Exercise 1: Performance Measures

(a) Given the following confusion matrices

$$M_1 = \begin{pmatrix} 0 & 10 \\ 0 & 990 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 10 & 0 \\ 10 & 980 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 10 & 10 \\ 0 & 980 \end{pmatrix},$$

each of which corresponds to a classifier. Compute the accuracy, F_1 score, G measure/mean, BAC and MCC of each classifier.

(b) What are the population counterparts of the class-specific variants in the multiclass setting of true positive rate, positive predictive value and true negative rate?

Exercise 2: Tomek Links

Implement the Tomek Links subsampling technique. The implementation can be decomposed into the following steps:

(a) Write a function `find_tomek_links` that identifies all samples belonging to Tomek links. The function accepts the input features `x` with shape `(num_samples, num_features)` and the class labels `y` with shape `(num_samples,)`. We assume that class label 1 represents the positive class, and 0 represents the negative class. In addition, the function should return a **binary** array with shape `(num_samples,)`, in which 1 means that the corresponding sample belong to a Tomek link. For example, if there are 6 samples in a dataset, and there are two Tomek links connecting samples (with 0-based indices) `0-1, 4-5`, respectively. Then, the returned indicator array should be `[1, 1, 0, 0, 1, 1]`.

(b) Write a function `find_kept_samples` that identifies the samples to be kept when performing subsampling. This function takes `x` and `y`, along with a binary array `is_tomek_sample` of shape `(num_samples,)` that indicates whether each sample belong to a Tomek link. The function should return a binary array of shape `(num_samples,)`, in which 1 means the sample should be kept, while 0 indicates it should be removed (because it is a sample of majority class and it belongs to a Tomek link).

(c) Write a script that run the experiment as follows:

(i) Generate an imbalanced dataset.

(ii) Visualize the imbalanced dataset, with positive and negative samples plotted in two distinct colors.

(iii) Find out the samples of Tomek links and visualize them in a different color.

(iv) Identify the samples that should be removed during subsampling. Visualize them in a different color.