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: Minimum Expected Cost Principle

Given a cost matrix $\mathbf{C}$

		True class	
		$y = 1$	$y = -1$
Predicted	$\hat{y} = 1$	$C(1,1) = 0$	$C(1,-1) = 1$
class	$\hat{y} = -1$	$C(-1,1) = \frac{n_-}{n_+}$	$C(-1,-1) = 0$

where n_+, n_- denote the number of testing samples with label 1 and -1, respectively. We further assume that we have knowledge of $p(\cdot|\mathbf{x})$. According to the Minimum Expected Cost Principle, $\mathbf{x}$ as class 1 if

$$\mathbb{E}_{K\sim p(\cdot|\mathbf{x})}(C(1,K)) \leq \mathbb{E}_{K\sim p(\cdot|\mathbf{x})}(C(-1,K)).$$

This yields the optimal threshold c^* for the probabilitic classifier $h(\mathbf{x}) = 2 \cdot \mathbb{1}_{[\pi(\mathbf{x}) \geq c^*]} - 1$. If we recompute the optimal threshold for $h(\mathbf{x})$ on a more imbalanced dataset with larger n_-/n_+, how will the testing PPV and TPR change?

Exercise 3: MetaCost

Implement the MetaCost algorithm and use it with some classifier of your choice on an imbalanced data set of your choice, where the cost-matrix is given by the cost-sensitive heuristic we saw in the lecture. Compare the confusion matrices of the underlying classifier and the MetaCost classifier as well as their total costs.