AIIM 113-1 Homework 1

Due: October 22, 2025

This homework aims to review the procedure of benchmark construction and better understand the evaluation metrics/tools for your final project.

Instructions

Please submit your answers in an English report (`.pdf`) and code (`.py`/`.ipynb`) on COOL. Name your files: "`studentid`-hw1.pdf" and "`studentid`-hw1.py/.ipynb". Remember to include your name and student ID in your report and in a comment in your code. Additionally, remember that your homework has to be able to run on Colab or Kaggle environments. You can check "HW1 grading policy (template)" for the grading policy. Requirements not met will result in a grade deduction. **Plagiarism and late submission are strictly prohibited.**

Problem 1 (35%) Benchmark Construction

Note that the datasets and models built from this problem will be used in the subsequent problems.

Load `breast_all.npz` attached or from Files/113-1 Final project: open datasets/Breast Cancer on COOL. Retrieve `x_train`, `c_train`, and `y_train` from `breast_all.npz`, which are training samples' gene expressions, clinical data, and the breast cancer prognosis labels, respectively. Finish the following steps:

1(a) (5%) Concatenate each sample's gene expression and clinical data as input features. Show a screenshot in the report illustrating the resulting training data and its shape.

1(b) (5%) Split the training data and labels into 75% training and 25% validation sets with a random seed of 0. Stratify your train-test-split with `y_train`. Show screenshots in the report illustrating the resulting dimensions of each set.

1(c)(5%) Conduct simple data exploration, checking for the distribution of clinical variables and labels in your training and testing datasets, and checking the imbalance rate in each.

1(d) (20%) Train the following models with your training set: logistic regression (LR), support vector machine (SVM), random forest (RF), and a feed-forward neural network (NN). Try different settings (e.g., number of estimators, regularisations, activation functions, or even data augmentation **on the training set**) until all models reach a validation AUROC above 0.730. Show each model's resulting AUROC in your report.

Note: You are allowed to adjust the random seed, but you'll need to describe the reason for doing so and your observation/comment in the report. Otherwise, you will lose 5% of the score.

Problem 2 (20%) Optimal Threshold

Retrieve `x_test`, `c_test`, and `y_test` from `breast_all.npz` to form your test set. Obtain the predicted probabilities of the validation and test set with the NN built from Problem 1. For this problem, you can choose to either use the F-1 or Youden's index to compute the threshold.

2(a) (10%) Find the optimal threshold for the F1-score/Youden's index through the validation set.

2(b) (5%) Find the corresponding score (F1-score$_{\text{opt}}$/Youden's Index$_{\text{opt}}$) of the validation and test sets.

2(c) (5%) Show how your confusion matrices for the validation and test sets vary using at least three different thresholds. Explain which threshold you prefer to choose and justify your answer. Hint: think about what kind of data is used for this assignment and identify which mistakes are more costly in a clinical setting.)

Problem 3 (20%) Precision-recall Curve

With the chosen threshold from Problem 2:

3(a) (10%) Plot the precision-recall curve (PRC) and obtain the area under the PRC (AUPRC) of the test set.

3(b) (5%) Failure analysis: Pick 5–10 misclassified samples and inspect their clinical vs genomic features. You can use plots to illustrate this and then further argue what might have caused the misclassification.

3(c) (5%) Find the corresponding PRC baseline and elaborate on its relationship with the proportion of class-1 samples in the test set. You may explain it mathematically or illustrate it through the scatter plot.

Problem 4 (25%) Concordance Index

Use the chosen threshold in Problem 2 and retrieve the event times `o_train` and `o_test` from `breast_all.npz`:

4(a) (5%) Calculate the concordance index (C-index) of the test set.

4(b) (5%) Calculate the AUROC of the test set.

4(c) (15%) Assume the dataset has infinite samples, and no sample receives exactly the same predicted probability. Examine the relationship between the C-index and AUROC based on the definition on p.26 of the workshop slides. Are they asymptotically equal? You may explain it mathematically or illustrate it through a scatter plot.