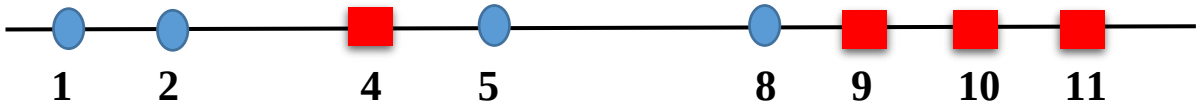
Test 2 seminar ML – 12 decembrie 2021, ora 16:00 (100 minute)

1. (1.8pt) Considerăm setul de date de antrenament din figura de mai jos. Rulați $T = 3$ iterații ale algoritmului AdaBoost, folosind drept clasificatori slabi compași de decizie (engl., decision stumps). Completaţi pentru acest set de date un tabel similar cu cel descris în enunţul problemei 24 (pag. 489). La final scrieţi formula analitică a clasificatorului H. Punctele marcate cu albastru le consideraţi ca având eticheta +1 iar dreptunghiurile roşii au eticheta -1.



2. (2.4pt) Aplicați algoritmul de clusterizare ierarhică aglomerativă folosind metrica lui Ward şi clusterizarea complete-linkage pe următorul set de date: A(0,0), B(3,0), C(0,2), D(4,0), E(4,2). Scrieți/calculați matricele de distanțe la fiecare pas de agregare. În caz de egalitate, agregați în ordine alfabetică. Desenaţi dendogramele rezultate. Înălţimea corespunzătoare fiecărui cluster non-singleton (adică, a fiecărui nod intern) din dendrogramă va fi considerată ca fiind egală cu distanţa (i.e., conform măsurii de similaritate) dintre cele două sub-clustere constitutive.

3. (1.8pt) Fie setul de date din $\mathbb{R}$: *-3, -2, -1, 1, 1, 2, 2, 3, 3*. Considerăm următorii centroizi de start pentru metoda *2-means*: $\mu_1^{(0)} = -5$ şi $\mu_2^{(0)} = -6$.

   i. Rulați algoritmul *2-means* până când componența clusterelor nu se mai schimbă.
   ii. Demonstrați în manieră analitică (NU numeric!) că pentru $t=1$ are loc relația:

$$J(C^{(t)}, \mu^{(t)}) \leq J(C^{(t-1)}, \mu^{(t-1)}).$$

   Definiţia criteriului J este următoarea:

$$J(C^{(t)}, \mu^{(t)}) = \sum_{i=1}^{n} \left(x_i - \mu^{(t)}_{C^{(t)}(x_i)} \right)^2 ,$$

   unde $C^{(t)}$ este ansamblul clusterelor la momentul/iteraţia t, $\mu^{(t)}$ desemneaza mulţimea centroizilor de la pasul t iar $\mu^{(t)}_{C^{(t)}(x_i)}$ este centroidul clusterului la care este asignată instanţa x_i, la iteraţia t.