)) Report:
  - (25% of report total): Description of problem and justification and description of model used.
  - (25% of report total): Discussion and justification of assumptions in model and how parame- ters were optimised.
  - (25% of report total): Scientific quality of answer to research question.
  - (25% of report total): Overall written report quality. Clarity of writing, use of visualisations etc.

6 When should you do this?

Whilst we will be happy to answer coursework-related questions in the lab sessions, you should not use the lab sessions to actually do the work. Feel free to use the discussion forum on Moodle/Piazza to post any questions. Remember that this must be your own work.

Reference

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., ... & Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science translational medicine, 3(108), 108ra113-108ra113.

Fuchs, T. J., & Buhmann, J. M. (2011). Computational pathology: Challenges and promises for tissue analysis. Computerized Medical Imaging and Graphics, 35(7), 515-530.