Question 1 (SVM) [30 pts]

In this question, you will train one soft margin and one hard margin SVM classifiers on the UCI Breast Cancer dataset [1]. The dataset, provided in `breast_cancer.csv`, contains 8 predictors for breast cancer. The last column of the dataset is the class label. In terms of class labels, 0 denoted benign tumor and 1 denotes malignant tumor.

You must perform 10-fold cross validation WITHOUT using any libraries but you CAN use libraries or software packages to train your SVM. Take first 500 rows as your training set and the rest as the test set.

Question 1.1 [15 points] In this part, you will train a linear SVM model with soft margin without using any kernels. Your model's hyper-parameter is C. Using 10-fold cross validation on your *training set*, find the optimum C value of your model. Look for the best C value with line search in the following range $[10^{-3}, 10^{-2} 10^{-1}, 10, 10^{1}, 10^{2}]$ and calculate accuracy on the left-out fold. For each value of C, calculate mean cross validation accuracy by changing the left-out fold each time and plot it in a nice form. Report your optimum C value. Then, run your model on the *test set* with this C value and report test set accuracy with the confusion matrix. Calculate and report micro and macro averages of precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1 and F2 scores.

Question 1.2 [15 pts] This time, use radial basis function (RBF) kernel to train your hard margin SVM model on the processed (discretized) dataset from Question 3. RBF kernel is defined as

$$K(x, x') = exp\left(- \frac{||x - x'||^2}{2\sigma^2} \right) \tag{0.1}$$

In RBF kernel formula, $\gamma = -\frac{1}{2\sigma^2}$ is a free parameter that can be fine-tuned. This parameter is the inverse of the radius the influence of samples selected by the model as support vectors. Similar to linear SVM part, train a SVM classifier with RBF kernel using same training and test sets you have used in linear SVM model above. In addition to the penalty parameter C, γ is your new hyper-parameter that needs be optimized. Using 10-fold cross validation and calculating mean cross validation accuracy as described in Question 4.1, find and report the best γ within the interval from the logarithmic scale $[2^{-4}, 2^{-3}, 2^{-2} 2^{0}, 2^{1}]$. After tuning γ on your *training set*, run your model on the *test set* and report your accuracy along with the confusion matrix. Calculate and report micro and macro averages of precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1 and F2 scores.

References

[1] D. Dua and C. Graff, "UCI machine learning repository," 2017.