

Problem Set 5

Posted: Tuesday, March 21, 2023

Due: Tuesday 11:59 pm, April 4, 2023

The relevant material for Pset 5 was covered in Lectures 9 and 10.

6.8300 students are expected to finish **all** problems.

6.8301 students are expected to finish all problems, **except 4c, 5c**.

Please note that 6.8301 students will *not* receive credit for 4c, 5c.

We provide a Python notebook with the code to be completed. You can run it locally or on Google Colab. To use Colab, upload it to Google Drive and double-click the notebook (or right-click and select Open with Google Colaboratory), which will allow you to complete the problems without setting up your own environment.

Submission Instructions: To submit your problem set, please navigate to the Pset 5 assignment submission window on Canvas and you will find a link that takes you to **Gradescope**. All you will need to submit is your **Python notebook**. Please make sure to run all the cells before submitting your notebook, so that we can inspect and grade your output in addition to your code. If a problem requires you to write text or you would like to write comments, the easiest way to do so is by adding a new text cell and writing into it.

Attention: Failure to follow the submission instructions will result in point deductions. For example, not running all the cells before submitting your notebook or submitting a zip file instead of just your notebook will be penalized.

Late Submission Policy: If your problem set is submitted within 7 days (rounding up) of the original deadline, you will receive partial credit. Such submissions will be penalized by a multiplicative coefficient that linearly decreases from 1 to 0.5.

Owing to the success of machine learning, vision algorithms have made rapid headway in various aspects of our life, from video surveillance and automated resume screening to clinical diagnostics. Due to their usage in a wide variety of situations, ethical concerns arise during each step of development, including data collection, model development, and model deployment. All these issues oftentimes lead to unfair products: for example, racial and gender biases have been reported in many instances of face detection algorithms [1]. Similarly, a recent study has shown that some computer vision models detecting COVID-19 from chest radiographs rely on confounding factors rather than medical pathology, creating an undesired situation in which the systems may appear accurate, but could lead to catastrophic failure when tested in new hospitals [2]. Under certain circumstances, these biases can lead to harmful consequences, which makes it crucial for CV practitioners to be aware of the Social and

Ethical Responsibilities of Computing (SERC) ¹.

The aim of this problem set is to study some common issues that may appear when developing a clinical decision support model. We will use the CBIS-DDSM/DDSMM dataset, which is a database of scanned film mammography images. It contains normal, benign, and malignant cases². You will train a simple model to classify these images, learn to evaluate the model using different metrics, and analyze some of the problems that arise when using these systems in practice. This Pset is not a comprehensive summary of the potential sources of bias/harm, and the problems studied are heavily simplified so that they can be studied concisely. However, we hope that this pset and the references provided will give you a start to delve deeper into this important set of topics.

Problem 1 *Train an Image Classifier (2 points)*

You will use the experience you gained in Pset 4 to train a simple image model that learns to produce a diagnosis in the categories: normal, malignant and benign. Experiment with some of the hyperparameters, techniques, and model types that are proposed in the notebook.

List the final model type, set of techniques, and hyperparameters you choose (e.g., ResNet-18, data augmentation by rotation, SGD optimizer, learning rate of 0.5, 10% dropout) and explain how they affect model performance (theoretically or intuitively). Report the train, validation, and test top-1 accuracy of your final network. Your trained network should achieve a test accuracy of at least 90% to get a full score on the problem.

For the rest of the problem set, you can either use your model or the one trained by the instructors (which is automatically downloaded in the notebook). If the test performance of your own model is less than 93%, please use the latter.

Problem 2 *Metrics (2 points)*

In the previous question, you may have achieved high accuracy, but that doesn't necessarily mean that the model is useful for clinic decision support in practice. Suppose that your model just learns nothing and outputs "normal" for every image. If 90% of the patient images in your dataset are normal, the accuracy is still as high as 90%! In this question, we will introduce some other metrics commonly used when your dataset is imbalanced, like in biomedical imaging tasks. **For a better understanding of these metrics, please don't use pre-defined functions for these metrics from other libraries.**

(a) Confusion Matrix (1 point)

Classification accuracy alone can hide important details if you have an unequal number of observations in each class or if you have more than two classes in your dataset. A confusion matrix helps you summarize the performance of a classification algorithm and can give you a better idea of what your classification model is getting right and what types of errors it is making.

¹<https://computing.mit.edu/cross-cutting/social-and-ethical-responsibilities-of-computing/>

²<https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM>

In the provided notebook, **build the function `plot_confusion_matrix()` to plot a confusion matrix for your selected model**. Your confusion matrix should approximately follow the template given in Figure 1, but the format doesn't need to match exactly.

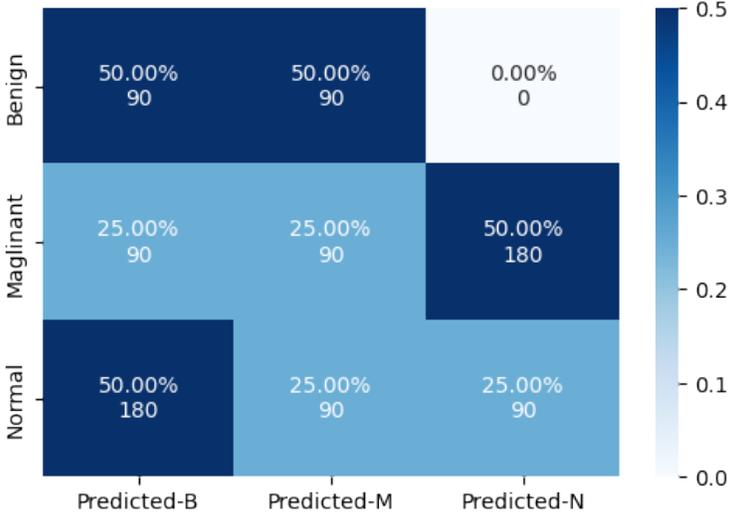


Figure 1: Confusion Matrix Template

(b) Precision-Recall Curve (1 point)

A precision recall curve is a 2D plot showing precision and recall values for different decision thresholds in a binary classification problem. Typically, the decision threshold is 0.5: if the predicted probability of class A is, say, 0.7, then the probability of class B will be 0.3 (they always sum to 1), and we choose class A as our predicted class.

For those classification problems that have a severe class imbalance, the default threshold can result in poor performance. As such, a simple and straightforward approach to improving the performance of a classifier that predicts probabilities on an imbalanced classification problem is to tune the threshold used to map probabilities to class labels.

Our dataset has 3 classes and in order to be able to compute metrics commonly used in binary classification, you need to create binary labels from the given class labels. We will assume a scenario where we really care about malignant cases. **Please consider the malignant class (label 1) as the positive class and consider the benign (label 0) and normal (label 2) classes together as the negative class.**

In this scenario, if we use the default threshold of 0.5, and our classifier predicts a 0.6 for class malignant, 0.3 for class benign and 0.1 for class normal (always sums to 1!), we would confidently assign a positive class to that example. If we change our threshold to 0.7, however, we will assume that our example is part of the *negative class* for the purposes of calculating precision and recall.

Before proceeding, please familiarize yourself with the following terms commonly used in binary classification: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Here, we are interested in two combinations of these quantities: Precision ($P = TP / (TP + FP)$) and Recall ($R = TP / (TP + FN)$).

Which one should you look at when you are trying to improve the model? It really depends on the problem at hand. For example, if you are trying to detect people with allergies, you don't want to alert too many people who may not have allergies (FP) and therefore you want a high precision. However, in cases such as cancer prevention, a false negative is usually more disastrous than a false positive so you want a high recall. You could have 100% recall by predicting the positive class for everyone but then you would have a completely uninformative model. This is where the precision-recall curve can be helpful in balancing the trade-off between precision and recall by using different probability cutoffs, or thresholds.

As we indicated above, the malignant class is our positive class, and the rest are negative classes. In this scenario, a positive prediction happens when our model's prediction for class malignant is above a user-defined threshold. That means that a True Positive happens when we have a positive prediction, and the ground truth label is malignant. A False Positive happens when we have a positive prediction and the ground truth label is something other than malignant. True negatives and false negatives follow the same logic.

In the provided notebook, calculate precision and recall for 1000 evenly spaced thresholds between 0 and 1, and plot the precision-recall curve of your selected model. Then, compute the AUC leveraging the function `trapz` from `numpy`, and display the AUC on the plot. Your output for this question should be one plot, with precision on the Y axis and recall on the X axis.

Problem 3 *What is the model looking at? (1 point)*

A common technique used to delve deeper into trained models is visualizations. As studied in the previous problem set, filter visualizations are sometimes useful to understand how the network behaves, but it is hard for non-expert users to interpret them.

Another technique previously studied is CAM visualizations, which provide a simple way to interpret what regions of the image contribute to the prediction for each output target. This is a useful tool to check whether the model is picking up spurious correlations or the predictions focus on regions of the image that are consistent with expert knowledge. Run Grad-CAM (or another CAM visualization) on several images for the different test sets, and explain the differences you see between correctly and incorrectly classified samples.

Problem 4 *Evaluation on an External Dataset (6.8301: 2 points, 6.8300: 3 points)*

In this question, we will study a problem that occurs frequently in practice when deploying systems. Oftentimes, the data distribution at test time is different from the one used during training.

(a) (1 point) To begin, visually compare the performance with the original model against an external dataset of the same type, provided in `external_dataset_dir`.

(b) (1 point) Compare samples of the two datasets visually, and report the differences that you observe. (Hint: if you need inspiration, check typical image transformations in PyTorch under `torchvision.transforms`.)

(c) (1 point) **[6.8300 only]** Implement a set of transformations that can be applied to the original dataset at train time that would mitigate this issue. Plot the original dataset with the proposed transformations, which should produce qualitatively similar samples as the external dataset. Although you don't need to retrain the model, you are free to do so to further check that the transformations are correct (performance on the external dataset should improve using the original data with the extra transformations). Use some of the available PyTorch transformations under `torchvision.transforms`.

Problem 5 *Class Imbalance (6.8301: 2 points, 6.8300: 4 points)*

An issue typically present in medical datasets is class imbalance: patients that do not have the disease are more frequent than patients that do have the disease. One way to make the trained model perform differently is by simulating a dataset that contains a different number of samples of each category. This can be achieved by at least two methods: reweighting the loss and changing the sampling procedure.

(a) (1 point) Implement reweighting of the loss so that all classes contribute equally to the loss.

(b) (1 point) Modify the dataset so that samples for the three classes are balanced.

(c) (2 points) **[6.8300 only]** Finally, train your original model with the two modifications, plot the confusion matrix for them, and answer the following questions:

- What are the empirical differences you see (if any) between the two approaches for addressing class imbalance?
- What are the theoretical differences between the two techniques?
- What are the strengths and weaknesses of the techniques?

Hint: think about gradient estimation in the extreme case where the number of samples of the less predominant class is low.

Reading list

- Standard deep learning models can be trained to predict race from medical images with high performance even when models are optimized to perform clinically motivated tasks: Reading Race: AI Recognises Patient's Racial Identity In Medical Images(<https://arxiv.org/abs/2107.10356>)
- Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle by Harini Suresh and John Gutttag. (<https://mit-secr.pubpub.org/pub/potential-sources-of-harm-throughout-the-machine-learning-life-cycle/release/1>)

References

- [1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.
- [2] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.