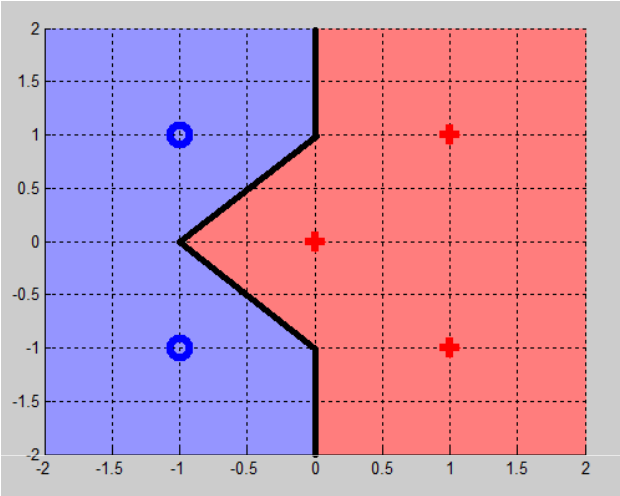
Zadatak 2: Algoritam k-NN

(a) Raspolažemo skupom primjera za učenje $\mathcal{D} = \mathbb{R}^2 \times \{0, 1\}$:

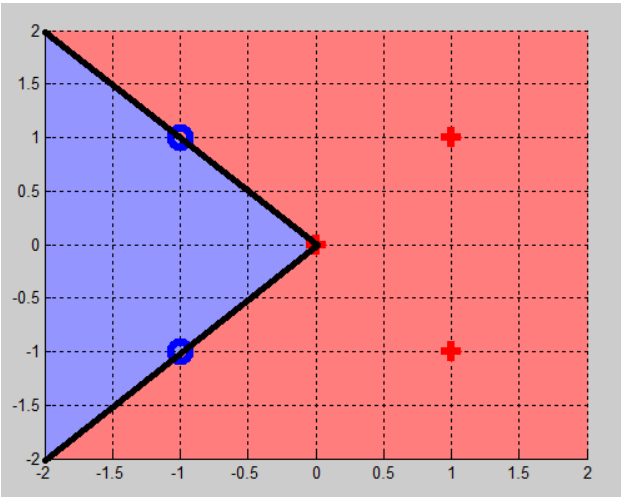
$$\mathcal{D} = \{((-1, 1), 0), ((-1, -1), 0), ((0, 0), 1), ((1, 1), 1), ((1, -1), 1)\}.$$

Skicirajte primjere u ulaznom prostoru te granicu između klasa za klasifikatore 1-NN i 3-NN.

Ovo su SKICE, nisam ništa računao, nemojte se ljutiti ako nije baš točno.



Slika 1. Zadani primjeri, i granica korištenjem 1-NN



Slika 2. Zadani primjeri, i granica korištenjem 3-NN

(b) Koristeći primjere iz zadatka 1, odredite klasifikaciju 8. primjera pomoću klasifikatora 3-NN naučenog na prvih sedam primjera uz uporabu (1) euklidske mjere udaljenosti i (2) Jaccardove mjere udaljenosti.

$$d(x^a, x^b) = \sqrt{\sum_{i=1}^{n} (x_i^a - x_i^b)^2} \qquad J'(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 1 - J(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = 1 - \frac{\sum_{k=1}^{n} \mathbf{1}\{x_k^{(i)} \wedge x_k^{(j)}\}}{\sum_{k=1}^{n} \mathbf{1}\{x_k^{(i)} \vee x_k^{(j)}\}}$$

Gore su formule za Euklidsku(lijeva) i Jaccardovu(desna) mjeru udaljenosti.

i	Euklid	Jaccard	$y^{(i)}$
1.	2	0.57	1
2.	2.45	0.75	0
3.	3.16	1	0
4.	1.41	0.33	1
5.	2.45	0.75	0
6.	2.45	0.75	1
7.	2.45	0.75	0

Tablica 1. Euklidskih i Jaccardovih udaljenosti

Tablica lijevo sadrži udaljenosti od prvih sedam primjera od osmog primjera. Kao što vidimo, 2 najbliža primjera su 1. i 4., a kako su oni označeni s 1, ne moramo znati koji je treći najbliži susjed jer već ova prva dva određuju da je oznaka ($y^{(8)}$) = 1.

(c) U nekim situacijama međutim ne želimo razlike u vrijednostima tretirati jednako i tada treba drugačije definirati mjeru udaljenosti. Je li to slučaj s problemom iz zadatka 1a? Ako smatrate da jest, predložite bolju mjeru udaljenosti.

Ako je svaki primjer sastavljen od n multinomijalnih značajki, a dva primjera se razlikuju u jednoj multinomijalnoj značajki, udaljenost je $\sqrt{2}$ u slučaju Euklidske mjere udaljenosti, a $\frac{2}{n+1}$ u slučaju Jaccardove mjere udaljenosti.

Pa vidite, nije nam to toliko bitno, ali mogli bi to koristiti kod nas, recimo za prosjek, srednji prosjek je više sličan visokom prosjeku nego je to nizak prosjek. Ako bi htjeli nešto takvo, mogli bi koristiti euklidsku udaljenost, ali umjesto binarnim vektorima, različite vrijednosti istog atributa bi predstavili realnim brojevima. Ili još bolje, vektorima realnih brojeva.