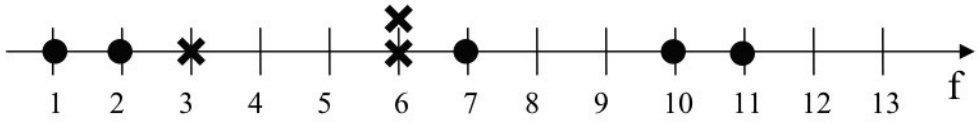
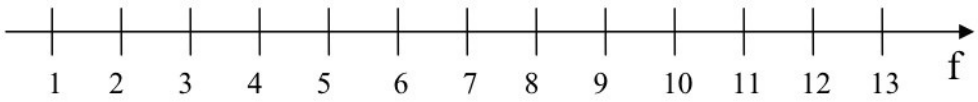
Raport z przebiegu laboratorium: Uczenie maszynowe Data:
Zespół: (imiona i nazwiska):

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.

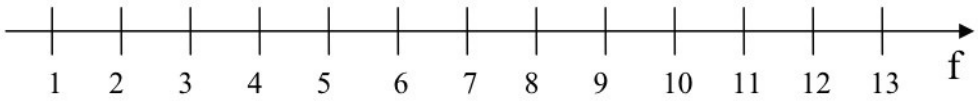


1.1. Powyższe dane użyte są do nauczenia klasyfikatora minimalnoodległościowego **1-NN**, a następnie klasyfikacji nieznanych wcześniej danych. Jakie decyzje zostaną podjęte przez klasyfikator dla danych o wartościach atrybutu *f* równych: f_1=1, f_2=2.4, f_3=6.5, f_4=9, f_5=12. Zaznacz decyzje na osi.

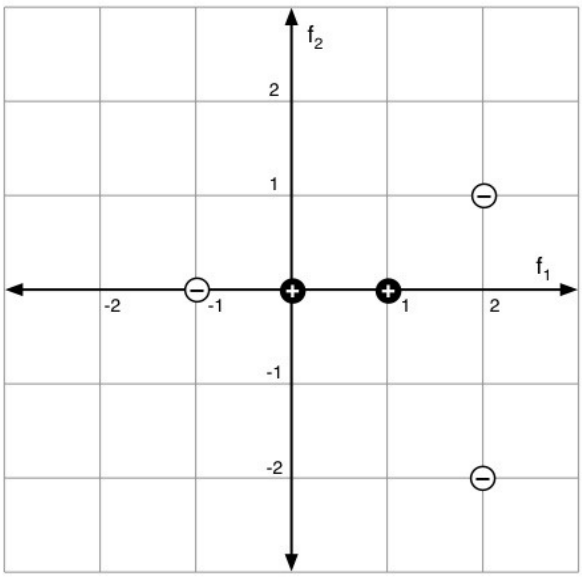


Zaznacz na osi przedziały wartości cechy f, przy których nowy obiekt zostanie zaklasyfikowany do klasy O.

1.2. Powyższe dane użyte są do nauczenia klasyfikatora minimalnoodległościowego **5-NN.** Zaznacz na osi przedziały wartości cechy f, przy których nowy obiekt zostanie zaklasyfikowany do klasy O.



2. Na poniższym wykresie zaznaczone zostały dane uczące. Naszkicuj granice decyzyjne, których używałby klasyfikator minimalnoodległościowy **1-NN** do klasyfikacji nowych przykładów. Do której klasy decyzyjnej zaklasyfikowany zostałby nowy obiekt o atrybutach f_1=1 i f_2=-1.01 przy użyciu reguły **1-NN**, a do jakiej używając 3-NN?



3. Poniższa tabela przedstawia dane uczące dla problemu klasyfikacji binarnej o dwóch atrybutach numerycznych (A i B). Jaka jest skuteczność (trafność) klasyfikatora na zbiorze uczącym przy zadanej wielkości sąsiedztwa (1, 3 i 11)?

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)? Co się dzieje gdy dla tych danych sąsiedztwo jest zbyt małe lub zbyt duże?

3.2. Użyj oprogramowania WEKA aby zbadać trafność klasyfikacji na zbiorze testowym wygenerowanym jako podzbiór oryginalnego zbioru danych (percentage split) oraz przy użyciu 15-krotnej walidacji krzyżowej (cross-validation). Jakie wartości k (wielkość sąsiedztwa) pozwalają osiągnąć maksymalną skuteczność?

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