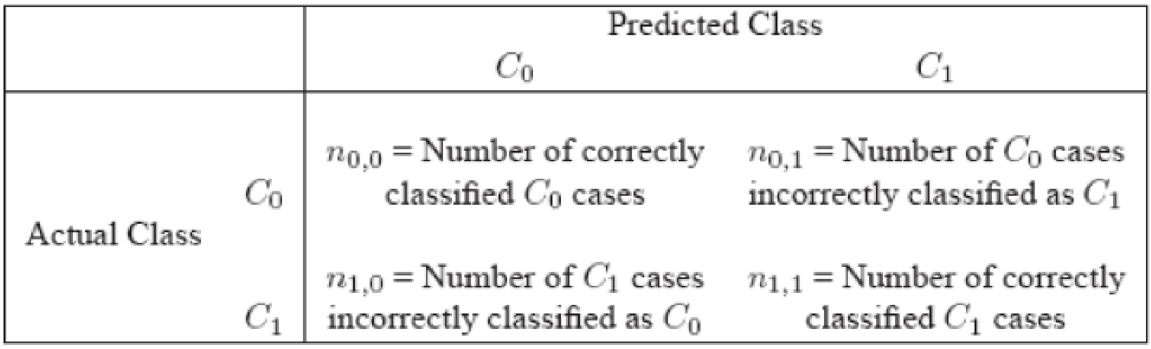
## Problem 5.1

5.1 A data mining routine has been applied to a transaction dataset and has classified 88 records as fraudulent (30 correctly so) and 952 as non-fraudulent (920 correctly so). Construct the classification matrix and calculate the error rate.

**Answer to 5.1:**

Explanation



| | | Predicted Class | |
|---|---|---|---|
| | | $C_0$ | $C_1$ |
| Actual Class | $C_0$ | $n_{0,0}$ = Number of correctly classified $C_0$ cases | $n_{0,1}$ = Number of $C_0$ cases incorrectly classified as $C_1$ |
| | $C_1$ | $n_{1,0}$ = Number of $C_1$ cases incorrectly classified as $C_0$ | $n_{1,1}$ = Number of correctly classified $C_1$ cases |

Therefore in our problem the confusion matrix is

| **Classification Confusion Matrix** | | |
|---|---|---|
| | **Predicted Class** | |
| **Actual Class** | 1 (fraudulent) | 0 (non-fraudulent) |
| 1 (fraudulent) | 30 | 32 |
| 0 (non-fraudulent) | 58 | 920 |

Error Rate = (n0,1 + n1,0) / n = (32 + 58) / 1040 = 0.0865 = 8.65%

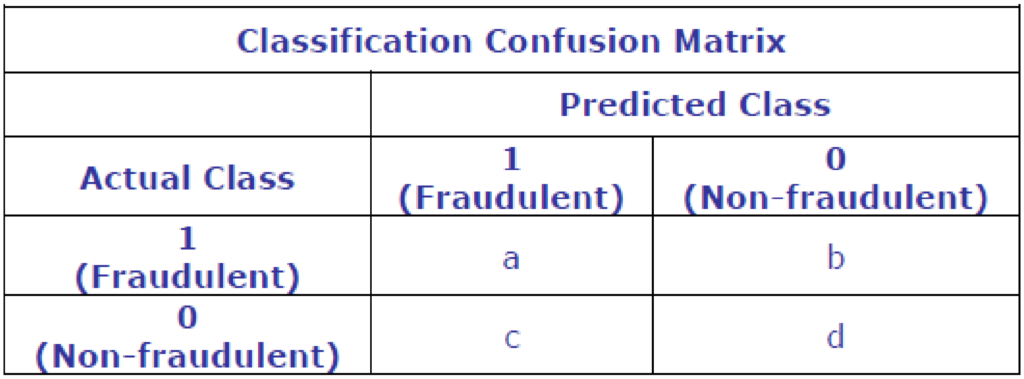
## Problem 5.2

5.2 Suppose that this routine has an adjustable cutoff (threshold) mechanism by which you can alter the proportion of records classified as fraudulent. Describe how moving the cutoff up or down would affect

a. The classification error rate for records that are truly fraudulent.

**Answer to 5.2.a:**

Let us assume a given cutoff, say, 0.5.



| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | 1<br>(Fraudulent) | 0<br>(Non-fraudulent) |
| 1<br>(Fraudulent) | a | b |
| 0<br>(Non-fraudulent) | c | d |

The classification error rate for truly fraudulent records (with this 0.5 cutoff) is a / (a+b)

The classification error rate for truly non-fraudulent records (with this 0.5 cutoff) is c / (c+d)

**Lowering the cutoff** (here, below 0.5) leads to classifying more non-fraudulent records as fraudulent records (more zeros misclassified as 1).
a. The numerator of the classification error rate for truly fraudulent records becomes less than or equal to b. In general, lowering the cutoff will result in a decrease (or no change) in the classification error rate for truly fraudulent records.
b. The numerator of the classification error rate for truly non-fraudulent records becomes greater than or equal to c. In general, decreasing the cutoff will result in an increase (or no change) in the classification error rate for truly non-fraudulent records.

**Increasing the cutoff** (here, above 0.5) leads to classifying more fraudulent records as non-fraudulent records (more 1 misclassified as 0).
a. The numerator of the classification error rate for truly fraudulent records becomes greater than or equal to b. In general, increasing the cutoff will result in an increase (or no change) in the classification error rate for truly fraudulent records..
a. The The numerator of the classification error rate for truly non-fraudulent records is less than or equal to c. In general, increasing the cutoff will result in a decrease (or no change) in the classification error rate for truly non-fraudulent records.