BMI 260 ASSIGNMENT #2 | Mammogram Benign/Malignant Classification

Spring 2017

Darvin Yi (darvinyi[at]Stanford.edu)

Introduction

Breast cancer has the highest incidence and second highest mortality rate for women in the US. Our study aims to utilize deep learning for benign/malignant classification of mammogram tumors using a subset of cases from the Digital Database of Screening Mammography (DDSM).

Alright folks, congratulations on getting done with the segmentation assignment. As this is the second assignment, this one will be much less hand-holding than the previous assignment. Refer back to the lectures to get ideas about how to build your algorithm. Though we'll still give you some sense of possible backbone structure for this assignment, really do consider this an opportunity to get creative and create a truly spectacular machine learning algorithm.

Similar to the previous assignment, you'll be linked to the data below, but you aren't given any starter code or anything. You're 100% free to do whatever you want. Again, you don't even have to do benign/malignant classification. We give you the segmetations, so you could create a segmentation program again. The recommended project is that you use the given ROI's to extract features, and then train a classifier based on those features using the algorithms presented to you in the machine learning lectures. However, feel free to completely deviate from that if you think you know better. For example, you might know about the general success convolutional neural networks has had in the field of image classification. Feel free to build a deep learning approach for the problem if you see that as a better fit for this project. As usual, we'll be grading you much more on your report than your code or your results.

Assignment Description

This recommended assignment will have you classify masses in mammograms as either malignant or benign, an extremely difficult problem from a computational point of view. You will be given a set of mammography images and the segmentations of the masses, as well as ground truth labels for whether they are malignant or benign. Your job is to build a classifier that can take the given information and report whether a mass is malignant or benign. For test cases, you may assume that you'll get a mammogram of similar size as your given training set and a segmentation file as well. We hope you can put together all the things you've learned in lecture so far, including (but not limited to) morphological feature extraction, texture features, and machine learning.

Similar to the first assignment, you can work on any other project as long as it uses the mammogram data. Feel free to approach the mammogram benign/malignant classification from a different approach than the ones listed below as well (for example, with neural networks). You're all the lords and ladies of your assignment, and I can't wait to read all that you've written on this topic.

The Data

The data you're given comes from the Digital Database for Screening Mammography. This is an open source database of mammograms. Databases like the DDSM and TCGA/TCIA, which give wide access to these complex medical images are crucial for the advancement of research and algorithm development. You can read more about the DDSM at `http://marathon.csee.usf.edu/Mammography/Database.html`.

You will be given a large amount of data (approximately 400 images) pertaining to a subset of patients available in the DDSM. These images from been saved as .tif's with the naming convention P_⟨PatientNumber⟩_⟨PatientSize⟩_⟨View⟩.tif. The segmentation ROI's have been given to you in a similar naming convention, also as binary .tif files. The view's for each patient can be "CC" or "MLO." CC stands for Cranial-Caudal, and can be thought of as a view from above. MLO stands for MedioLateral-Oblique, and is an oblique or angled view.

Finally, you will also get a file named labels.csv. This CSV file contains the descriptions for the masses of all the patients in DDSM (a superset of patients that you will be given). The label "0" corresponds to benign cases and the label "1" corresponds to malignant cases.

You can download the data from Canvas.

Group Work

You can work in groups of 2 without any change in how we'll grade your report. If you're really endeavoring to do something grand, you can work in bigger groups (for example, tackling deep learning). However, with 3 or more people in a group, we'll expect you to have novel results, maybe at the level of publishing in a real journal/conference. This scales with people. With 4 people, I'm gonna expect a Nature level publication. With 5, you better win a Nobel Prize.

Submission Description

For this project, you will submit a .zip file containing everything you'd like to be considered for grading. Please name it "assignment1_⟨suid1⟩-⟨suid2⟩.zip" where you replace ⟨suid1⟩ with your actual Stanford University net id of the first person, ⟨suid2⟩ with the second person and so on. Your .zip file should contain, but should not be limited to

- report_⟨suid1⟩-⟨suid2⟩.pdf: The main publication camera-ready report showing off your findings and explaining your algorithm. Format your report so that it follows some major journal/conference format. Completely acceptable formats include MICCAI, IEEE, NIPs, and ICCV. Other formats are valid too; you're free to roll the dice on that one.
- `main.py` (`.sh`, `.m`, etc...): Some file we can run to produce the results detailed in your report. All your dependencies should be called by this main file. Your README.txt should give some detail as to how to run your programs.
- `README.txt`: A general guide as to how we should run your code. Also, if you added supplementary material, explain to us what it is and why you think its cool.
- Code Dependencies
- Supplementary Material: whatever else you want us to consider

Your total .zip file should be less than 20mB in size. If your .pdf is large, consider putting it through some compressing program. This is a hard limit.

The assignment .pdf will follow the format of a submittable paper in a reputable journal/conference. Thus, feel free to focus more on what you consider the more important aspects of what you discovered during the course of this assignment (i.e. following normal conventions of a paper, you don't/shouldn't focus on how you used `scipy.misc.imread` to read in the .tif's). Completing the tasks outlined in the assignment alone will be enough to write a paper around; however, we highly encourage you to pursue one of the extra areas outlined below. We are pursing a pure paper format for the submission reports for the following reasons:

- thinking about the assignment as a submittable paper will help focus your work
- writing an introduction will force you to do a literature review of what you're working on
- problem sets suck and aren't applicable to life. papers suck less. get used to writing papers.
- if you do indeed do something cool, you're ready to submit this right away, even if it's just to arXive.

As for the grade, we will resort to the tasks outlined below as a rough backbone. However, please do try to experiment more and explore this segmentation task or 3D data for more novel discovery without fear of falling behind in terms of the grade. Highlight these extra novel contributions in your paper and we'll do our best to credit everything.

Though the majority of your grade will be dependent on your .pdf, we will still take a close look at your program, and we will attempt running it. Please make understanding your code as painless as possible via comments and your README.txt file.

Understand your Data

The first and arguably most important step as a data scientist is for you to understand what and how your data is. I know that was a really clunky sentence, but I do mean it in the way I said. You should know how much of your data is benign vs. malignant. How many images do you have? How big are each image? How many patients do you have? What views are each of the images? The more you know about your data, the more you can form a concrete research plan. Most importantly, check out how valid you think the ROI's are. The data that you're given is real life research data. Thus, maybe it's not 100% accurate. Describing your data may help you to develop exclusion criteria, should you deem that necessary.

Extract Features: Recommended Approach

In the old-school paradigm of machine learning with medical images, there are three general parts: (1) segmentation, (2) feature extraction, and (3) learning. If you wish to follow this paradigm, we have given you the first part, the segmentation. There are a plethora of features that you can extract. Try extracting:

- **Morphological Features** You are given a segmentation of the mass. Thus, before even going into the actual mammograms, you can calculate a whole bunch of features on the mammograms. These include things we've covered in lecture, like perimeter and area. For binary images, MATLAB has a very nice command called `regionprops` that can take in a binary image and a list of features you'd like to extract, and spit you them out. I definitely recommend looking into this.
- **Histogram/Statistics Based Features** You have a whole bunch of pixels in your image. You got the actual pixel values in your total image. You could probably do some sort of breast segmentation, and then you'd have the pixel values in that region of interest. And finally, you have the pixel values in the mass ROI as well. Thus, even without taking into account neighboring pixel information, you could gather a whole bunch of statistics just based on those subsets of pixels. For example, you could calculate the mean, median, and variance of the pixel values. You could also bin the pixel values into a histogram, and use those bins as features themselves. If you do decide to go for a histogram based result, make sure you think it through. For example, should use normalize that histogram? How many bins should you use? Should you use a constant number of bins and let MATLAB choose the bin centers/edges for you, or should you use a constant set of bin centers?
- **Derivative Information** There's a whole set of algorithms that I'd like to call derivative based, in that they're all derivative of the derivatives. You can transform any image into a whole bunch of other images. You've learned about the curvature images and edge images. You can possible use these transformations to your advantage. For example, if you have a transformed gradient-based image, you could use the above methods to extract the histogram/statistics based features from those images.
- **Texture Information** You've learned in lecture two strong methods of texture feature extraction: Gray Level Co-occurence Matrix based texture (e.g. Haralick Features) and Convolution based texture (e.g. Gabor Features). MATLAB has a function `graycomatrix` that extracts a GLCM for you, from which you can calculate all the Haralick Texture Features as well. You can also use a bank of filters, like Gabor Filters, and make multiple convolution products of your original image, in which you can extract more statistical feature.s

Learning Benign vs. Malignant

No matter what approach you take, if you set out to classify mammograms as benign or malignant, you'll at some point need to build a model to learn which mammography masses are benign and which are malignant. Thus, to do that, you'll need to do a few things. Overall, everyone who chose to do a predictive modeling homework should be creating some mapping from some $\mathbf{X}$ to some $\mathbf{y}$, i.e. $\mathbf{X} \mapsto \mathbf{y}$.

Extract the ground truth data that is your $\mathbf{y}$. In this case, your $\mathbf{y}$ is going to be the benign/malignant labels that you'll extract from the CSV file. The way I would do it is to create a hash-table that maps the patient ID to the pathology column, and then upon training, you can easily extract what your ground truth is. Let's just say this: if you don't extract the correct information for your ground truth, you can have the best algorithm in the world, and it won't count for anything.

Clearly define what your $\mathbf{X}$ is. The starting point for everyone's assignment will be the images and the segmentations. However, depending on your model, your $\mathbf{X}$ that will be mapped onto your ground truth $\mathbf{y}$ will be wildly different. For example, if you do feature extraction, you should end up with a n by p matrix $\mathbf{X}$, which will be have n rows for each patient and p columns for each feature that you've extracted. If you're running off of a more deep learning frame work, you're $\mathbf{X}$ might be a three dimensional matrix of size $n \times m \times m$, where you'll have n patient's worth of $m \times m$ images that you might cropped out of the original images given the information of the mass ROI. This also includes more in-depth choices based upon your analysis from the "Describe Your Data" section. For example, should you use all the data? Should you create some sort of exclusion criteria? Should use use both CC and MLO views? Should use use both Left and Right breast images? Make sure that you're ready to defend your choices.

Choose a learning framework. In lecture, we've gone over a few classification algorithms. However, there are a whole lot more. Clearly define how you'll go about doing your classification. Maybe you'll try a few, in which case, report which ones have worked the best. Clearly define for me all your hyperparameters. Maybe you're going for more of an ensemble type thing, where you'll train many models together. If that's the case, define your voting metric. Basically, I want a data-scientist to be able to read your report and repeat your work exactly.

Results and Interpretation

If you set out to do mass classification, you will be judged mainly based on how you show us results. For example, if you chose to do some sort of linear fit (e.g. logistic regression), you can possible give us p-values on your fit. However, if you're using a slightly more complex model, we understand that it might be harder to come up with a Null Hypothesis for a p-value, so maybe you decided to give us a test accuracy. We won't be grading you on your actual numerical results (i.e. p ¿ 0.05 is fine, we just want you to acknowledge it).

The next important thing we would like is for you to interpret your actual results. If you did feature extraction and learning, can you tell us which features were the most important? Or maybe you can tell us the relationship between features. If you've taken a more deep-learning approach, maybe you can do some neural net visualization to figure out how the conv-net thinks. Either way, want you to go one more step than a black-box implementation.

Statistical Validity

This is going to be an interesting section, mainly because we want you to argue for the actual validity of your results. Basically, we want to make sure everyone is reporting results that are 100% statistically accurate even if your results do not show 100% testing accuracy. For example, if you chose some sort of hyperparameter, did you choose them randomly in order to make sure your you can be statistically kosher, or did you do some sort of k-fold Cross Validation separate from your test-set to make sure your final accuracy measurements were not down-biased.

Again, I know a lot of you guys might be a bit confused about this nebulous section of "statistical validity." The idea is that we want to make sure you take some time to think about why your results are actually valid. Are they repeatable? Do you think there might be some pit-fall? Or maybe you think you created the end all classifier for mammogram masses. Basically, reporting results is easy. Make sure you're ready to defend them though.

As an aside, by the design of problem sets, we need to make your score a linear combination of a lot of sections. Basically, for each section score S_i w.r.t. some rubric (see rubric below), your final grade for an assignment will be $\sum_i S_i$. However, you should know that in a more research oriented system, there are some binary switches. In my opinion, statistical validity is one of those switches. No matter how beautiful the story you weave about the interpretation of your hard-trained model, no matter how many GPU's you've used, if you mixed up training set and validation set, your research will never be accepted. Thus, a more accurate interpretation might be $\mathbf{1}\{\text{valid?}\}\sum_i S_i$, where if your approach wasn't statistically valid, you're left with nothing. Clearly, this isn't a way we can grade the course. That being said, make sure you're ready to defend all of the results you have reported and all the decisions you've made.

Grading Rubric

Table 1: **Grading Rubric** As stated in multiple places in this problem set, this rubric is only meant for your defense. We can be much more generous with the rubric. This rubric should just be used as a general guide to ensure that you have some goal.

Section	Description	%
Title	Choose a good title that is concise, clear, and fully demonstrates the problem you are attempting to solve.	5
Abstract	A proper scientific abstract that discusses • the general problem to be solved (2) • the main contribution to science (2) • the summary results (1).	5
Introduction	This will mainly deal with doing some literature research. This is an extremely loose rule of thumb, but aim to have maybe 10 citations in this section. We want to see • background research on the problem (10) • background research on other solutions (10) • background information on your methods (5)	25
Dataset	Talk about the dataset that you are using. Give some background on where the data came from and let the reader understand what the data your using is. Cite the data source.	10
Methods	Explain the approach you used. You don't have to go into too much of the detail if the detail is trivial. You can assume the general knowledge of the field you're writing your scientific paper to.	15
Results	Showcase your results. This is a very visual project. Figure out a way to create a wonderful figure that really proves that you've accomplished what you want to create.	15
Discussion	Discuss your algorithm. Tell us where it stands within the current field. What value have you added to the medical imaging community. Finally, make sure you do talk about the limitations of your research. Nothing is worse than overstating your results. This really is a tightrope walk. You want to showcase how sexy your results are, but you don't want to overstate them either. Make sure you present your results in a statistically valid way.	20
Style	Basically, aim to write well and keep all your code clean. If you do switch off writing some sections, make sure that the voices between sections don't clash in a jarring way.	5

NOTE: Again, the above is a proposed rubric for what I (Darvin) think is a good scientific paper. The main requirement is that you write a mock-paper (in any reputable journal). There are different theories on how papers should be formatted and what should go in each section. If you have a preferred style, do whatever you want. Obviously, some parts of the rubric (e.g. the title, abstract, and introduction) are a bit more standard, but format the rest however you want. We're really not trying to trap or hurt you here.

Frequently Asked Questions

- **How does grading work?** *Don't worry about grading so much.* But I know you will. If getting or securing that perfect hw score is an important thing in the world to you, just do the minimal tasks outlined in this problem set and showcase your segmentation results together with your pipeline and focus on the write-up as a minimum, since you can use that to defend why you deserve full credit later on.
- **Is —— a good addition/substitution to the assignment?** Ask on piazza.
- **Do I have to use LaTeX?** *No.* You just have to use the style of a reputable journal/conference. Many journals/conferences have templates in word or whatever. Just submit a .pdf (i.e. print to pdf). We don't want to deal with figures being out of place.
- **Is —— a reputable journal/conference?** Ask on Piazza.
- **Access to computational resources?** The current assignment should be able to be done on a simple computer or the shared Stanford resource of `corn.stanford.edu`. If you are attempting to do something on a grander scale and would like additional computational resources, please contact the TA's. We have some cloud computing credits we are reserving for projects and future assignments as well as lab resources that we could possibly support you with.
- **Do we have late days?** *No.* You're an adult. Submit your stuff on time or fail gracefully.
- **What if my computer froze last minute, and that's why my assignment's late.** *Don't do this.* We'll only grade the most recently submitted .zip file after the assignment is due, so just submit your assignments often. Aim to submit your assignments at 9pm instead of midnight to give yourself a 3hr grace period in case of computer troubles.
- **What if I go through an emergency?** Obviously, that's different. Contact the TA's privately via email or Piazza for extensions. Life happens. we understand that and by no means want to make your life worse.
- **Can I contact someone to get extra help for my idea?** Yes. Feel free to come by on the Friday sessions to ask open forum questions. You can also use Piazza. Please email the TA's to schedule more help if you think you need something more akin to office hours.
- **This is homework is going great. Can I turn it into a more research-like project?** Oh, god yes. Please feel free to treat the assignment like a sandbox to explore new discovery. This is the whole point of creating a report akin to a publication. Just because this is a homework for a class shouldn't limit you to staying on some pre-specified tracks.