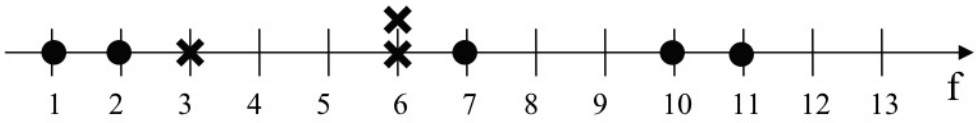
Raport z przebiegu laboratorium: Uczenie maszynowe Data:
Zespół (imiona i nazwiska):

Dzień i godzina zajęć: Grupa: .....

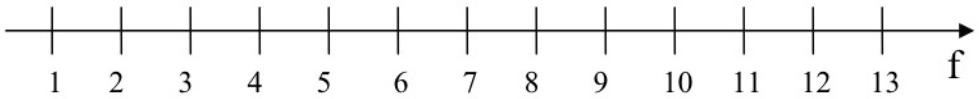
1. Poniższy rysunek prezentuje zbiór danych uczących opisanych pojedynczym atrybutem warunkowym ***f***. Atrybut decyzyjny przyjmuje jedną z dwóch wartości (klas decyzyjnych) – X i O (●).



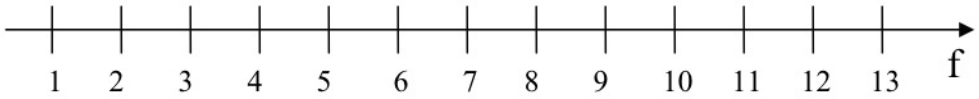
1.1. Powyższe dane zostały użyte do nauczenia klasyfikatora minimalno-odległościowego **1-NN**, a następnie do klasyfikacji nieznanych wcześniej danych. Jakie decyzje zostaną podjęte przez klasyfikator dla danych o wartościach atrybutu ***f*** równych:

f_1=1: _____, f_2=2.4: _____, f_3=6.5: _____, f_4=9: _____, f_5=12: _____

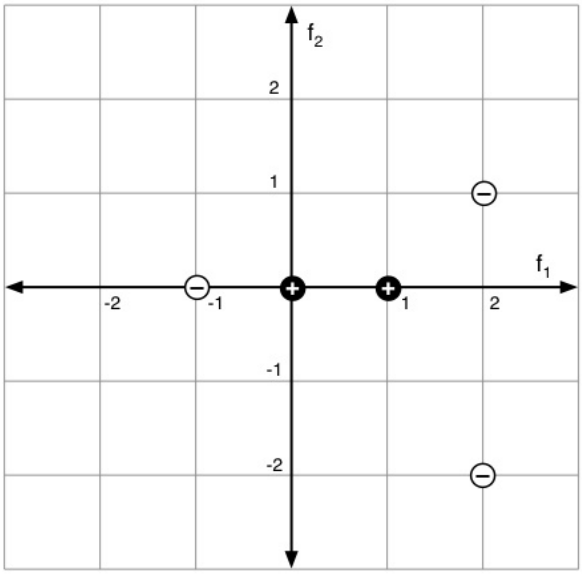
1.2. Zaznacz na poniższej osi przedziały wartości cechy ***f***, dla których nowy obiekt zostanie zaklasyfikowany do klasy O ( ).



1.3. Dane w punkcie 1 zostały użyte do nauczenia klasyfikatora minimalno-odległościowego **5-NN.** Zaznacz na osi przedziały wartości cechy ***f***, przy których nowy obiekt zostanie zaklasyfikowany do klasy O ( ).



2. Na poniższym wykresie przedstawione są inne dane uczące. Naszkicuj granice decyzyjne, których niejawnie używałby klasyfikator minimalno-odległościowy **1-NN** do klasyfikacji nowych przykładów.



2.1. Do której klasy decyzyjnej (+,-) zaklasyfikowany zostałby nowy obiekt o atrybutach f_1=1 i f_2=-1.01 przy użyciu klasyfikatora: a) **1-NN: _____** b) **3-NN**: _____. Rozpatrujemy odległości euklidesowe.

Do wykonania poniższych zadań wykorzystaj program WEKA. Zapoznaj się wcześniej z krótką instrukcją obsługi zamieszczoną na stronie (plik *instrukcja_weka.pdf*).

3. Poniższa tabela przedstawia dane uczące dla problemu klasyfikacji binarnej (2 klasy decyzyjne) o dwóch atrybutach numerycznych **A** i **B**. Dane te pochodzą z pliku *zad3.arff*. Korzystając z programu WEKA uzupełnij tabelę decyzjami zwracanymi przez klasyfikatory dla poszczególnych przypadków testowych i podaj skuteczności (trafności) poszczególnych klasyfikatorów.

   Konfiguracja: By WEKA wypisywała odpowiedź klasyfikatora dla poszczególnych przypadków (predykcje), należy w zakładce '*Classify*' znaleźć panel '*Test options*', kliknąć w przycisk '*More options…*' i zaznaczyć opcję '*Output predictions*'. W tym samym panelu powinno być zaznaczone '*Use training set*' – dzięki temu skuteczność klasyfikatora będziemy badać na tych samych danych, na których się uczył (w ogólności jest to zła praktyka ze względu na przeuczenie).

Dane uczące			**Odpowiedzi klasyfikatora**			
A	**B**	**Decyzja**	**1-NN**	**3-NN**	**11-NN**	**15-NN**
1	5	0				
2	6	0				
2	7	1				
3	7	0				
3	8	1				
4	8	0				
5	1	1				
5	9	0				
6	2	1				
7	2	0				
7	3	1				
8	3	0				
8	4	1				
9	5	1				
10	6	1				
		Skuteczność				

   3.1. Dlaczego trafność klasyfikacji na zbiorze uczącym nie jest dobrą miarą oceny tego klasyfikatora (zwłaszcza przy wielkości sąsiedztwa 1)?

   3.2. Co się dzieje, gdy sąsiedztwo zawiera wszystkie przykłady uczące?

   3.3. Podaj trafności klasyfikatorów **k-NN** dla k=1, 3, 11, 15 na zbiorze testowym wygenerowanym jako:
   a) podzbiór oryginalnego zbioru danych (percentage split, 33%)

   b) 15-krotna walidacja krzyżowa (cross-validation, 15 folds)

4. Używając oprogramowania WEKA zbuduj klasyfikator minimalno-odległościowy **k-NN** dla problemu diagnozowania białaczki, dla którego dane umieszczone są w pliku *leukemia.csv*. Przy jakim **k** trafność klasyfikatora (oceniana na podstawie 10-krotnej walidacji krzyżowej) jest największa? Ile ona wynosi?