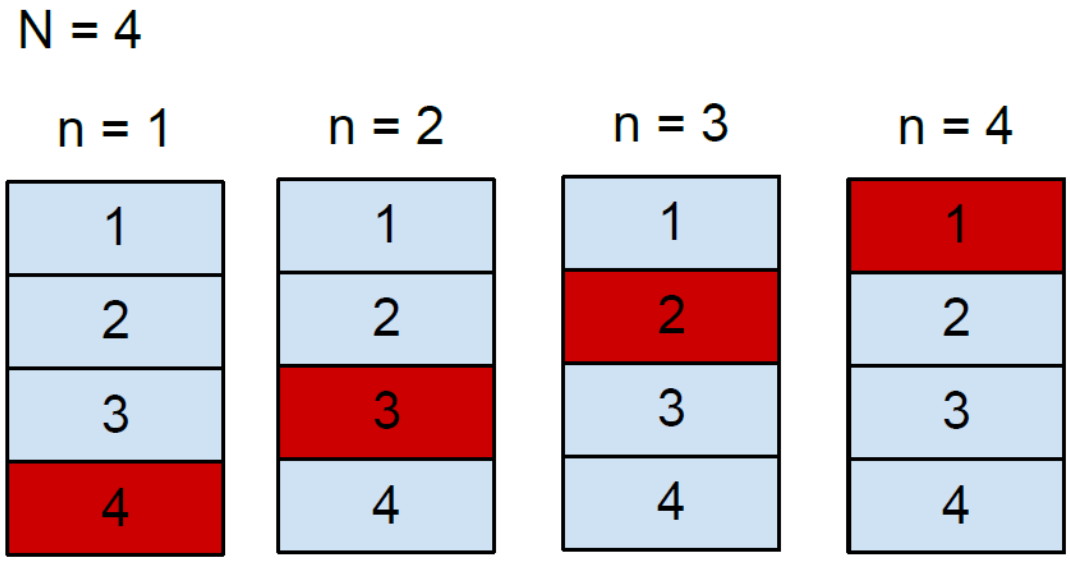
Klasyfikacja

W ramach każdej z metod klasyfikacji wyróżnia sie dwie operacje:
- Budowania (uczenia) klasyfikatora.
- Klasyfikacji nowych obserwacji.

W ramach pierwszej operacji konstruowany jest model na podstawie danych zawartych w zbiorze uczącym. Konstrukcja modelu może odbywać sie poprzez znalezienie parametrów funkcji separującej (sieci neuronowe, SVM), wygenerowanie zestawu reguł bądź drzew decyzyjnych, czy tez znalezieniu parametrów rozkładu (regresja logistyczna). W ramach drugiej operacji skonstruowany w procesie model klasyfikatora jest wykorzystywany do klasyfikacji nowych obiektów o nieznanych etykietach klas.

Walidacja krzyżowa
Celem oceny jakości klasyfikacji proponuje się metodykę walidacji krzyżowej (*ang. cross-validation*). Polega ona na losowym podziale zbioru danych na *N* (Najczęściej przyjmuje sie *N* = 10) w miarę równo rozłożonych części (tzn. foldów). Walidacja odbywa sie poprzez *N*-krotne wyuczenie klasyfikatora na zbiorze składającym się *N*− 1 części i przetestowaniu go na *N*-tej, nie wykorzystanej w uczeniu części. Istota tej metodyki testowania jest to, że w każdym kroku proces testowania odbywa się na innej części zbioru, a każda obserwacja ze zbioru będzie dokładnie raz przetestowana w procesie walidacji. Przykład działania metody walidacji krzyżowej (dla 4 foldów) obrazuje rysunek poniżej:



W pierwszym kroku (n=1) klasyfikator jest uczony z wykorzystaniem elementów 1,2,3 (kolor niebieski) a testowanie odbywa sie na elemencie 4 (kolor czerwony). W następnym kroku (n=2) do testowania brany jest zbiór, który nie był jeszcze

testowany, przykładowo ten o indeksie 3, a pozostałe części wykorzystywane są do uczenia. Proces jest powtarzany do momentu w którym każda z części nie zostanie wykorzystana do testowania.

	Zaklasyfikowany do klasy pozytywnej	Zaklasyfikowany do klasy negatywnej
Należy do klasy pozytywnej	TP (*True positive*)	FN (*False negative*)
Należy do klasy negatywnej	FP (*False positive*)	TN (*True negative*)

Miary jakości metod klasyfikacji

Podstawa oceny jakości metod klasyfikacji jest macierz konfuzji (*ang. confusion matrix*). Macierz konfuzji odpowiada na pytanie, jakie były tendencje w klasyfikacji pomiędzy klasami w odniesieniu do rzeczywistych etykiet klas obiektów. Typowym kryterium do oceny jakości jest poprawność klasyfikacji:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Innym wskaźnikiem oceny metod klasyfikacji jest wskaźnik specyficzności (znamienności, *ang. specificity*), nazywany również wskaźnikiem TN (*ang. TN rate*), i definiuje sie go w następujący sposób:

$$TNrate = \frac{TN}{TN + FP}$$

Kolejnym wskaźnikiem jest wskaźnik czułości (*ang. sensitivity*), bądz też wskaźnikiem TP (*ang.TP rate*), i wyrażony jest wzorem:

$$TPrate = \frac{TP}{TP + FN}$$

Bardzo ważnym wskaźnikiem jest wskaźnik średniej geometrycznej czułości i specyficzności:

$$GMean = \sqrt{TPrate * TNrate}$$

oraz wskaźnik AUC:

$$AUC = \frac{1 + TPrate - FPrate}{2}$$

Oba zadania zostaną wykonane na pliku *XXXXXXL4 1.arff*.

Ćwiczenie 1. Należy zaimplementować w Javie (z wykorzystaniem biblioteki Weka) program który będzie przeprowadzał testowanie jakości klasyfikatora z wykorzystaniem krzyżowej walidacji. Założenia programu:

(a) Program powinien działać niezależnie od metody klasyfikacji i wybranego zbioru uczącego (Należy rozważyć wykorzystanie klas *Classifier*, oraz *Instances*).
(b) Jak parametr programu należy zadać liczbę foldów dla walidacji krzyżowej oraz liczbę powtórzeń eksperymentu.
(c) Podział zbioru na równoliczne foldy musi być realizowany losowo.
(d) Program powinien w wyniku przeprowadzonego testu zwrócić otrzymana macierz konfuzji (będącą suma macierzy konfuzji zwracanych dla zbioru testowego w każdej iteracji walidacji krzyżowej, w przypadku większej niż 1 liczby powtórzeń elementy macierzy należy uśrednić), wartości *Accuracy*, *TPrate*, *TNrate*, *GMean*, oraz *AUC*.

Ćwiczenie 2. Wykorzystując program z poprzedniego punktu należy przeprowadzić badania dla zbioru z pliku analizę jakości metod klasyfikacji, takich jak **ZeroRule**, **JRip**, **J48**,**SMO**, **MultilayerPerceptron**, oraz **NaiveBayes** (*Uwaga !* przyjmujemy *status pożyczki* jako klasę, klasa *pozytywna* jest *zły* klient). Dla wybranych metod badania przeprowadzić dla różnych wartości parametrów i zidentyfikować najlepsze parametry ze względu na wskaźnik *GMean*, oraz *AUC*. Dla każdej metody należy przedstawić wyniki i dokonać ich interpretacji